A First Course in
Design and Analysis
of Experiments

A First Course in
Design and Analysis
of Experiments

Gary W. Oehlert
University of Minnesota

Minitab is a registered trademark of Minitab, Inc.
SAS is a registered trademark of SAS Institute, Inc.
S-Plus is a registered trademark of Mathsoft, Inc.
Design-Expert is a registered trademark of Stat-Ease, Inc.

For Becky
who helped me all the way through
and for Christie and Erica
who put up with a lot while it was getting done

Contents

Preface to the Second Edition

It's been more than twenty years since the first edition came out. In that time, the hardcopy edition of the book sold reasonably well, but never in vast numbers; it then went out of print, and I distributed it for free in pdf format. Why a new edition now? There are several things that the first edition skimped on or left out entirely; there are more modern ways of doing some things; computing marches on; some mention of the so-called "replication crisis" needs to be included. Specifically, the second edition contains:

- Expanded coverage of response surfaces and mixture designs.

- Some discussion of optimal design and "computer designs," primarily in the contexts of non-regular fractional factorials, response surface designs, and mixture designs.

- Reduced emphasis on traditional p-value criteria such as .05 or .01 together with more discussion on replication of experiments and the hidden multiplicities of analysis.

- A broader array of analysis approaches including Bayesian methods, (restricted) maximum likelihood, and generalized linear models.

- Computing examples done primarily in **R**, with the vast bulk of the computational examples shifted a companion book *Extended* **R** *Examples for A First Course in Design and Analysis of Experiments*. The companion text is and e-book in HTML format (technically, GitBook) that includes not just the examples, but much more explication on using **R**. There are links from the main text to the companion text, and they even work on some platforms.

- Many more problems and data sets, and an **R** package that contains all of the data sets.

- Changes to the typography reflecting the fact the book will almost always be viewed on a screen rather than on paper (for example, no more recto-verso).

Finally, I would like to thank John Corbett, who has been a valued critic, cheerleader, and friend during the revision process.

Preface

This text covers the basic topics in experimental design and analysis and is intended for graduate students and advanced undergraduates. Students should have had an introductory statistical methods course at about the level of Moore and McCabe's *Introduction to the Practice of Statistics* (Moore and McCabe 1999) and be familiar with t-tests, p-values, confidence intervals, and the basics of regression and ANOVA. Most of the text soft-pedals theory and mathematics, but Chapter 19 on response surfaces is a little tougher sledding (eigenvectors and eigenvalues creep in through canonical analysis), and Appendix A is an introduction to the theory of linear models. I use the text in a service course for non-statisticians and in a course for first-year Masters students in statistics. The non-statisticians come from departments scattered all around the university including agronomy, ecology, educational psychology, engineering, food science, pharmacy, sociology, and wildlife.

I wrote this book for the same reason that many textbooks get written: there was no existing book that did things the way I thought was best. I start with single-factor, fixed-effects, completely randomized designs and cover them thoroughly, including analysis, checking assumptions, and power. I then add factorial treatment structure and random effects to the mix. At this stage, we have a single randomization scheme, a lot of different models for data, and essentially all the analysis techniques we need. I next add blocking designs for reducing variability, covering complete blocks, incomplete blocks, and confounding in factorials. After this I introduce split plots, which can be considered incomplete block designs but really introduce the broader subject of unit structures. Covariate models round out the discussion of variance reduction. I finish with special treatment structures, including fractional factorials and response surface/mixture designs.

This outline is similar in content to a dozen other design texts; how is this book different?

- I include many exercises where the student is required to *choose* an appropriate experimental design for a given situation, or *recognize* the design that was used. Many of the designs in question are from earlier chapters, not the chapter where the question is given. These are important skills that often receive short shrift. See examples on pages 627 and 622.

- I use Hasse diagrams to illustrate models, find test denominators, and compute expected mean squares. I feel that the diagrams provide a much easier and more understandable approach to these problems than the classic approach with tables of subscripts and live and dead indices. I believe that Hasse diagrams should see wider application.

- I spend time trying to sort out the issues with multiple comparisons procedures. These confuse many students, and most texts seem to just present a laundry list of methods and no guidance.

- I try to get students to look beyond saying main effects and/or interactions are significant and to understand the relationships in the data. I want them to learn that understanding what the data have to say is the goal. ANOVA is a tool we use at the beginning of an analysis; it is not the end.

- I describe the difference in philosophy between hierarchical model building and parameter testing in factorials, and discuss how this becomes crucial for unbalanced data. This is important because the different philosophies can lead to different conclusions, and many texts avoid the issue entirely.

- There are three kinds of "problems" in this text, which I have denoted exercises, problems, and questions. Exercises are intended to be simpler than problems, with exercises being more drill on mechanics and problems being more integrative. Not everyone will agree with my classification. Questions are not necessarily more difficult than problems, but they cover more theoretical or mathematical material.

This text contains many formulae, but I try to use formulae only when I think that they will increase a reader's understanding of the ideas. In several settings where closed-form expressions for sums of squares or estimates exist, I do not present them because I do not believe that they help (for example, the Analysis of Covariance). Similarly, presentations of normal equations do not appear. Instead, I approach ANOVA as a comparison of models fit by least squares, and let the computing software take care of the details of fitting. Future statisticians will need to learn the process in more detail, and Appendix A gets them started with the theory behind fixed effects.

Speaking of computing, examples in this text use one of four packages: MacAnova, Minitab, SAS, and S-Plus. MacAnova is a homegrown package that we use here at Minnesota because we can distribute it freely; it runs on Macintosh, Windows, and Unix; and it does everything we need. You can download MacAnova (any version and documentation, even the source) from `http://www.stat.umn.edu/~gary/macanova`. Minitab and SAS are widely used commercial packages. I hadn't used Minitab in twelve years when I started using it for examples; I found it incredibly easy to use. The menu/dialog/spreadsheet interface was very intuitive. In fact, I only opened the manual once, and that was when I was trying to figure out how to do general contrasts (which I was never able to figure out). SAS is far and away the market leader in statistical software. You can do practically every kind of analysis in SAS, but as a novice I spent many hours with the manuals trying

to get SAS to do any kind of analysis. In summary, many people swear by SAS, but I found I mostly swore at SAS. I use S-Plus extensively in research; here I've just used it for a couple of graphics.

I need to acknowledge many people who helped me get this job done. First are the students and TA's in the courses where I used preliminary versions. Many of you made suggestions and pointed out mistakes; in particular I thank John Corbett, Alexandre Varbanov, and Jorge de la Vega Gongora. Many others of you contributed data; your footprints are scattered throughout the examples and exercises. Next I have benefited from helpful discussions with my colleagues here in Minnesota, particularly Kit Bingham, Kathryn Chaloner, Sandy Weisberg, and Frank Martin. I thank Sharon Lohr for introducing me to Hasse diagrams, and I received much helpful criticism from reviewers, including Larry Ringer (Texas A&M), Morris Southward (New Mexico State), Robert Price (East Tennessee State), Andrew Schaffner (Cal Poly—San Luis Obispo), Hiroshi Yamauchi (Hawaii—Manoa), and William Notz (Ohio State). My editor Patrick Farace and others at Freeman were a great help. Finally, I thank my family and parents, who supported me in this for years (even if my father did say it looked like a foreign language!).

They say you should never let the camel's nose into the tent, because once the nose is in, there's no stopping the rest of the camel. In a similar vein, student requests for copies of lecture notes lead to student requests for typed lecture notes, which lead to student requests for more complete typed lecture notes, which lead ... well, in my case it leads to a textbook on design and analysis of experiments, which you are reading now. Over the years my students have preferred various more primitive incarnations of this text to other texts; I hope you find this text worthwhile too.

Gary W. Oehlert

Draft of April 22, 2022

Chapter 1

Introduction

How do you answer these questions?

- Is a new drug a safe, effective treatment for a disease?

- How much buffer is needed around a GMO corn field to prevent the spread of GMO pollen to surrounding corn fields?

- How will the click rate change depending on the placement of an advertisement on a web page?

- Will an ice cream manufactured with a new kind of stabilizer be as palatable as our current ice cream?

- Does short-term incarceration of spouse abusers deter future assaults?

- What are the optimal conditions for operating a chemical refinery, given this month's grade of raw material?

Experiments collect the data that help answer questions like these, and this book is meant to help decision makers and researchers design good experiments, analyze them properly, and answer their questions.

Experiments answer questions

Consider the spousal assault example mentioned above. Justice officials need to know how they can reduce or delay the recurrence of spousal assault. They are investigating three different actions in response to spousal assaults. The assailant could be warned, sent to counseling but not booked on charges, or arrested for assault. Which of these actions works best? How can they compare the effects of the three actions?

This book deals with *comparative experiments*. We wish to compare some *treatments*. For the spousal assault example, the treatments are the three actions by the police. We compare treatments by using them and comparing the outcomes. Specifically, we apply the treatments to *experimental units* and then measure one or more *responses*. In our example, individuals who assault their spouses could be the experimental units, and the response could be whether or not assault recurs within one year. We compare treatments by comparing the responses obtained from the experimental units in the different

Treatments, experimental units, and responses

treatment groups. This could tell us if there are any differences in responses between the treatments, what the estimated sizes of those differences are, which treatment has the greatest reduction in one-year recurrence, and so on.

> An experiment is characterized by the treatments and experimental units to be used, the way treatments are assigned to units, and the responses that are measured.

1.1 Why Experiment?

Experiments help us answer questions, but there are also non-experimental techniques. What is so special about experiments? Consider that:

Advantages of experiments

1. We can design experiments to compare treatments directly.

2. We can design experiments to minimize any bias in the comparison.

3. We can design experiments so that the error in the comparison is small.

4. We can design experiments so that error is accurately estimated.

5. Most important, we are in control of experiments, in the sense of controlling the assignment of treatments to units (or units to treatments, if you prefer) and having that control allows us to make stronger inferences about the nature of differences that we see in the experiment. Specifically, we may make inferences about *causation*.

This last point distinguishes an experiment from an *observational study*. An observational study also has treatments, units, and responses. However, in the observational study we merely observe which units are in which treatment groups; we don't get to control that assignment.

Control versus observation

Example 1.1 Does spanking hurt?

Let's contrast an experiment with an observational study described in Straus, Sugarman, and Giles-Sims (1997). A large survey of women aged 14 to 21 years was begun in 1979; by 1988 these same women had 1239 children between the ages of 6 and 9 years. The women and children were interviewed and tested in 1988 and again in 1990. Two of the items measured were the level of antisocial behavior in the children and the frequency of spanking. Results showed that children who were spanked more frequently in 1988 showed larger increases in antisocial behavior in 1990 than those who were spanked less frequently. Does spanking cause antisocial behavior? Perhaps it does, but there are other possible explanations. Perhaps children who were becoming more troublesome in 1988 may have been spanked more frequently, while children who were becoming less troublesome may have been spanked less frequently in 1988.
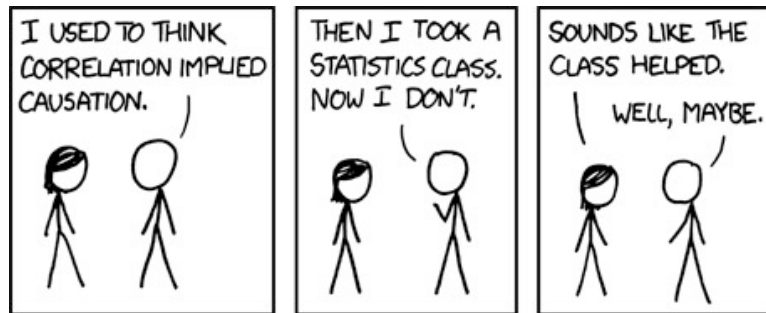
Figure 1.1: XKCD 552: Correlation. Accessed from `https://m.xkcd.com/552`, used under the Creative Commons license.

Example 1.2 Keep on smoking?

Freedman, Pisani, Purves, and Adhikari (1991) describe a large survey of households conducted by the Public Health Service. Men and women in those households were divided into groups by age and by whether they had never smoked, were current smokers, or had quit smoking. The nonsmokers were a little healthier than the smokers, but those who had quit smoking were much less healthy than the current smokers. Does this mean that stopping smoking makes you sick? No, there are several other potential explanations, the most likely of which is that many smokers who get very sick quit smoking, thus making the group of former smokers look less healthy than the smokers.

The drawback of observational studies is that the grouping into "treatments" is not under the control of the experimenter and its mechanism is usually unknown. Thus observed differences in responses between treatment groups could very well be due to other hidden mechanisms, rather than the treatments themselves. Observational studies can find correlation or association, but observational studies cannot, in and of themselves, find causation. See Figure 1.1 for an additional example.

It is important to say that while experiments have some advantages, observational studies are also useful and can produce important results. For example, studies of smoking and human health are observational, but the link that they have established is one of the most important public health issues in recent decades. Similarly, observational studies established an association between heart valve disease and the diet drug fen-phen that led to the withdrawal of the drugs fenfluramine and dexfenfluramine from the market (Connolloy et al. 1997 and US FDA 1997).

Observational studies are useful too

Mosteller and Tukey (1977) list three concepts associated with causation and state that at least two of the three are needed to support a causal relationship:

Causal relationships

- Consistency

- Responsiveness

- Mechanism.

Consistency means that, all other things being equal, the relationship between two variables is consistent across populations in direction and maybe in amount. Responsiveness means that we can go into a system, change the causal variable, and watch the response variable change accordingly. Mechanism means that we have a step-by-step mechanism leading from cause to effect.

In an experiment, we are in control, so we can achieve responsiveness. Thus, if we see a consistent difference in observed response between the various treatments, we can infer that the treatments caused the differences in response. We don't need to know the mechanism—we can demonstrate causation by experiment. (This is not to say that we shouldn't try to learn mechanisms—we should. It's just that we don't need mechanism to infer causation.)

Experiments can demonstrate consistency and responsiveness

We should note that there are times when experiments are not feasible, even when the knowledge gained would be extremely valuable. For example, we can't perform an experiment proving once and for all that smoking causes cancer in humans. We can observe that smoking is associated with cancer in humans; we have mechanisms for this and can thus infer causation. But we cannot demonstrate responsiveness, since that would involve making some people smoke, and making others not smoke. It is simply unethical.

Ethics constrain experimentation

Ethical issues in experimentation can be much more subtle than assigning people to smoke, and research institutions have review boards to ensure that experimentation maintains ethical standards. This involves many issues such as minimizing pain or trauma for experimental animals, ensuring that human subjects give informed consent to participate in experiments, setting up special safeguards when working with vulnerable populations (for example, children, the mentally ill, or trauma survivors), minimizing potential side effects (this could be drug side effects for humans or migration of genetically modified pollen into the wild), and so on. Although this book will not have much further discussion regarding ethics, ethics must be a consideration in the design of any experiment. Be sure to follow your local review board's standards and instructions.

Ethical issues may be subtle

1.2   Experimental Design

An experiment has treatments, experimental units, responses, and a method to assign treatments to units.

Treatments are the different procedures we want to compare. These could be different kinds or amounts of fertilizer in agronomy, different advertisement placement in web design, or different temperatures in a reactor vessel in chemical engineering.

Experimental units are the things to which we apply the treatments. These could be plots of land receiving fertilizer, different articles in an online newspaper, or batches of feedstock at a refinery.

Responses are outcomes that we observe after applying a treatment to an experimental unit. That is, the response is what we measure to judge what happened in the experiment; we often have more than one response. Responses for the above examples might be nitrogen content or biomass of corn plants, click-through rate for the advertisement, or yield and quality of the product per ton of raw material.

Randomization is the use of a known, understood probabilistic mechanism for the assignment of treatments to units. Other aspects of an experiment can also be randomized: for example, the order in which units are evaluated for their responses.

Together the treatments, experimental units, responses, and a method to assign treatments to units constitute the *experimental design*. Some authors make a distinction between the selection of treatments to be used, called "treatment design," and the selection of units and assignment of treatments, called "experiment design." We will not maintain that formal distinction.

Note that there is no mention of a method for analyzing the results in our definition of experimental design. Strictly speaking, the analysis is not part of the design, but a wise experimenter will always consider the analysis when planning an experiment. Analyzing experiments would be easy if there were no *experimental error*.

Analysis not part of design, but consider it during planning

Experimental Error is the random variation present in all experimental results. Different experimental units will give different responses to the same treatment, and it is often true that applying the same treatment over and over again to the same unit will result in different responses in different trials. Experimental error does not refer to conducting the wrong experiment or dropping test tubes.

Making sense of an experiment can be very difficult if there is *confounding*. Except in very special circumstances, confounding should be avoided, because no amount of fancy analysis will overcome confounding.

Confounding occurs when the effect of one factor or treatment cannot be distinguished from that of another factor or treatment. The two factors or treatments are said to be confounded. Consider planting corn variety A in Minnesota and corn variety B in Iowa. In this experiment, we cannot distinguish location effects from variety effects—the variety factor and the location factor are confounded.

Whereas the design determines the proper analysis to a great extent, we will see that two experiments with similar designs may be analyzed differently, and two experiments with different designs may be analyzed similarly. Proper analysis depends on the design and the kinds of statistical model assumptions we believe are correct and are willing to assume.

Not all experimental designs are created equal. A good experimental design must do the following:

Be accurate (avoid bias/systematic error/confounding) Comparative experiments look at differences in response between treatments. If our experiment has systematic error, then our comparisons will be biased, no matter how precise our measurements are or how many experimental units we use. For example, if responses for units receiving treatment one are measured with instrument A, and responses for treatment two are measured with instrument B, then we don't know if any observed differences are due to treatment effects or instrument miscalibrations. Randomization, as will be discussed more below, is our main tool to combat systematic error. Blinding (see below) can also be important.

Be precise (reduce variability) Even without systematic error, there will be random error in the responses, and this will lead to random error in the treatment comparisons. Experiments are precise when this random error in treatment comparisons is small. Precision depends on the size of the random errors in the responses, the number of units used, and the experimental design used. Several chapters of this book deal with designs to improve precision.

Allow estimation of error Experiments must be designed so that we have an estimate of the size of random error. This permits statistical inference: for example, confidence intervals or tests of significance. We cannot do inference without an estimate of error. Sadly, experiments that cannot estimate error continue to be run.

Have broad validity The conclusions we draw from an experiment are applicable to the experimental units we used in the experiment. If the units are actually a statistical sample from some population of units, then the conclusions are also valid for the population. Beyond this, we are extrapolating, and the extrapolation might or might not be successful. For example, suppose we compare two different drugs for treating attention deficit disorder. Our subjects are preadolescent boys from our clinic. We might have a fair case that our results would hold for preadolescent boys elsewhere, but even that might not be true if our clinic's population of subjects is unusual in some way. The results are even less valid for older boys or for girls. Thus if we wish to have wide validity—for example, broad age range and both genders—then our experimental units should reflect the population about which we wish to draw inference.

We need to realize that some compromise will probably be needed between these goals. For example, broadening the scope of validity by using a variety of experimental units may decrease the precision of our comparisons.

Compromise often needed

1.3   More About Randomization

We characterize an experiment by the treatments and experimental units to be used, the way we assign the treatments to units, and the responses we measure. An experiment is *randomized* if the method for assigning treatments to units involves a known, well-understood probabilistic scheme. The probabilistic scheme is called a *randomization*. As we will see, an experiment may have several randomized features in addition to the assignment of treatments to units. Randomization is one of the most important elements of a well-designed experiment.

Let's emphasize first the distinction between a random scheme and a "haphazard" scheme. Consider the following potential mechanisms for assigning treatments to experimental units. In all cases suppose that we have four treatments that need to be assigned to 16 units.

Randomization to assign treatment to units

Haphazard is not randomized

- We use sixteen identical slips of paper, four marked with A, four with B, and so on to D. We put the slips of paper into a basket and mix them thoroughly. For each unit, we draw a slip of paper from the basket and use the treatment marked on the slip.

- Treatment A is assigned to the first four units we happen to encounter, treatment B to the next four units, and so on.

- As each unit is encountered, we assign treatments A, B, C, and D based on whether the "seconds" reading on the clock is between 1 and 15, 16 and 30, 31 and 45, or 46 and 60.

The first method clearly uses a precisely-defined probabilistic method. We understand how this method makes it assignments, and we can use this method to obtain statistically equivalent randomizations in replications of the experiment.

The second two methods might be described as "haphazard;" they are not predictable and deterministic, but they do not use a randomization. It is difficult to mathematically model the mechanism that is being used. Assignment here depends on the order in which units are encountered, the elapsed time between encountering units, how the treatments were labeled A, B, C, and D, and potentially other factors. I might not be able to replicate your experiment, simply because I tend to encounter units in a different order, or I tend to work a little more slowly. The second two methods are not randomization.

> Haphazard is not randomized!

Introducing more randomness into an experiment may seem like a perverse thing to do. After all, we are always battling against random experimental error. However, random assignment of treatments to units has two useful consequences:

Two reasons for randomizing

1. Randomization protects against confounding.
2. Randomization can form the basis for inference.

We will discuss randomization for inference in the next chapter, but it is rarely used for inference in practice. However, the success of randomization in the protection against confounding is so overwhelming that randomization is almost universally recommended.

1.3.1   Randomization Against Confounding

We defined confounding as occurring when the effect of one factor or treatment cannot be distinguished from that of another factor or treatment. How does randomization help prevent confounding? Let's start by looking at the trouble that can happen when we don't randomize.

Consider a new drug treatment for coronary artery disease. We wish to compare this drug treatment with bypass surgery, which is costly and potentially dangerous. We have 100 patients in our pool of volunteers that have agreed via informed consent to participate in our study; they need to be assigned to the two treatments. We then measure five-year survival as a response.

What sort of trouble can happen if we fail to randomize? Bypass surgery is a major operation, and patients with severe disease might not be strong enough to survive the operation. It might thus be tempting to assign the stronger patients to surgery and the weaker patients to the drug therapy. This confounds strength of the patient with treatment differences. The drug therapy would likely have a lower survival rate because it is getting the weakest patients, even if the drug therapy is every bit as good as the surgery.

Failure to randomize can cause confounding

Alternatively, perhaps only small quantities of the drug are available early in the experiment, so that we assign more of the early patients to surgery, and more of the later patients to drug therapy. There will be a problem if the early patients are somehow different from the later patients. For example, the earlier patients might be from your own practice, and the later patients might be recruited from other doctors and hospitals. The patients could differ by age, socioeconomic status, and other factors that are known to be associated with survival.

There are several potential randomization schemes for this experiment; here are two:

- Toss a coin for every patient; heads—the patient gets the drug, tails—the patient gets surgery.

- Make up a basket with 50 red balls and 50 white balls well mixed together. Each patient gets a randomly drawn ball; red balls lead to surgery, white balls lead to drug therapy.

Note that for coin tossing the numbers of patients in the two treatment groups are random, while the numbers are fixed for the colored ball scheme. (Both of these designs are gross oversimplifications. A real experimental design would include considerations for age, gender, health status, and so on.)

Here is how randomization has helped us. No matter which features of the population of experimental units are associated with our response, our

randomizations put approximately half the patients with these features in each treatment group. Approximately half the men get the drug; approximately half the older patients get the drug; approximately half the stronger patients get the drug; and so on. The beauty of randomization is that it helps prevent confounding, even for factors that we do not know are important. These are not exactly 50/50 splits, but the deviation from an even split follows rules of probability that we can use when making inference about the treatments.

<div align="right">Randomization balances the population on average</div>

Here is another toy example of randomization. A company is evaluating two different accounting packages for use by its staff. Part of the evaluation is how quickly a set of transactions can be entered correctly using the two programs. We have 20 test accountant specialists, and each will enter the transactions twice, using each program once.

As expected, there are potential pitfalls in nonrandomized designs. Suppose that all account specialists did the evaluation in the order A first and B second. Does the second program have an advantage because the accountant will be familiar with the transactions and thus enter them more quickly? Or maybe the second program will be at a disadvantage because the accountants will be tired and thus slower.

Two randomized designs that could be considered are:

1. For each accountant, toss a coin: the accountant will use the programs in order AB or BA according to whether the coin is a head or a tail, respectively.

2. Choose 10 accountants at random for the AB order, the rest get the BA order.

Both these designs are randomized and will help guard against confounding, but the designs are slightly different and we will see that they should be analyzed differently.

<div align="right">Different randomizations are different designs</div>

Cochran and Cox (1957) draw the following analogy:

> Randomization is somewhat analogous to insurance, in that it is a precaution against disturbances that may or may not occur and that may or may not be serious if they do occur. It is generally advisable to take the trouble to randomize even when it is not expected that there will be any serious bias from failure to randomize. The experimenter is thus protected against unusual events that upset his expectations.

Randomization generally costs little in time and trouble, but it can save us from disaster.

1.3.2 Randomizing Other Things

We have taken a very simplistic view of experiments; "assign treatments to units and then measure responses" hides a multitude of potential steps and

choices that will need to be made. Many of these additional steps can be randomized, as they could also lead to confounding. For example:

- If the experimental units are not used simultaneously, you can randomize the order in which they are used.

- If the experimental units are not used at the same location, you can randomize the locations at which they are used.

- If you use more than one measuring instrument for determining response, you can randomize which units are measured on which instruments.

When we anticipate that one of these might cause a change in the response, we can often design that into the experiment (for example, by using blocking; see Chapter 12). Thus I try to design for the known problems, and randomize everything else. In sum,

> Randomize! Randomize! Randomize!

Example 1.3 One tale of woe

I once evaluated data from a study that was examining cadmium and other metal concentrations in soils around a commercial incinerator. The issue was whether the concentrations were higher in soils near the incinerator. They had eight sites selected (matched for soil type) around the incinerator, and took ten random soil samples at each site.

The samples were all sent to a commercial lab for analysis. The analysis was long and expensive, so they could only do about ten samples a day. Yes indeed, there was almost a perfect match of sites and analysis days. Several elements, including cadmium, were only present in trace concentrations, concentrations that were so low that instrument calibration, which was done daily, was crucial. When the data came back from the lab, we had a very good idea of the variability of their calibrations, and essentially no idea of how the sites differed.

The lab was informed that all the trace analyses, including cadmium, would be redone, all on one day, in a random order that we specified. Fortunately I was not a party to the question of who picked up the $75,000 tab for reanalysis.

1.3.3   Performing a Randomization

Once we decide to use randomization, there is still the problem of actually doing it. Randomizations usually consist of choosing a random order for a set of objects (for example, doing analyses in random order) or choosing random subsets of a set of objects (for example, choosing a subset of units for treatment A). Thus we need methods for putting objects into random orders and choosing random subsets. When the sample sizes for the subsets are fixed

Random orders and random subsets

and known (as they usually are), we will be able to choose random subsets by first choosing random orders.

Randomization methods can be either physical or numerical. Physical randomization is achieved via an actual physical act that is believed to produce random results with known properties. Examples of physical randomization are coin tosses, card draws from shuffled decks, rolls of a die, and tickets in a hat. I say "believed to produce random results with known properties" because cards can be poorly shuffled, tickets in the hat can be poorly mixed, and skilled magicians can toss coins that come up heads every time. Large scale embarrassments due to faulty physical randomization include poor mixing of Selective Service draft induction numbers during World War II (see Mosteller, Rourke, and Thomas 1970). It is important to make sure that any physical randomization that you use is done well.

Physical randomization

Physical generation of random orders is most easily done with cards or tickets in a hat. We must order N objects. We take N cards or tickets, numbered 1 through N, and mix them well. The first object is then given the number of the first card or ticket drawn, and so on. The objects are then sorted so that their assigned numbers are in increasing order; this puts the objects into random order. With good mixing, all orders of the objects are equally likely.

Physical random order

Once we have a random order, random subsets are easy. Suppose that the N objects are to be broken into g subsets with sizes n_1, …, n_g, with $n_1 + \cdots + n_g = N$. For example, eight students are to be grouped into one group of four and two groups of two. First arrange the objects in random order. Once the objects are in random order, assign the first n_1 objects to group one, the next n_2 objects to group two, and so on. If our eight students were randomly ordered 3, 1, 6, 8, 5, 7, 2, 4, then our three groups would be (3, 1, 6, 8), (5, 7), and (2, 4).

Physical random subsets from random orders

Numerical randomization uses numbers taken from a table of "random" numbers or generated by a "random" number generator in computer software. The word *random* is quoted because these numbers are not truly random. The numbers in the table are the same every time you read it; they don't change unpredictably when you open the book. The numbers produced by the software package are from an algorithm; if you know the algorithm you can predict the numbers perfectly. They are technically *pseudorandom* numbers; that is, numbers that possess many of the attributes of random numbers so that they appear to be random and can usually be used in place of random numbers, for example, for randomly assigning treatments to units.

Numerical randomization

Pseudorandom numbers

Example 1.4 Random sampling in **R**.

Random sampling is easy in **R**. In principle, we need to assign random numbers to very unit, and then arrange the units in a random order by putting their random numbers in increasing order. However, functions in **R** can hide that and move directly to the samples.

```
1 > sample(1:8,8)
 [1] 5 4 8 1 7 3 2 6
2 > subject.names <- c("Christine","Erica","Rebecca","Debbie","Laura","Karen",
 + "Sarah","Ellen","Alice","Amanda")
3 > sample(subject.names,10)
 [1] "Laura"     "Erica"     "Karen"     "Alice"     "Amanda"     "Rebecca"
 [7] "Ellen"     "Sarah" "Christine"    "Debbie"
4 > sample(c(1,1,1,1,2,2,3,3),8)
 [1] 2 1 1 3 1 2 1 3
```

In line 1, we ask for a sample of size 8 from the numbers 1 through 8 without
replacement. This produces the numbers arranged in random order. Lines 2
and 3 show that we can do the same thing with subject names instead of sub-
ject numbers, putting the names in random order. Finally, line 4 approaches
the issue differently. Suppose we want 4 units in treatment one, and 2 in
each of treatments two and three. Line 4 puts 4 ones, 2 twos, and 2 threes in
random order. In this case, subjects 2, 3, 5, and 7 would get treatment one.

1.4   More About Units

To this point, we have discussed experimental units as the items to which we
apply the treatments. However, there are also *measurement units*.

Measurement units (or response units) are the actual objects on which the
response is measured. These may differ from the experimental units.
For example, consider the effect of different fertilizers on the nitrogen
content of corn plants. Different field plots are the experimental units,
because we apply the fertilizers to the plots, but the measurement units
might be a subset of the corn plants on the field plot, or a sample of
leaves, stalks, and roots from the field plot.

A common source of difficulty is failing to recognize the distinction be-
tween experimental units and measurement units. Consider an educational
study, where six classrooms of 25 first graders each are assigned at random
to two different reading programs, with all the first graders evaluated via a
common reading exam at the end of the school year. Are there six experi-
mental units (the classrooms) or 150 (the students)?

Experimental and measurement units

One way to determine the experimental unit is via the consideration that
an experimental unit should be able to receive any treatment. Thus if students
were the experimental units, we could see more than one reading program in
each classroom. However, the nature of the experiment makes it clear that all
the students in the classroom receive the same program, so the classroom as
a whole is the experimental unit. We don't measure how a classroom reads,
though; we measure how students read. Thus students are the measurement
units for this experiment.

Experimental unit could get any treatment

There are many situations where a treatment is applied to group of ob-
jects, some of which are later measured for a response. For example,

- Fertilizer is applied to a plot of land containing corn plants, some of which will be harvested and measured. The plot is the experimental unit and the plants are the measurement units.

- Ingots of steel are given different heat treatments, and each ingot is punched in four locations to measure its hardness. Ingots are the experimental units and locations on the ingot are measurement units.

- Mice are caged together, with different cages receiving different nutritional supplements. The cage is the experimental unit, and the mice are the measurement units.

Treating measurement units as experimental units makes us think that we have more information that we actually have, and this generally leads to overly optimistic analysis. For example, we will reject null hypotheses more often than we should, and our confidence intervals will be too short and will not have their claimed coverage rates. The usual way around this is to determine a single response for each experimental unit. This single response is a summary of the responses in each measurement unit in the experimental unit, typically the average or total of the responses for the measurement units within an experimental unit, but the median, maximum, minimum, variance or some other summary statistic could also be appropriate depending on the goals of the experiment.

<div align="right">Use summary of measurement unit responses as experimental unit response</div>

A second issue with units is determining their "size" or "shape." For agricultural experiments, a unit is generally a plot of land, so size and shape have an obvious meaning. For an animal feeding study, size could be the number of animals per cage. For an ice cream formulation study, size could be the number of liters in a batch of ice cream. For a cloud computing configuration study, size could be the length of time the computer cluster is observed under load conditions.

<div align="right">Size of units</div>

Not all potential measurement units in an experimental unit will be equivalent. For the ice cream, samples taken near the edge of a carton (unit) may have more ice crystals than samples taken near the center. Thus it may make sense to plan the units so that the ratio of edge to center is similar to that in the product's intended packaging. Similarly, in agricultural trials, guard rows are often planted to reduce the effect of being on the edge of a plot. You don't want to construct plots that are all edge, and thus all guard row, so this constrains the size and shape of the experimental units. For experiments that occur over time, such as the computer network study, there may be a transient period at the beginning before the network moves to steady state. You don't want time units so short that all you ever measure is transient.

<div align="right">Edge may be different than center</div>

Financial resources are always limited, but one common situation is that is a limit on some other resource, such as a fixed area, a fixed amount of time, or a fixed number of measurements. This fixed resource needs to be divided into units and perhaps measurement units. How should the split be made? In general, more experimental units with fewer measurement units per experimental unit works better (see, for example, Fairfield Smith 1938). However, smaller experimental units are inclined to have greater edge effect problems

<div align="right">More experimental units, fewer measurement units usually better</div>

than are larger units, so this recommendation needs to be moderated by consideration of the actual units.

A third important issue is that the response of a given unit should not depend on or be influenced by other units, either the treatments given other units or the responses of other units. This is usually ensured through some kind of separation of the units, either in space or time. For example, a forestry experiment would provide separation between units, so that a fast-growing tree does not shade trees in adjacent units and thus make them grow more slowly; and a drug trial giving the same patient different drugs in sequence would include a washout period between treatments, so that a drug would be completely out of a patient's system before the next drug is administered.

When the response of a unit is influenced by the treatment given to other units, we get confounding between the treatments, because we cannot estimate treatment response differences unambiguously. In some cases, we can design around this problem. When the response of a unit is influenced by the response of another unit, we get a poor estimate of the precision of our experiment unless we modify our analysis to account for the correlation between responses. In particular, we usually overestimate the precision. Failure to achieve this independence can seriously affect the quality of any inferences we might make.

A final issue with units is determining how many units are required. We consider this in detail in Chapter 7.

1.5 More About Responses

We have been discussing "the" response, but it is a rare experiment that measures only a single response.

Primary responses Experiments often address several questions, and we may need a different response for each question. Responses such as these are often called *primary* responses, because they measure the quantity of primary interest for a unit.

Surrogate responses We cannot always measure the primary response. For example, a drug trial might be used to find drugs that increase life expectancy after initial heart attack: thus the primary response is years of life after heart attack. This response is not likely to be used, however, because it may be decades before the patients in the study die, and thus decades before the study is completed. For this reason, experimenters use *surrogate* responses. (It isn't only impatience; it becomes more and more difficult to keep in contact with subjects as time goes on.)

Surrogate responses are responses that are supposed to be related to— and predictive for—the primary response. For example, we might measure the fraction of patients still alive after five years, rather than wait for their actual lifespans. Or we might have an instrumental reading of ice crystals in ice cream, rather than use a human panel and get their subjective assessment of product graininess.

Surrogate responses are common, but not without risks. In particular, we may find that the surrogate response turns out not to be a good predictor of the primary response.

Predictive responses In addition to responses that relate directly to the questions of interest, some experiments collect *predictive* responses. We use predictive responses to model the primary response. The modeling is done for two reasons. First, such modeling can be used to increase the precision of the experiment and the comparisons of interest. In this case, we call the predictive responses *covariates* (see Chapter 17). Second, the predictive responses may help us understand the mechanism by which the treatment is affecting the primary response. Note, however, that since we observed the predictive responses rather than setting them experimentally, the mechanistic models built using predictive responses are observational.

Audit responses A final class of responses is *audit* responses. We use audit responses to ensure that treatments were applied as intended and to check that environmental conditions have not changed. Thus in a study looking at nitrogen fertilizers, we might measure soil nitrogen as a check on proper treatment application, and we might monitor soil moisture to check on the uniformity of our irrigation system.

Blinded responses Blinding occurs when the evaluators of a response do not know which treatment was given to which unit. Blinding helps prevent bias in the evaluation, even unconscious bias from well-intentioned evaluators. Double blinding occurs when both the evaluators of the response and the (human subject) experimental units do not know the assignment of treatments to units. Blinding the subjects can also prevent bias, because subject responses can change when subjects have expectations for certain treatments.

Example 1.5 Cardiac arrhythmias

Acute cardiac arrhythmias can cause death. Encainide and flecanide acetate are two drugs that were known to suppress acute cardiac arrhythmias and stabilize the heartbeat. Chronic arrhythmias are also associated with sudden death, so perhaps these drugs could also work for nonacute cases. The Cardiac Arrhythmia Suppression Trial (CAST) tested these two drugs and a placebo (CAST Investigators 1989). The real response of interest is survival, but regularity of the heartbeat was used as a surrogate response. Both of these drugs were shown to regularize the heartbeat better than the placebo did. Unfortunately, the real response of interest (survival) indicated that the regularized pulse was too often 0. These drugs did improve the surrogate response, but they were actually worse than placebo for the primary response of survival.

By the way, the investigators were originally criticized for including a placebo in this trial. After all, the drugs were *known* to work. It was only the placebo that allowed them to discover that these drugs should not be used for chronic arrhythmias.

1.6   More About Treatments

There are a couple special forms of treatments that deserve special mention.

Control Beyond the idea of a controlled experiment, we also can have *control* treatments. A control treatment is a "standard" treatment that is used as a baseline or basis of comparison for the other treatments. This control treatment might be the treatment in common use, or it might be a null treatment (no treatment at all). For example, a study of new pain killing drugs could use a standard pain killer as a control treatment, or a study on the efficacy of fertilizer could give some fields no fertilizer at all. This would control for average soil fertility or weather conditions.

In general, if you want to compare treatments to some kind of standard treatment, that standard/control treatment should be in the experiment. The only exception might occur when there is very strong prior knowledge that responses to the control treatment behave in a quantifiably consistent and predictable way. That prior knowledge is almost never available.

Placebo A *placebo* is a null treatment that is used when the act of applying a treatment—any treatment—has an effect. Placebos are often used with human subjects, because people often respond to any treatment: for example, reduction in headache pain when given a sugar pill. Blinding is important when placebos are used with human subjects. Placebos are also useful for nonhuman subjects. The apparatus for spraying a field with a pesticide may compact the soil affecting crop growth. Thus we drive the apparatus over the field, without actually spraying, as a placebo treatment.

Factors In many cases a treatment is actually the combination of two or more aspects. For example, the baking treatment for a cake involves a given time at a given temperature. The treatment is the combination of time and temperature, but we can vary the time and temperature separately. Thus we speak of a time factor and a temperature factor. Individual settings for each factor are called *levels* of the factor.

1.7   Problems

Suppose we are studying the effect of diet on height of children, and we have two diets to compare: diet A (a well balanced diet with lots of broccoli) and diet B (a diet rich in potato chips and candy bars). We wish to find the diet that helps children grow (in height) fastest. We have decided to use 20 children in the experiment, and we are contemplating the following methods for matching children with diets: **Problem 1.1**

1. Let them choose.
2. Take the first 10 for A, the second 10 for B.

3. Alternate A, B, A, B.

4. Toss a coin for each child in the study: heads → A, tails → B.

5. Get 20 children; choose 10 at random for A, the rest for B.

Describe the benefits and risks of using these five methods.

Problem 1.2

Human organs for transplantation have a very limited shelf life. The only seemingly viable method to extend that life is via cryopreservation, but this requires that the organ be frozen and thawed at appropriate rates. One problem with this approach is that the organ needs to be thawed uniformly enough that the early thawing tissues are not aging out of usefulness while other parts of the organ are still frozen, and it must be thawed slowly enough that the tissues are not damaged. Iron oxide nanoparticles (IONP) offer the possibility of uniformly thawing organ tissues at controllable rates, because they give off heat when placed in an alternating magnetic field (AMF). In principle, organs are placed in a solution of IONP so that the IONP are absorbed fairly uniformly into the tissue. The organ is frozen, and then thawed by putting it in an AMF.

One practical problem is that the IONPs will clump in the tissues, reducing the rate of heating. This experiment examines additives that are hoped to reduce the clumping and thus speed the thawing. Four organs are randomized to four different treatments, namely IONP dispersed in filtered water, filtered water and FBS, filtered water and PBS, and filtered water and agarose. Each organ was split into three samples. The twelve samples were then subjected (in random order) to AMF and the resulting specific absorption rate (rate of temperature change per gram of iron) was measured.

How many experimental units are there in this design? Explain your answer.

Problem 1.3

Time: the early 2000s. Place: Minnesota. The high school graduation requirements are widely reviled and are being replaced. The governor and the state House of Representatives have proposed one new set of standards, whereas the state Senate has proposed a different set of standards. Neither group wants to give in, so suppose that the governor proposes the following compromise.

Each school district in the state can choose between the two competing sets of standards, and students in those districts must meet the standards chosen by the district. In 10 years time, 3 complete cohorts of students will have moved from freshman year in high school through a nominal 4 years of college. The response for any given school district will be the percentage of students in those three cohorts who graduate from that district who also graduate from college before the end of the 10 year time limit. After the 10 years, the two sets of standards will be compared on this (and other) responses.

Comment on the design of the study; tell me what is good and what is bad. (You should ignore the political implausibility of this compromise and the

near certainty that the new rules would be changed multiple times in the next
10 years.)

Chapter 2

On Inference

Statistical inference is the process of taking the information that we have in data and making statements about the underlying processes that generated the data. For example, inferential statements could be estimates of means, variances, or other aspects of the underlying process, or they could be statements about evidence relating to certain hypotheses, such as two treatments producing the same mean response. In the inference step, we take the data we collected in our experiment and try to answer the questions that originally prompted us to run the experiment.

Inference is moving from data to answers

2.1 Schools of Inference

While it would be nice to have a one size fits all approach to inference, learners of statistics may be disappointed to discover that there are multiple ways to approach inference. Usually these multiple approaches lead to similar inferential results, but the philosophies behind the approaches, the difficulty of implementing the approaches, and the kinds of inferential statements that can be made differ.

Multiple approaches to inference

Example 2.1 Collar Runstitch Times

Bezjak and Knez (1995) provide data on the length of time it takes garment workers to runstitch a collar on a man's shirt, using a standard workplace and a more ergonomic workplace. Each worker sewed two sets of collars, one set with each system, with the order standard then ergonomic or the reverse determined by the toss of a coin. Table 2.1 gives the "auxiliary manual time" per collar in seconds for 30 workers using both systems. One question of interest is whether the times are the same on average for the two workplaces. Alternatively, one might wish to make an interval estimate of the difference in average runstitch times between the two workplaces. Employee

Table 2.1: Auxiliary manual times runstitching a collar for 30 workers under standard (S) and ergonomic (E) conditions.

#	S	E	#	S	E	#	S	E
1	4.90	3.87	11	4.70	4.25	21	5.06	5.54
2	4.50	4.54	12	4.77	5.57	22	4.44	5.52
3	4.86	4.60	13	4.75	4.36	23	4.46	5.03
4	5.57	5.27	14	4.60	4.35	24	5.43	4.33
5	4.62	5.59	15	5.06	4.88	25	4.83	4.56
6	4.65	4.61	16	5.51	4.56	26	5.05	5.50
7	4.62	5.19	17	4.66	4.84	27	5.78	5.16
8	6.39	4.64	18	4.95	4.24	28	5.10	4.89
9	4.36	4.35	19	4.75	4.33	29	4.68	4.89
10	4.91	4.49	20	4.67	4.24	30	6.06	5.24

Table 2.2: Differences in runstitching times (standard − ergonomic).

1.03	-.04	.26	.30	-.97	.04	-.57	1.75	.01	.42
.45	-.80	.39	.25	.18	.95	-.18	.71	.42	.43
-.48	-1.08	-.57	1.10	.27	-.45	.62	.21	-.21	.82

#1 (Mr. Skeptical) thinks the workplaces will make no difference in times. Employee #2 (Mr. Enthusiastic) thinks that the new environment will shave half a second off the stitching time.

These data are *paired*, because each worker was measured twice, once for each workplace, so the observations on the two workplaces are dependent. Fast workers are probably fast for both workplaces, and slow workers are slow for both. Because the mean of differences is the same as the difference of means, what we do is compute the difference (standard − ergonomic) for each worker, and work with the differences. This gets rid of much of the dependence in the data. Table 2.2 gives the differences between standard and ergonomic times.

2.1.1 Standard Frequentist Approach

Frequentist approaches to statistics are by far the most common, and this is what most people think of when they think of statistics.[1] They assume that unknowns (means, variances, regression coefficients, and so on) are fixed quantities and relate the observed data to these unknowns through a probability distribution for the data given the values of the parameters; this is called the *likelihood*. For the runstitching data we assume that the differences (standard − ergonomic) are independent from worker to worker, have mean μ and variance σ^2, and follow a normal distribution giving us a probability distribution:

Models often based on normal distribution

$$f(y_i; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

---

[1] Indeed, these were the only approaches mentioned in the first edition of this book.

for one data point or

$$f(y_1, \ldots, y_n; \mu, \sigma) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right]^n \exp \left[-\sum_{i=1}^n \frac{(y_i - \mu)^2}{2\sigma^2} \right] \qquad (2.1)$$

for a set of n independent data points. This is the simplest example of a standard linear model, of which we will see many generalizations later in this book. Procedures designed specifically for this set of assumptions, including t-tests and confidence intervals, F-tests, and so on are the basic, standard tools of statistical analysis.

Frequentists make inferential statements by comparing the results actually observed to other results that could have been observed when sampling data from the likelihood. This leads to confidence intervals for estimating parameters (in what fraction of repeated experiments would this procedure produce an interval that contains the parameter of interest), p-values for testing null hypotheses (in what fraction of repeated experiments when the null is true would we observe results this extreme or more extreme), and so on. Most uses of statistics involve frequentist methods, and many users of statistics are not even aware that other approaches are possible.

Confidence intervals and p-values

The null hypothesis of interest for Mr. Skeptical is that μ, the mean of the differences, is 0; Mr. Enthusiastic has a null hypothesis that the mean is .5. In fact, most of the time we will be joining Mr. Skeptical in assuming no effect as a null hypothesis. While one might hope that the ergonomic workplace shortened the time to complete a collar, it is best to check for changes in both directions. With these model assumptions, we would typically use a one-sample t-test on the differences (the same thing as a paired t-test). A t-based confidence interval is the standard approach for an interval estimate of μ.

Let $d_1, d_2, \ldots, d_n$ be the n differences in the sample (standard $-$ ergonomic in our example). Our null hypothesis is that the mean μ equals prespecified value μ_0 (H_0: $\mu = \mu_0$, here $\mu_0 = 0$ or $\mu_0 = .5$), and our alternative is H_1: $\mu \neq \mu_0$.

The formula for a one sample t-test is

One-sample t-test

$$t = \frac{\bar{d} - \mu_0}{s/\sqrt{n}} \quad ,$$

where $\bar{d}$ is the mean of the data (here the differences $d_1, d_2, \ldots, d_n$), n is the sample size, and s is the sample standard deviation (of the differences)

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2} \quad .$$

If our null hypothesis is correct and our assumptions are true, then the t-statistic follows a t-distribution with $n - 1$ degrees of freedom.

The p-value for a test is the probability, assuming that the null hypothesis is true, of observing a test statistic as extreme or more extreme than the one

The p-value

we did observe. "Extreme" means away from the null hypothesis toward the alternative hypothesis. Our alternative here is that the true average is different than the null hypothesis value, so larger or smaller values of the test statistic are extreme. Thus the p-value is the area under the t-curve (with $n - 1$ degrees of freedom) for the region that is at least as big in absolute value as the absolute value of the observed t.

A t-based confidence interval for the mean of the differences with coverage rate $1 - \mathcal{E}$ is formed via

t confidence interval

$$\bar{d} \pm t_{\mathcal{E}/2,n-1} \frac{s}{\sqrt{n}}$$

where $t_{\mathcal{E}/2,n-1}$ is the $\mathcal{E}/2$ percent point of a t distribution with n-1 degrees of freedom. There are no hypotheses associated with a t confidence interval, although there is a close association: the points in a $1 - \mathcal{E}$ confidence interval are the potential null hypothesis values that would have a p-value of more than $\mathcal{E}$.

Example 2.2 Standard freqentist analysis of runstitching time differences.

Even though t-tests and confidence intervals are simple to do by hand, we will still usually do them in **R**. See the *Companion* example Paired Procedures for an exposition on the commands needed.

We can test the null hypothesis that the mean difference is 0 (Mr. Skeptical) or the null hypothesis that the mean difference is .5 (Mr. Enthusiastic). The mean difference is positive (.175), but much closer to 0 than to .5. The t-statistic testing $\mu_0 = 0$ is only 1.49 corresponding to a p-value of .147 (.074 above the observed 1.49, plus .074 from below -1.49 for the two-sided alternative); the t-statistic for testing $\mu_0 = .5$ is -2.76 with a p-value of .01. There is effectively no evidence against $\mu_0 = 0$, and there is reasonable, but not overwhelming, evidence against $\mu_0 = .5$.

The data are much more in alignment with Mr. Skeptical than with Mr. Enthusiastic.

2.1.2 Likelihood Approach

The *Likelihood* approach forms a special subset of frequentist methods that is widely applicable with quasi-automatic methods of inference that generally work well for large sample sizes. Likelihood methods sometimes match standard frequentist methods, but often they are slightly different.

The likelihood is essentially the same thing as the probability function, except now we think of the data as being fixed and the distributional parameters as quantities that we can vary. Continuing the example in equation 2.1:

$$L(\mu, \sigma; y_1, \ldots, y_n) = \left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^n \exp\left[-\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

We usually work with the *log likelihood* $\ell = \ln(L)$:

$$\ell(\mu, \sigma; y_1, \ldots, y_n) = -\frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2} \qquad (2.2)$$

A maximum likelihood estimate (MLE) chooses parameters to be those that make the data most likely. (See, for example, Casella and Berger 2002 chapter 6.) In our example, we choose μ and σ to be those that maximize $\ell(\mu, \sigma)$. In simple situations, the MLE can often be written as a simple formula, but this is not true in many of the situations we will see. Maximum likelihood estimates will always lie in the domain of the unknown parameter. For example, variances will always be estimated to be nonnegative. Some standard frequentist methods for estimating variances in complex experimental designs can lead to negative estimates of variance. This embarrassment is one of the principal reasons we use MLEs in analyzing some experimental results.

Maximum likelihood estimate (MLE)

Under certain conditions (which do not always hold!), MLEs are approximately normally distributed for large enough sample sizes and have a variance that can be computed from the data. The most worrisome non-normal case for us is estimating a parameter on the boundary of its domain, especially estimating a variance, which might be estimated as 0, but cannot be less than 0.

The Likelihood ratio test statistic (LRT) is twice the difference between the log likelihood at the MLE and the log likelihood at the null hypothesis (see Figure 2.1). As the size of the data set increases, the distribution of the LRT under the null hypothesis approaches chi-square with degrees of freedom equal to the number of parameters being tested. (See, for example, Casella and Berger 2002 chapter 7.) Larger values of the LRT lead to smaller p-values. As with estimation, tests of parameter values at the boundary of the possible parameter values cause the chi-square approximation to fail. Likelihood confidence intervals can be constructed from likelihood ratio tests as the set of parameters for which the LRT does not reject the null hypothesis. These can be written out in a formula in simple cases, but, in general, these intervals need to be computed in software.

Likelihood ratio test (LRT)

Likelihood interval

This likelihood formulation is very general, but inferential statements from likelihood methods are typically only approximate, improving as the sample size grows. We will use likelihood methods primarily for estimating variance components in complex designs.

Example 2.3 Likelihood ratio tests for runstitching time differences.

We can fit null models and full models in **R**, and then extract the log likelihoods using the `logLik` (sic) function. See *Companion* section Likelihood and Predictive Procedures.

The log likelihoods for models with 0 and .5 means as well as the model with the estimated mean are:
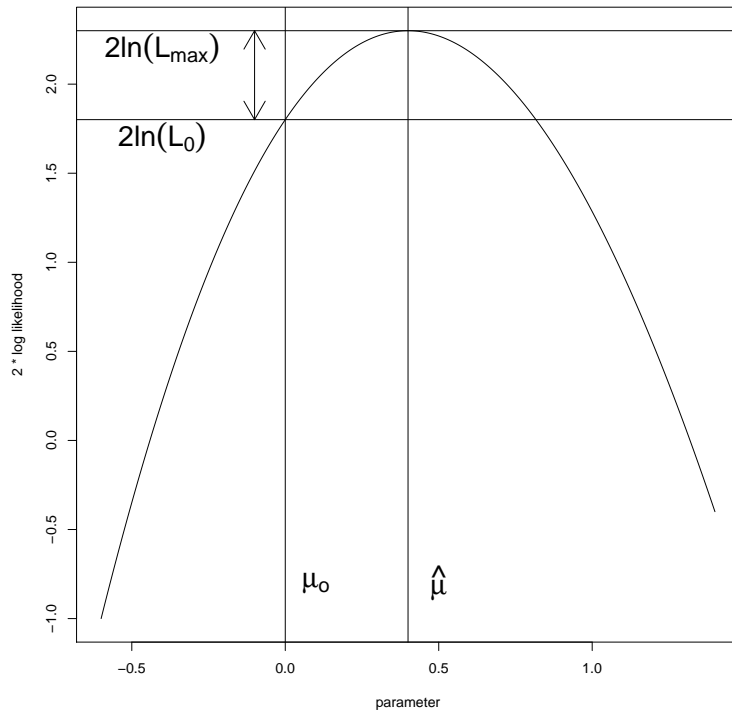
Figure 2.1: Two times a generic log likelihood curve with, the null value and MLE marked by vertical lines, and (twice) the log likelihood at the MLE and null shown by horizontal lines. The LRT is the length of the arrow.

Model	Log likelihood	LRT
$\mu = 0$	-29.99	$2(-28.88 - -29.99) = 2.21$
$\mu = .5$	-32.38	$2(-28.88 - -32.38) = 6.99$
μ estimated	-28.88	

The two null models fit one parameter each (the variance), and the unrestricted model fits two parameters (the mean and the variance). Thus the likelihood ratio tests should be compared to a chisquare distribution with $2 - 1 = 1$ degrees of freedom. The resulting p-values are .137 for the null of 0 and .008 for the null of .5. As we would hope, the p-values for these LR tests are very similar to those from the t-tests (ote that the square root of the LRT is approximately equal to the t-test).

Although we will typically not be obtaining the MLE or LRT by hand, it is worthwhile to look at our simple example in Equation 2.2 in a bit more

detail. First, it is clear that in order to get the maximum likelihood, we will choose $\hat{\mu}$ to make $\sum_i (y_i - \mu)^2$ as small as possible (the sum of squared differences enters with a negative coefficient, so minimizing the sum of squares maximizes its negative). This explains why "least squares" is such a prevalent technique.

Least squares

Second, the values of μ and σ^2 that maximize the likelihood are

$$\hat{\mu} = \bar{d} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (d_i - \bar{d})^2 = \frac{n-1}{n} s^2$$

Notice that the maximum likelihood estimate (MLE) for the variance is slightly smaller than the unbiased estimate s^2 used in the t-test (the MLE divides by n instead of $n-1$). Many people prefer unbiased estimates of variances, and this had led to modifications of maximum likelihood that provide unbiased estimates of variances. It is in this guise of *restricted* maximum likelihood (REML) that we will be using likelihood techniques.

n not $n-1$

REML

Third, a bit of algebra gives us that for our example

$$LRT = n \log(1 + t^2/(n-1))$$

showing the functional relationship between the t and the LRT in this problem: as the sample size gets bigger, the LRT will be closer and closer to t^2.

LRT related to t^2

2.1.3 Predictive Model Selection

Predictive model selection says that the best model for data is the one that would best predict future data from the same data generating system. Such model selection does not involve testing hypotheses about parameters in models. It is instead concerned with figuring out how well a model predicts future data based only on the data at hand. The overriding problem is that a model will not predict future data as well as it predicts the data on which it is fit; it will be too optimistic. Thus the task is to use the current data for both a model fit and an adjustment for the unfounded optimism.

Seek best prediction

Cross-validation is a tool that directly attacks the problem of using the same data for fitting a model and assessing its predictive performance. As an example, 10-fold cross-validation divides the data into 10 randomly chosen subsets. It then fits the model ten times, each time leaving out one of the subsets of data. Predictive accuracy is judged by predicting the values of the left out subset based on the model fit to the other nine subsets. N-fold cross-validation is similar, except it leaves out one data point at a time, and predicts with the model fit to the other $N-1$ data points. The "quality of the prediction" is sometimes fairly simple to define (for example, squared residuals in regression-like situations), but it not always obvious. Cross-validation works fairly well, but it involves a lot of computation. This has led researchers to consider other approaches.

Cross-validation

One generic approach to assessing model predictions would be to look at the expected log likelihood of future data, conditional on the parameters. Of course, we usually don't know the parameters, so we substitute the MLE for the unknown parameters. We also don't have future data, so we just evaluate the criterion at the data that we do have. This is simply L_{max}, the same maximized log likelihood that we used in the LRT. Fortunately, we can calculate the expected amount by which L_{max} over-estimates the log likelihood of future data; it is simply p, the number of parameters in the model. Thus the maximized log likelihood on the current data less the number of parameters in the model is an unbiased estimate of the log likelihood of future data based on the same model.

For historical reasons, the difference of L_{max} and p is multiplied by –2, obtaining AIC, the Akaike Information Criterion. There is also a version of AIC that works a little better in small samples called corrected AIC, or AICc. Suppose we have several models to compare based on a data set of size n. Let L_{max}^k be the maximized likelihood for model k; this is the likelihood evaluated at the maximum likelihood estimates of the parameters in the kth model. Let p_k be the number of parameters that we fit for model k. Then the Akaike Information Criterion is

AIC and AICc

$$AIC = 2p_k - 2\log(L_{max}^k),$$

and the AIC corrected for small sample size is

$$AICc = 2p_k\frac{n}{n - p_k - 1} - 2\log(L_{max}^k) \quad.$$

We want models with small values of AIC (AICc). A large likelihood makes these criteria small as does a small number of parameters. The criteria try to balance between fitting well and using too many parameters.

Minimize AIC

In information theory, if two models have AIC values that differ by δ, then the model with the lower AIC is $\exp(|\delta|/2)$ times as likely as the model with the higher AIC to be the model that minimizes information loss (does the best prediction). Thus an AIC difference of 2 gives approximately 3:1 odds in favor of the model with the smaller AIC.

Model odds

Example 2.4 Comparing models for runstitching time differences using information criteria.

We fit the full and null (both 0 and .5 nulls) models for the runstitch data and compute the `AICc` criterion (see *Companion* section Likelihood and Predictive Procedures):

Model	AICc
$\mu = 0$	62.12
$\mu = .5$	66.90
Fit μ	62.21

AICc ever so slightly prefers the null model with mean of zero, although the difference between this and the model with estimated μ is not material. The model with mean .5 should probably not be considered further.

For two models that differ by just one parameter, AIC will select the larger model if the LRT is 2 or greater. Using the chi-square approximation to the LRT, AIC is adding an additional variable when the p-value for the LRT is .157 or smaller, so AIC is not a stringent filter for adding single variables. However, AIC becomes more stringent as you consider adding more and more variables at once, being roughly equivalent to testing at the .05 level for 7 variables at a time, and roughly equivalent to testing at the .01 level for 16 variables at a time.

AIC differs from testing parameters

2.1.4 Subsampling Approach

Subsampling methods, including randomization tests and bootstrap methods, do not use probability functions or likelihood functions but instead use some kind of subsampling from the original data to generate reference distributions for inference. Randomization tests are rarely used in practice, but they have the advantage that basically their only assumption is that a randomization was performed in setting up the experiment. This makes them useful in legal settings, or other settings where the assumptions of the inference may be subject to dispute. We will use Bootstrap procedures in some cases where the theory for a test or confidence interval is difficult to work out.

Nearly all the analysis that we will do in this book will be parametric, with nearly all of it making assumptions like "The responses in treatment group A are independent from unit to unit and follow a normal distribution with mean μ and variance σ^2." Nowhere in the design of our experiment did we do anything to make this so; all we did was randomize treatments to units and observe responses.

In fact, randomization itself can be used as a basis for inference. The advantage of this randomization approach is that it relies only on the randomization that we performed. It does not need independence, normality, and the other assumptions that go with linear models. The disadvantage of the randomization approach is that it can be tedious to implement, even in relatively small problems, though computers make it much easier. Furthermore, the inference that randomization provides is often indistinguishable from that of standard techniques such as t-tests or F-tests.

Randomization inference makes few assumptions

Randomization tests look at the world "backwards" from the parametric procedures we used in frequentist and likelihood methods. In frequentist methods, we know what groups or treatments the data belong to; what is random is the value of the response. In randomization methods, we know the responses that we observed; what is random is the groups or treatments that the data belong to. The randomization null hypothesis is that the labelling as to groups makes no difference on the response; it's just a label. The variability in any test statistic is simply generated by the random assignment of treatments to units.
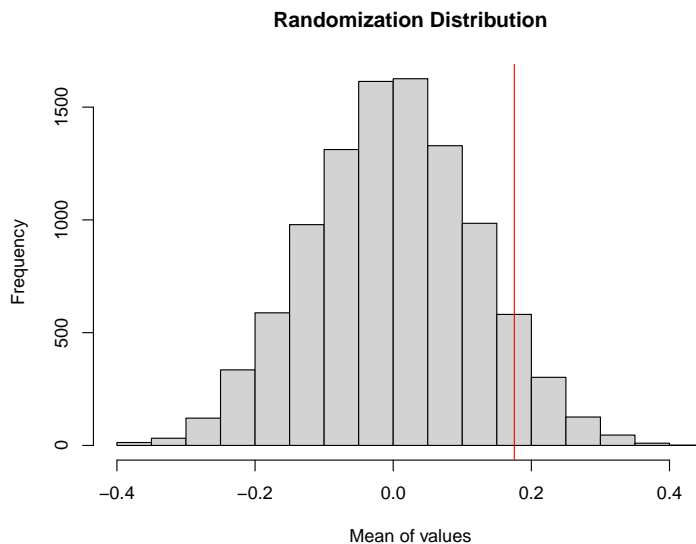
Figure 2.2: Histogram of 10,000 samples from the randomization distribution of the sum of the differences for runstitching, with vertical line added at the observed sum.

To construct a randomization test, we choose a descriptive statistic for the data and then get the distribution of that statistic under the randomization null hypothesis. The randomization p-value is the probability (under this randomization distribution) of getting a descriptive statistic as extreme or more extreme than the one we observed.

Example 2.5 A randomization alternative to the paired t-test for runstitching time differences.

The randomization null hypothesis is that the two workplaces are completely equivalent; we would have observed the responses we did observe regardless of the treatments, and the treatments merely act to label the responses. For example, the first worker had responses of 4.90 and 3.87, which we have labeled as standard and ergonomic. Under the randomization null, the responses would be 4.90 and 3.87 no matter how the random assignment of treatments turned out. The only thing that could change is which of the two is labeled as standard, and which as ergonomic. Thus, under the randomization null hypothesis, we could, with equal probability, have observed 3.87 for standard and 4.90 for ergonomic.

What does this mean in terms of the differences? We observed a difference of 1.03 for worker 1. Under the randomization null, we could with equal probability have observed the difference −1.03, and similarly for all the

other differences. Thus in the randomization analogue to a paired t-test, the absolute values of the differences are taken to be fixed, and the signs of the differences are random (because of the randomization), with each sign independent of the others and having equal probability of positive and negative.

For this problem, we take the sum of the differences as our descriptive statistic. The average would lead to exactly the same p-values, and we could also form tests using the median or other measures of center. With 30 workers, there are $2^{30} = 1,073,741,824$ different ways that the random assignment of signs could turn out. In principle, we need to evaluate all billion plus possibilities to compute the p-value for our test (please recall that I did describe this approach as tedious). In practice, we take a random sample of possibilities from the total list of possibilities and determine the p-value from this random subsample. See *Companion* section Paired Procedures.

Figure 2.2 shows a histogram of the randomization distribution for 10,000 random sample configurations of signs. The observed value of .175 is not far into the tail of the distribution, and the (two-sided) randomization p-value is .145. The similarity of the randomization p-value and the typical t-test p-value is typical, and they get closer as sample sizes increase. Keep in mind that we sampled our reference distribution, so if we do the test another time we will get a slightly different sample from the reference distribution and a slightly different p-value.

We can also construct a confidence interval for the mean difference. How much larger could the mean of differences be before it becomes significant (in a two-sided 5% test) in the randomization distribution? Just .061. How much smaller could it be before it becomes significant in the randomization distribution? About .411. Thus the randomization confidence interval of the mean of the differences is $(-.061, .411)$.

Randomization interval estimate

2.1.5   Bayesian Approach

Bayesian methods assume that all unknowns are random variables that follow probability distributions. In addition to the likelihood for the data given the unknowns (shared by frequentists and Bayesians), Bayesians express their prior beliefs about the unknown parameters via a *prior probability distribution*. After observing data, Bayesians update probability distributions for the unknowns using Bayes rule (whence Bayesian statistics) to obtain the *posterior distribution*, and construct inference based on the posterior distribution of the unknowns. Prior and posterior in this context mean before and after seeing the data.

Prior and posterior

> A bit more mathematically, we have:
>
> $f_{like}(y|\theta)$ a likelihood for the data y given the parameters θ;
>
> $f_{prior}(\theta)$ a prior distribution for the parameters; and
>
> $f_{post}(\theta|y)$ the posterior for the parameters given the data.

Draft of Feb 25, 2021

They combine via Bayes rule

$$f_{post}(\theta|y) = \frac{f_{like}(y|\theta)f_{prior}(\theta)}{f(y)} \tag{2.3}$$

$$= \frac{f_{like}(y|\theta)f_{prior}(\theta)}{\int f_{like}(y|\theta)f_{prior}(\theta)d\theta} \tag{2.4}$$

Frequentist inferential statements are somewhat awkward. For example, for a 95% confidence interval for a mean, the statement says that in a long run of identical experiments, the procedure that generated the confidence interval will produce intervals containing the mean in 95% of all repetitions. It says nothing in particular about this particular repetition that we have observed. In contrast, Bayesian inferential statements are generally the kind of statements we would like to make. For example, for a 95% Bayesian credible interval, the statement says that the unknown parameter is in the computed interval with probability 95%.

Bayesians make probability statements

One of the challenges with Bayesian analysis is that the denominator in Equation 2.3 is generally very difficult to compute. Instead, statisticians use approximate samples from the posterior to do inference, with these samples coming from a technique called *Markov Chain Monte Carlo* (MCMC). To get the posterior mean, we take the mean of the MCMC samples. To get a posterior probability interval, we take the corresponding quantiles of the MCMC samples. MCMC algorithms allow us to make progress using only the numerator in Equation 2.3 without ever needing to evaluate the denominator.

MCMC

There are several methods available for comparing Bayesian models, but they fall into two groups: the Bayes factor and various predictive measures such as WAIC (Widely Applicable Information Criterion) and LOOCV (Leave One Out Cross-Validation). Both WAIC and LOOCV are more computationally intensive than AIC, but it turns out there is a trick that allows us to estimate LOOCV without actually doing multiple fits of the data (Vehtari, Gelman, and Gabry 2017). WAIC and LOOCV estimate the same quantity and will typically be nearly equal in well-behaved cases. We will use LOOCV and the Bayes factor with the important caveat that the Bayes factor is *highly* dependent on how the prior is specified.

LOOCV

In the Bayesian context, we can construct a "posterior predictive" distribution. This is the probability distribution of a future data point given the data we have seen and the priors we are using. In formulae, the posterior predictive distribution $f_{postpred}$ is

$$f_{postpred}(y_{new}|y) = \int f_{like}(y_{new}|\theta)f_{post}(\theta|y)\,d\theta$$

LOOCV estimates N times the expected value of $\log f_{postpred}(y_{new}|y)$.

Prior distributions can keep parameters from varying freely. If a parameter is not completely free to adapt to the data, it does not contribute as much to the model fit. In this way, the effective number of parameters in a Bayesian model can differ from the evident/explicit number of parameters. WAIC estimates and uses an effective number of parameters; LOOCV does not explicitly use an effective number of parameters, but such a value can be derived.

Effective number of parameters

The second Bayesian approach to comparing models is to compute the *Bayes Factor*. The Bayes factor gives the evidence, based solely on the data, for preferring one model relative to another model; the Bayes factor does not take into account any prior information we might have about which *model* is more likely (it does take priors on the parameters into account).

Compare models via Bayes factor

The Bayes factor for model 1 relative to model 2 is denoted BF_{12}. If $BF_{12} > 1$, then the data favor model 1 over model 2.

> Technically, the Bayes factor for model 1 relative to model 2 is the (marginal) probability of the data under model 1 divided by the (marginal) probability of the data under model 2. Each of these probabilities is defined as the likelihood of the data given the parameters times the prior probability of the parameters, that product then integrated (averaged) across all possible values of the parameters. Somewhat more mathematically, each of these probabilities takes the form
>
> $$f(y) = \int f_{like}(y|\theta) f_{prior}(\theta) \, d\theta$$
>
> The discerning reader will recognize this quantity as the denominator of Equation 2.3, which value we just said was very difficult to compute. In practice, we will also need tools to approximate the Bayes factor.

There is a relationship between between a Bayes factor and the analogous LRT. The LRT takes the ratio of the peak or maximum values of the likelihoods under the two models whereas the Bayes factor takes the ratio of the average values of the likelihood, with each average weighted according to the corresponding prior distribution for the unknown parameters.

Bayes factor and LRT

Here are some useful facts about Bayes factors:

- The Bayes factor internally accounts for the number of parameters, so no further parameter count adjustment is needed.

- $BF_{12} = 1/BF_{21}$, so the Bayes factor can give evidence in favor of either model.

- If model 1 is nested in model 2 and model 1 is correct, then BF_{12} will go to infinity as the sample size increases. (Note that this kind of thing does not happen with frequentist tests; model 1 will still be rejected at rate $\mathcal{E}$ even for very large sample sizes.)

- If we have BF_{12} and BF_{32}, then $BF_{13} = BF_{12}/BF_{32}$. This can be be useful if we have Bayes factors for multiple models compared to a single model.

- Bayes factors are much more sensitive to how the prior distributions were specified than are the estimates of parameters or LOOCV.

- While it is easy to define these marginal probabilities, they can be difficult to compute due to the integration/averaging.

The Bayes factor can be combined with prior probabilities of models to get the (posterior) probability of models given the data:

$$\frac{P(\text{Model 1 given data})}{P(\text{Model 2 given data})} = BF_{12} \, \frac{P(\text{Model 1})}{P(\text{Model 2})} \qquad (2.5)$$

where the probabilities in the rightmost fraction are prior probabilities for the models. Posterior odds for models are thus a combination of the evidence in the data (via the Bayes factor) and the prior odds for the models. To make an extraordinary claim (select model 1 when the prior probability for model 1 is much less than model 2), you need to have extraordinary evidence (BF_{12} very large). See Figure 2.3. If our prior probabilities for all models are the same, then the Bayes factor tells us the relative probability of each model to other models.

Model comparison with priors on models

One can do a formal Bayesian model selection using *decision theory*, or one can do an informal Bayesian model selection using the Bayes factor alone. The formal approach leads to a decision regarding two models, taking into account the prior probabilities for the models as well as the losses one might incur from making the wrong decision. The informal approach looks only at what the data have to say about the odds of the two models given the data (that is, the Bayes factor).

Informally, if BF_{12}, the Bayes Factor for model 1 relative to model 2, is greater than 1, then the data say that model 1 is the preferred model. The greater BF_{12}, the more strongly the data argue for model 1 over model 2. There is no hard and fast rule for interpreting Bayes Factors, but Table 2.3 gives one scale for interpreting them. If $BF_{12} < 1$, then you can either flip the models to obtain BF_{21} (which will be greater than 1) and interpret using Table 2.3 or you can take the reciprocals of the ranges and say, for example, that a Bayes factor between 1 and 1/3 is not worth more than a bare mention.

Scale for Bayes factor

Even though $BF_{12} > 1$ argues for model 1, most researchers do not feel that this is sufficient evidence to make a model selection statement with confidence. Thus you will sometimes see an informal policy that says to select model 1 if $BF_{12} > K$, select model 2 if $BF_{12} < 1/K$, and remain undecided if the Bayes factor lies between the two cutoffs. Typical values of K might be 3, 5, or even 10.

> The more formal approach to choosing between models requires that the researcher provide additional information. This includes the prior probability that model 1 is the true model (P_1, so that $P_2 = 1 - P_1$ is
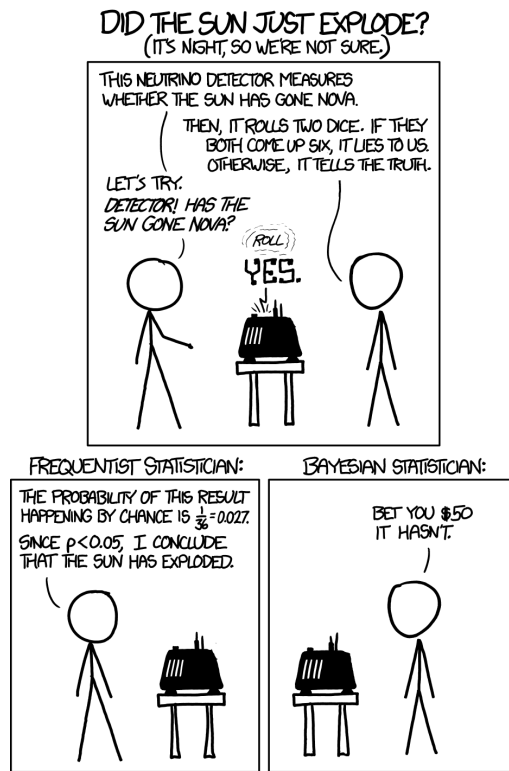
Figure 2.3: XKCD 1132: Frequentists vs. Bayesians. Accessed from `m.xkcd.com/1132`, used under the Creative Commons license.

Table 2.3: Kass and Raftery (1995) guidelines for interpreting the Bayes factor of model 1 relative to model 2.

BF_{12}	Evidence against model 2
1 to 3	Not worth more than a bare mention
3 to 20	Postive
20 to 150	Strong
>150	Very strong

the prior probability that model 2 is correct), the loss (cost) for choosing model 2 when model 1 is correct (C_2), and the loss (cost) for choosing model 1 when model 2 is correct (C_1). Using equation 2.5, we can

rearrange it to get the posterior probability of model 1 given the data:

$$P(\text{Model 1 given data}) = \frac{BF_{12}\frac{P_1}{P_2}}{1 + BF_{12}\frac{P_1}{P_2}}$$

If you choose model 1, the average cost will be $P(\text{Model 2 given data})C_2$; if you choose model 2, the average cost will be $P(\text{Model 1 given data})C_1$.

The decision theory approach says to make the decision with lowest average cost, so choose model 1 if

$$\text{Average loss if choosing model 2} \quad > \quad \text{Average loss if choosing model 1}$$
$$P(\text{Model 1 given data})C_1 \quad > \quad P(\text{Model 2 given data})C_2$$
$$\frac{BF_{12}\frac{P_1}{P_2}}{1 + BF_{12}\frac{P_1}{P_2}}C_1 \quad > \quad \frac{1}{1 + BF_{12}\frac{P_1}{P_2}}C_2$$
$$BF_{12} \quad > \quad \frac{C_2}{C_1}\frac{P_2}{P_1}$$

Thus the formal Bayesian model choice is also based on the Bayes factor, but the cutoff for which model is chosen is adjusted to account for prior probabilities of models and the costs of choosing the wrong model. The formal approach becomes the informal approach if we assume the ratio on the right is 1 (for example, equal prior probabilities and equal costs of mis-selection).

Decision theoretic model choice minimizes expected cost

Bayesian model selection is about minimizing the expected loss when choosing a model. It does not reject or fail to reject hypotheses. It does not privilege one model over the other (as opposed to frequentist methods, which privilege the null model). It makes no claims about the rate at which it falsely selects certain models (that is, it has no analogue of the p-value). In fact, if the cost ratio is sufficiently far from 1, the Bayesian approach can select a model that is not well supported by the data, simply to hedge against a large potential cost.

Example 2.6 A Bayesian analysis for the runstitching time differences.

Please see the *Companion* section One-Sample Bayesian Procedures for the computational aspects of this example.

We will use the same likelihood that we used in standard frequentist and likelihood methods, but we must also specify prior distributions. Assume that the prior distribution for μ is normal with mean 0 and standard deviation .5 (in the *Companion* computations, we only assume that the prior standard deviation is very close to .5, but not exactly equal to .5). This reasonably sums up our prior belief that the different environments could have somewhat different means, but probably not vastly different means. We will assume that our prior belief about σ is summarized as σ could be about 1, but we are 99%

Table 2.4: Bayesian estimation and model comparison results for the runstitching experiment. Bayes factors computed as first model (with point prior at 0) relative to other model.

μ Prior Mean	μ Prior SD	σ Prior Range	μ Post. Interval	LOOCV	Bayes Factor
.0	.0	(.02,4.3)	(.0,.0)	62.10	1.00
.0	.1	(.02,4.3)	(−.08,.22)	61.55	0.83
.0	.5	(.02,4.3)	(−.06,.39)	61.78	1.54
.0	.5	(.26,2.4)	(−.08,.40)	61.67	1.04
.0	.8	(.02,4.3)	(−.07,.41)	61.90	2.31
.5	.5	(.02,4.3)	(−.05,.43)	61.98	1.76
.5	.1	(.02,4.3)	(.21,.54)	63.33	3.69
.5	.0	(.02,4.3)	(.5,.5)	66.69	10.46

sure it is between .02 and 4.3; this is a much broader prior spread on σ than is likely in the data.

On the basis of MCMC samples, we can say that the probability that μ is in the interval (−.06, .39) is 95%. Note that this is actually a probability statement, not a confidence statement. Furthermore, we estimate the posterior mean for μ to be .164. Notice that the posterior mean is between the mean of the prior and the mean of the data. The Bayesian analysis compromises between the information in the prior and the information in the data.

Suppose now that Mr. Enthusiastic complains that the prior distribution for μ should be normal with mean .5 and standard deviation .5. He is a Bayesian and is allowed to say that is his prior belief. We can refit with this prior, and we get a posterior mean for μ of .193 with a 95% credible interval of (−.048, .427). We can see that this larger prior mean has pulled the posterior mean up a little bit. Suppose instead that Mr. Enthusiastic is not only enthusiastic but also more certain in his prior beliefs; in that case he might use a prior for μ that is normal with mean .5 and standard deviation .1. Implementing this model yields a posterior mean estimate of .374 and a 95% credible interval of (.214, .540). In the limit, someone might express complete certainty about the mean. He or she might wish to fit with a prior for μ that has mean 0 and standard deviation 0, or perhaps mean .5 and standard deviation 0. No amount of data is going to overcome a prior with standard deviation 0, so the posterior estimates for μ are *exactly* 0 and .5 with no uncertainty.

Table 2.4 shows the posterior credible intervals for μ and LOOCV values for eight different priors. Five of the priors give very similar intervals for μ, with the prior expectation of .5 and standard deviation of .1 shifting the interval up noticeably. Six of the priors have sufficiently similar LOOCV values as to be effectively interchangeable. The priors centered at .5 with small standard deviation have higher LOOCV and would not be considered (sorry, Mr. Enthusiastic).

Table 2.4 also shows the Bayes factors for model one (with a point prior

at 0) relative to the other seven models. Values greater than one favor the point prior at 0, values less than one favor the other prior. Only the model with prior mean 0 standard deviation .1 for μ is favored to the point prior at zero, and that only just slightly. The point prior at 0 is strongly preferred over the point prior at .5. This is fully in agreement with LOOCV.

Look now at the Bayes factors for the two models that use a prior for μ with mean 0 and standard deviation .5. These two models differ by the prior that is placed on the error variability. For these models, the credible intervals for the error standard deviation are (.51, .87) and (.54, .89), so either prior is consistent with the data. However, the Bayes factors for these two priors differ by a factor of 1.5. This is an example of how the Bayes factor can be much more sensitive to the choice of prior than are the estimates or the LOOCV. Thus while there are strong philosophical justifications for using the Bayes factor as a measure of model preference based on the data, the Bayes factor is only as good as your belief in the priors that go into the models. Priors yielding roughly equivalent estimates can yield radically different Bayes factors.

One final approach to mention is the *Region Of Practical Equivalence*, or ROPE. This approach is appropriate in the common situation where one of the models (say model 1) being compared is simply a version of the other model with the parameters fixed at certain values. For example, model 2 might have a prior for μ that is N(0,.5), and model 1 might specify that $\mu = 0$. Suppose that in our heart of hearts we believe that the parameter values cannot be *exactly* as specified in model 1. They might be very close, but they simply cannot ever match the specification of model 1 exactly. On the other hand, we might be able to say that there is a region (interval for a single parameter) around the model 1 values that is for all practical purposes equivalent to model 1. For example, perhaps any μ between −.01 and .01 is close enough to 0 to ignore the differences. This is our ROPE.

ROPE: Region of Practical Equivalence

The ROPE approach says to get the probability of the ROPE under the posterior distribution: P_{ROPE}. For a suitable $\mathcal{E}$, select model 1 if $P_{ROPE} > 1 - \mathcal{E}$, select model 2 if $P_{ROPE} < \mathcal{E}$, and remain undecided if $\mathcal{E} < P_{ROPE} < 1 - \mathcal{E}$.

2.1.6   Wrap up

All of the inferential approaches we examined led to similar conclusions in our simple example. This will usually be the case, with most differences occurring in small data sets (or with very concentrated priors). So why are there so many different ways to do inference? Because every method has its weaknesses.

Strengths and weaknesses

The t-confidence intervals and F tests of the standard frequentist methods are widely understood and accepted, "exact" in many situations, broadly developed, and generally easy to compute. However, they work better in what we will call *fixed* effects analysis and start running into problems in complex *random* effects models. They do not provide probability statements as part of inference.

Likelihood methods solve some of the problems that classical methods experience in random effects, for example, you cannot get negative estimates of variances with likelihood methods. However, their standard methods for tests and confidence intervals are not applicable when testing whether variances are zero, which will be an area of key interest.

Randomization methods are fine in simple problems and are nearly bulletproof from an applicability/are-the-assumptions-correct point of view. But they do not generalize easily to more complicated designs.

Bayesian methods allow us to make the kind of inferential statements that we would like to make (for example, probability rather than confidence), but these statements are based on our subjectively-chosen prior. They are never easy to implement, but they do not get much more difficult (at least in principle) as designs and models get more complicated. In fact, you can do Bayesian analysis in extremely complex models.

Why don't we all use Bayesian procedures all the time? The two main reasons are subjectivity and difficulty. A Bayesian must specify a prior distribution for the unknowns in the model. My prior might not be the same as your prior, and as we saw with Mr. Skeptical and Mr. Enthusiastic, that means that my inference will not be the same as your inference. Many researchers find that deeply troubling. However, *everyone* uses prior information to *design* experiments; any experiment conducted without reference to the prior information held by the experimenter and the literature is a poor experiment indeed. In addition, the effect of the prior diminishes as the amount of data increases. Conversely, the so-called objective methods used by non-Bayesians are not nearly as objective in practice as one would wish them to be, as there are many different ways that the objective methods can be subjectively selected and used (Gelman and Loken 2014). Bayes methods are becoming more prevalent in applied work, but they remain an exception when analyzing designed experiments.

Some dislike subjectivity

Difficulty takes several forms. A Bayesian must elicit the prior distribution; doing this well is not straightforward. The prior can be specified in a routine/mechanistic fashion, and there has been a lot of work over the years on determining so-called objective priors. However, the fact remains that one only gets the full benefit of the Bayesian perspective if one has a genuine prior. And, except in the simplest of toy problems, Bayesian solutions are challenging to compute. It is only in the last two or three decades, when computers have become fast enough and algorithms have become clever enough, that one can approximate the Bayesian solution in realistic models. It is this evolution of our computing capabilities that has brought Bayesian statistics out of the shadows.

Doing Bayes well requires extra effort

2.2 The Talk

It is time to have "the talk," that difficult, embarrassing discussion about the statistical facts of life. And like some other talks we might have experienced,

Figure 2.4: XKCD 892: Null Hypothesis. Accessed from `m.xkcd.com/882`, used under the Creative Commons license.

this is not a one time issue but rather something we need to keep in mind throughout our work.

Considering biomedical research, Ioannidis (2005) has the provocative title "Why Most Published Research Findings are False." An effort to repeat 100 experiments in social psychology with statistically significant results found only 36 achieved significance when repeated[2] (Open Science Collaboration 2015), and the estimated effects averaged about half the size of the original studies. Baker (2016) reports on an informal poll of more than 1,500 scientists, saying "More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments." The clamor reached a point that the American Statistical Association felt the need to release "The ASA's Statement on p-Values: Context, Process, and Purpose" (Wasserstein and Lazar 2016). What in the world is going on? Why is Figure 2.4 so funny?

Many "significant" results not repeatable

Regression to the mean. In any measurement/remeasurement or test/retest situation, the second measurement is typically closer to the overall mean than the first measurement. This is the source of the term regression as it is used in statistics, and it means that the flashy result from your first experiment will usually not look so flashy when you repeat the experiment.

---

[2]Although bear in mind that the p-value is itself a random variable; it will not always be significant in a repeat experiment even absent any other issues.

Publication bias. Legions of valid, well-designed and well-executed experiments have been run that are never published or even discussed outside of the lab that ran them. Publication is biased in favor of experiments that have statistically significant results, even though non-significant (sometimes called negative) results are also informative.

Given the nature of significance testing, repeated testing of a true null hypothesis will eventually yield a significant result rejecting the null hypothesis. See Figure 2.5 for an application of this principle. If the earlier negative results are not publicly available, then that first significant result appears to stand on its own, when, in fact, a great deal of non-public evidence points in the other direction. Such results will not stand the test of time.

Options for analysis. This is a broad category that ranges from diligent analysis of experimental results all the way to scientific fraud. Experimenters can approach their data in many different ways, and it is responsible to attempt to extract as much information as possible from experimental results. However, every additional approach to the data is an opportunity to adapt the analysis to quirks of this particular set of results, quirks that are unlikely to be repeated in future data. Perhaps the treatments only look different for women, or perhaps the results appear to be more clear after removal of outliers. The more options you have in your analysis and the more hidden multiplicities of testing in your analysis, the less likely your "significant" result is to stand up to repeated experiments.

How bad can this be? Simmons, Nelson, and Simonsohn (2011) describe a simulation study wherein there were two completely random responses, three "treatment" conditions, and twenty observations per treatment with subjects of each gender. Suppose the researchers have the options of choosing among dependent variables, adding 10 more observations per treatment (after seeing the results of the first 20 observations), controlling for gender or its interaction with treatment, or dropping one of the three treatments. If they take advantage of all of their possibilities, the probability that at least one of their analysis choices leads to a p-value less than .05 is .61. Three times out of five they will find something "significant" when looking at totally random data with no treatment effects whatsoever. That is how bad it can be.

The more extreme aspect of options for analysis is dredging the data for something, anything, that looks "significant." This is sometimes called *p-hacking*. Results from data dredging are less likely to be repeatable. Data dredging/p-hacking is fine if reported as such and described as being part of exploration of the data. It can be an important hypothesis-generating method for further experimentation.

What is not fine is pretending that a result from data dredging was originally hypothesized. This is sometimes called *HARKing* (hypothesis after results known). The difference between diligent analysis or data dredging and outright fraud is often in the reporting. Saying that treatments one and four are different for women in the full data set is
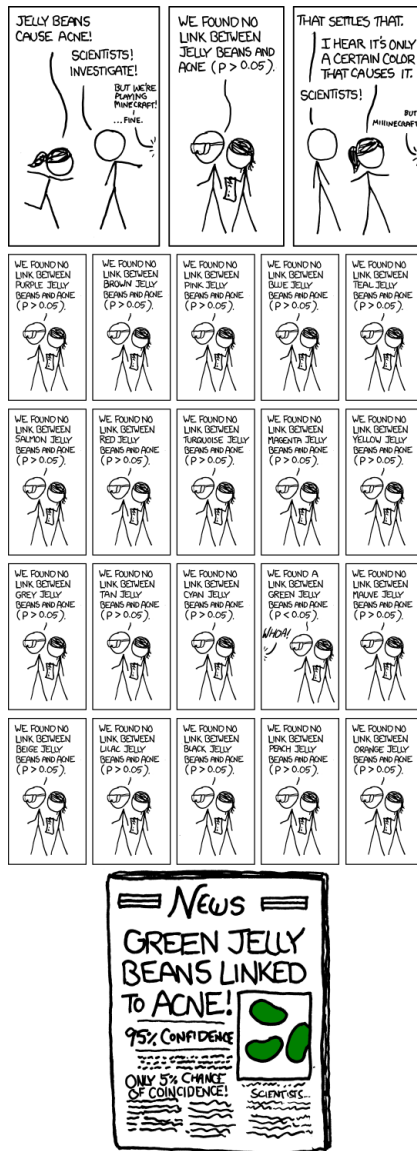
Figure 2.5: XKCD 882: Significant. Accessed from
`m.xkcd.com/882,` used under the Creative Commons license.

one thing; pretending that the experiment only contained women and
only used treatments one and four is fraud.

Curiously, traditional practice in analyzing designed experiments is ex-
tremely diligent about some aspects of multiple testing (for example, in
the context of pairwise comparisons) and totally oblivious to multiple
testing in other contexts (for example, in analysis of factorial experi-

ments).

Disincentives to repeat experiments. Wouldn't you be more certain of your result if you repeated your experiment and go the same result? Of course you would. But repeating the experiment costs money, repeating the experiment costs time (and you want to get that publication or product out the door), and repeating the experiment runs the risk of not getting the same results.[3]

Misunderstanding of p-values. When testing a null hypothesis, the p-value is the probability, *computed assuming that the null is true*, of observing results as extreme or more extreme than those in the data at hand. Extreme needs to be defined, but it roughly means "away from the null;" for a two-sample t-statistic, more extreme might mean larger in absolute value. Another way of thinking about a small p-value for a test is that either (a) the null is true and you were unlucky, or (b) the null is not true. No one likes to be unlucky.

That is all fine. The problems come when we try to take a p-value and interpret it as the probability that the null is true, or we base our decisions solely on the p-value, or we engage in p-hacking, or we interpret statistical significance as practical significance, and so on.

Confusing statistical and practical significance. A small p-value does not mean that the discovered effect has practical significance, and a large p-value does not mean that the data are inconsistent with with effects of a practically significant size.

Let's talk more about p-values. By long tradition, tests with p-values less than 5% are deemed statistically significant, and those with p-values less than 1% are deemed highly significant. A p-value is a form of evidence, and it is a reasonably continuous form of evidence. Decisions, however, are discrete and often binary. Thus we need to be wary of situations where we are making decisions near any cutoff point. A p-value of .049 is not in any practical sense different from a p-value of .051 as far as level of evidence goes, but if you have to make a binary decision, then you need to draw the line somewhere.

Continuous evidence, binary decisions

In order to illustrate the relationship between the p-value and the probability that a null hypothesis is false, let's make the following assumptions:

1. A fraction τ of the potential null hypotheses are false, and $1 - \tau$ of them are true.

2. We will reject the null if the p-value is less than $\mathcal{E}_I$, so a fraction $\mathcal{E}_I$ of the correct nulls will be rejected.

3. We will reject a false null with probability $1 - \mathcal{E}_{II}$ (or fail to reject it with probability $\mathcal{E}_{II}$).

---

[3]I once heard a very famous scientist say to take only one data point, because the second one will just confuse you. He was kidding ... I think.

Table 2.5: Sellke *et al.* approximate lower bound on the probability that rejecting the null is an error as a function of the p-value.

p	.05	.01	.005	.001	.0005	.0001
$\mathcal{P}(p)$	.29	.11	.067	.018	.01	.0025

This is an over-simplification of real world practice, but it works for purposes of illustration.

Putting all these together, we can compute the probability that a randomly chosen null we are testing is false given that we rejected it; this is called the *positive predictive value* of the test:

<div align="right">Positive predictive value</div>

$$P(\text{Null false}|\text{Null rejected}) = \frac{(1 - \mathcal{E}_{II})\tau}{(1 - \mathcal{E}_{II})\tau + \mathcal{E}_{I}(1 - \tau)} \qquad (2.6)$$

Smaller values of $\mathcal{E}_I$ and $\mathcal{E}_{II}$ lead to larger values of this probability, as do larger values of τ.

Sometimes you are operating in a "confirmatory" mode; in such a case, you are attempting to verify a result that has previous evidence in its favor, and you would expect τ to be fairly large. In other cases you could be operating in an "exploratory" or "hypothesis generating" mode, and there could be many null hypotheses with little to no prior evidence that they are false; for these situations, τ is likely to be very small. For example, this might be true in a brain imaging experiment where we are examining tens of thousands of brain regions for the handful that might be involved in a cognitive process.

<div align="right">Confirmatory or exploratory mode?</div>

Here are a couple of numerical examples. Set $\mathcal{E}_I = .05$, which is the traditional cutoff for statistical significance. Suppose that $\mathcal{E}_{II} = .1$, meaning that we have a large enough sample size or large enough effect sizes that we are reasonably sure of detecting any null that is actually false. (Many real world experiments have $\mathcal{E}_{II}$ considerably larger than .1.) Finally, suppose that we are in a confirmatory mode with $\tau = .5$. For these values, the probability of a rejected null actually being false works out to .947, so rejection is fairly strong evidence that the null is false. But what if we are working in an exploratory mode? If we change τ to .001 to reflect one possible exploratory situation, then the probability that a rejected null is actually false is less than .02, and we find that nearly all rejections are actually incorrect rejections. If we reduce $\mathcal{E}_I$ to .001, then the probability that a rejected null is actually false increases to .47, which is a lot bigger than .02, but not in the same range as we saw for confirmatory situations.

Here is another approach to the issue of what p-values mean. What we would really like to know is the probability that rejecting the null is an error; of course, the p-value does **not** give us that information. Sellke, Bayarri, and Berger (1999) define an approximate *lower bound* on this probability, and they show that this lower bound works pretty well in a wide variety of problems. Suppose that before seeing any data you thought that the null and alternative each had probability .5 of being true. Then for p-values less than

<div align="right">Approximate error probability</div>

$e^{-1} \approx .37$, the Sellke *et al.* approximate error probability is

$$\mathcal{P}(p) = \frac{-ep\log(p)}{1 - ep\log(p)} \quad .$$

The interpretation of the approximate error probability $\mathcal{P}(p)$ is that having seen a p-value of p, the probability that rejecting the null hypothesis is an error is *at least* $\mathcal{P}(p)$. Table 2.5 shows that the probability that rejection is a Type I error is more than .1, even for a p-value of .01. This lower bound also suggests that .05 and .01 traditional criteria for significance are not sufficiently stringent, and .005 and .001 might be more in line with what we are looking for.

How can we move forward? What should we do? Here are a few suggestions.

Preregister your design and analysis plan. If you describe your method of analysis in detail prior to seeing the data and stick to it, then you have provided yourself some protection against how the multiplicity of options affects the probability that your result will stand up to scrutiny and replication. Some grants require this. Preregistration is typically not a contract, but it does mean that you will need to explain deviations from the plan.

Document your data and analysis. Even if you do not register your analysis plan, be sure to thoroughly document the analysis you did perform. This should include thorough documentation of any software commands you used to produce your results. The `rmarkdown` package in **R** can be very helpful here. Similarly, thorough documentation of your data works toward openness in research. This is in the spirit of full disclosure so that others can properly evaluate your results. It can also save you a lot of time down the road when you need to come back and reconsider a data set.

Repeat experiments when possible. This is potentially costly and sometimes logistically impossible, but the gold standard for whether your results will stand up to scrutiny and replication is to repeat them and see what happens. Detailed documentation of methods is a crucial step in making experiments reproducible.

Publish or publicize non-significant/negative results. Ideally this would be done at a disciplinary level, but you can start by creating your own online archive of non-significant results, or, better yet, working with the library at your institution. This is only useful if others can find and understand what you did, so thorough documentation, careful keyword indexing, and metadata will be key.

Some journals are beginning to use *results independent review,* also called *registered reports,* wherein papers are accepted based on the introduction and methods sections, before the results are known. This also encourages high quality research proposals.

Use appropriate levels of significance. If we want the probability of the null being false given a rejection to be reasonably high, then Equation 2.6 shows us that the type I error rate $\mathcal{E}_I$ will need to decrease as τ (the probability that a random null hypothesis is false) decreases. For experiments with probability of rejecting a false null (power) of at least .5 ($\mathcal{E}_{II}$ of at most .5), using $\mathcal{E}_I = \tau/6$ as a cut off for significance will give us probability of at least .75 that a rejected null is, in fact, false. Of course, this just pushes the problem back to specifying τ, but it does tell us that if we wish to test whether ESP exists, we will need to test with very small p-value cut offs for significance.

Understand p-values. This ought to be a given, but it is all too easy to misinterpret a p-value, even when you know what it means. Read and understand the American Statistical Association statement on p-values (Wasserstein and Lazar 2016), which pushes the following points:

1. p-values can indicate how incompatible the data are with a specified statistical model.
2. p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of an effect.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

2.3 Problems

We wish to evaluate a new textbook for a statistics class. There are seven class sections; four are chosen at random to receive the new book, three receive the old book. At the end of the semester, student evaluations show the following percentages of students rate the textbook as "very good" or "excellent" (data set `TextBooks`):

Exercise 2.1

	Section						
	1	2	3	4	5	6	7
Book	New	Old	Old	New	New	Old	New
Rating	46	37	47	45	32	62	56

Use standard frequentist and randomization approaches to (a) test the null hypothesis that the two texts have the same average rating, and (b) create confidence intervals for the difference in average rating.

Dairy cows are bred by selected bulls, but not all cows become pregnant

Exercise 2.2

at the first service. A drug is proposed that is hoped to increase the bulls fertility. Each of seven bulls will be bred to 2 herds of 100 cows each (a total of 14 herds). For one herd (selected randomly) the bulls will be given the drug, while no drug will be given for the second herd. Assume the drug has no residual effect. The response we observe for each bull is the number of impregnated cows under drug therapy minus the number of impregnated cows without the drug. The observed differences are -1, 6, 4, 6, 2, -3, 5 (data set `CalfCounts`). Find the p-value for the randomization test of the null hypothesis that the drug has no effect versus a one-sided alternative (the drug improves fertility).

As part of a larger experiment, Dale (1992) looked at six samples of a wetland soil undergoing a simulated snowmelt. Three were randomly selected for treatment with a neutral pH snowmelt; the other three got a reduced pH snowmelt. The observed response was the number of Copepoda removed from each microcosm during the first 14 days of snowmelt (data set `Copepoda`).

Exercise 2.3

Reduced pH			Neutral pH		
256	159	149	54	123	248

Compare frequentist and randomization methods for testing the null hypothesis that pH does not affect the count of Copepoda.

Chu (1970) studied the effect of the insecticide chlordane on the nervous systems of American cockroaches. The coxal muscles from one meso- and one metathoracic leg on opposite sides were surgically extracted from each of six roaches. The roaches were then treated with 50 micrograms of α-chlordane, and coxal muscles from the two remaining meso- and metathoracic legs were removed about two hours after treatment. The Na^+-K^+ATPase activity was measured in each muscle, and the percentage changes for the six roaches are given here:

Exercise 2.4

$$15.3 \quad -31.8 \quad -35.6 \quad -14.5 \quad 3.1 \quad -24.5$$

Data set `Cockroaches`. Test the null hypothesis that the chlordane treatment has not affected the Na^+-K^+ATPas activity. What experimental technique (not mentioned in the description above) must have been used to justify a randomization test?

Twenty-six boards are cut to dimension 26 inches, by 2.5 inches by .75 inches. Thirteen of the boards are randomly selected, and these boards are planed to a uniform thickness of .625 inches. The remaining thirteen boards have a notch cut in the center that is 1 inch wide and .125 inch deep (that is, .625 inch of wood remains under the notch). Each board is then supported at the ends and pressure is applied in the center until the board fails (center-point loading at .1 in/minute across a span of 24 inches). The response is the breaking strength of the boards, in pounds. Data from D. Shmulsky, data set `NotchedBoards`.

Exercise 2.5

Shape	Breaking strength (lbs)					
Uniform	243	229	305	395	210	311	289
	269	282	399	222	331	369	
Notched	215	202	273	292	253	247	350
	246	352	398	267	331	342	

Compare the results of standard frequentist and randomization procedures for testing the null hypothesis that the strength of the boards is equal for the two shapes.

McElhoe and Conner (1986) use an instrument called a "Visiplume" to measure ultraviolet light. By comparing absorption in clear air and absorption in polluted air, the concentration of SO_2 in the polluted air can be estimated. The EPA has a standard method for measuring SO_2, and we wish to compare the two methods across a range of air samples. The recorded response is the ratio of the Visiplume reading to the EPA standard reading. There were six observations on coal plant number 2: .950, .978, .762, .733, .823, and 1.011 (data set `VisiplumePlant2`).

Problem 2.1

If we make the null hypothesis be that the Visiplume and standard measurements are equivalent (and the Visiplume and standard labels are just labels and nothing more), then the ratios could (with equal probability) have been observed as their reciprocals. That is, the ratio of .950 could with equal probability have been 1/.950 = 1.053, since the labels are equivalent and assigned at random. Suppose we take as our summary of the data the sum of the ratios. We observe .95 + ... + 1.011 = 5.257. Test (using randomization methods) the null hypothesis of equivalent measurement procedures against the alternative that Visiplume reads higher than the standard.

Are you a frequentist, a predictivist, or a Bayesian? Why? Do you consider your reason to be a good reason?

Problem 2.2

What is the standard p-value cutoff used in your field of study in order to declare some result "significant"? Is this cutoff appropriate for the kinds of experiments conducted in your field?

Problem 2.3

In your field of study, where could you preregister a design and analysis plan for an experiment?

Problem 2.4

Your lab partner analyzes your results can says that according to a t-interval, the probability that the mean response is between 1.73 and 2.11 is .95; comment on this statement.

Problem 2.5

Consider data $x_1, x_2, \ldots, x_n$. Model 1 says that these data are independent, normally distributed, with mean 0 and variance 1. Model 2 says that given μ, the data are independent, normally distributed with mean μ and variance 1, and in addition, μ has a prior distribution that is normal with mean 0 and variance 1.

Question 2.1

(a) Compute the Bayes factor for model 2 relative to model 1.

(b) The z test statistic for testing the null hypothesis $\mu = 0$ is $z = \sqrt{n}\ \bar{x}$; z is normally distributed with mean 1 and variance 1 if model 1 is correct.

Rewrite the Bayes factor in terms of z^2, and find the range of z^2 for which the Bayes factor favors model 2.

(c) Show that the probability that the Bayes factor selects model 1 when model 1 is correct goes to 1 as the sample size tends to infinity.

(d) Explain why the z test has a positive probability of rejecting $\mu = 0$ when model 1 is correct, even for arbitrarily large sample sizes.

(e) Show that the probability that the Bayes factor selects model 2 when model 2 is correct goes to 1 as the sample size tends to infinity.

(f) The "Bayesian Information Criterion" (BIC) is

$$BIC = \log(n)p_k - 2\log(L^k_{max}) \ \ .$$

Using your answer to part (b), explain why the BIC multiplies the number of parameters by the log of the sample size.

Question 2.2

The Dickey-Savage ratio says that for certain priors, when the "null" model is a restriction of the alternative model to certain parameter values, the Bayes factor will be the ratio of the posterior distribution at the null values (under model 2) to the prior distribution of the null values (under model 2). See Dickey and Lientz (1970). If the null values become more likely after seeing the data, then model 1 is the preferred model.

Explain in a heuristic way how the ROPE criterion is related to the Bayes factor using the Dickey-Savage ratio.

Chapter 3

Completely Randomized Designs

The simplest randomized experiment for comparing several treatments is the Completely Randomized Design, or CRD. We will study CRD's and their analysis in some detail before considering any other designs, because many of the concepts and methods learned in the CRD context can be transferred with little or no modification to more complicated designs. Here, we define completely randomized designs and describe the initial analysis of results.

3.1   Structure of a CRD

We have g treatments to compare and N units to use in our experiment. For a completely randomized design:

<div align="right">All partitions of units with sizes n_1 through n_g equally likely in CRD</div>

1. Select sample sizes $n_1, n_2, \ldots, n_g$ with $n_1 + n_2 + \cdots + n_g = N$.

2. Choose n_1 units at random to receive treatment 1, n_2 units at random from the $N - n_1$ remaining to receive treatment 2, and so on.

This randomization produces a CRD; all possible arrangements of the N units into g groups with sizes n_1 though n_g are equally likely. Statistically, that is all there is to a CRD. Note, however, that there is a lot more to creating the experiment than the randomization of treatments to units; the experimenter must also select the treatments, experimental units, and responses. Doing these requires much non-statistical insight.

<div align="right">First consider a CRD</div>

Completely randomized designs are the simplest, most easily understood, and most easily analyzed designs; they are also most robust against experimental difficulties such as missing data. For these reasons, we consider the CRD first when designing an experiment. The CRD may prove to be inadequate for some reason, but I always consider the CRD when developing an experimental design before possibly moving on to a more sophisticated design.

Example 3.1 Acid rain and birch seedlings

Wood and Bormann (1974) studied the effect of acid rain on trees. "Clean" precipitation has a pH in the 5.0 to 5.5 range, but observed precipitation pH in northern New Hampshire is often in the 3.0 to 4.0 range. Is this acid rain harming trees, and if so, does the amount of harm depend on the pH of the rain?

One of their experiments used 240 six-week-old yellow birch seedlings. These seedlings were divided into five groups of 48 *at random*, and the seedlings within each group received an acid mist treatment 6 hours a week for 17 weeks. The five treatments differed by mist pH: 4.7, 4.0, 3.3, 3.0, and 2.3; otherwise, the seedlings were treated identically. After the 17 weeks, the seedlings were weighed, and total plant (dry) weight was taken as response. Thus we have a completely randomized design, with five treatment groups and each n_i fixed at 48. The seedlings were the experimental units, the mist pH levels were the treatments, and plant dry weight was the response.

This is a nice, straightforward experiment, but let's look over the steps in planning the experiment and see where some of the choices and compromises were made. It was suspected that damage might vary by pH level, plant developmental stage, and plant species, among other things. This particular experiment only addresses pH level (other experiments were conducted separately). Many factors affect tree growth. The experiment specifically controlled for soil type, seed source, and amounts of light, water, and fertilizer. The desired treatment was real acid rain, but the available treatment was a synthetic acid rain consisting of distilled water and sulfuric acid (rain in northern New Hampshire is basically a weak mixture of sulfuric and nitric acids). There was no placebo *per se*. The experiment used yellow birch seedlings; what about other species or more mature trees? Total plant weight is an important response, but other responses (possibly equally important) are also available. Thus we see that the investigators have narrowed an enormous question down to a workable experiment using artificial acid rain on seedlings of a single species under controlled conditions. A considerable amount of nonstatistical background work and compromise goes into the planning of even the simplest (from a statistical point of view) experiment.

Example 3.2 Resin lifetimes

Mechanical parts such as computer disk drives, light bulbs, and glue bonds eventually fail. Buyers of these parts want to know how long they are likely to last, so manufacturers perform tests to determine average lifetime, sometimes expressed as mean time to failure, or mean time between failures for repairable items. The last computer disk drive I bought had a mean time to failure of 800,000 hours (over 90 years). Clearly the manufacturer did not have disks on test for over 90 years; how do they make such claims?

One experimental method for reliability is called an *accelerated life test*. Parts under stress will usually fail sooner than parts that are unstressed. By modeling the lifetimes of parts under various stresses, we can estimate (ex-

Table 3.1: $\log_{10}$ times until failure of a resin under temperature stress. Data set `ResinLifetimes`.

Temperature (^{o}C)									
175		194		213		231		250	
2.04	1.85	1.66	1.66	1.53	1.35	1.15	1.21	1.26	1.02
1.91	1.96	1.71	1.61	1.54	1.27	1.22	1.28	.83	1.09
2.00	1.88	1.42	1.55	1.38	1.26	1.17	1.17	1.08	1.06
1.92	1.90	1.76	1.66	1.31	1.38	1.16			

trapolate to) the lifetime of parts that are unstressed. That way we get an estimate of the unstressed lifetime without having to wait the complete unstressed lifetime.

Nelson (1990) gave an example where the goal was to estimate the lifetime (in hours) of an encapsulating resin for gold-aluminum bonds in integrated circuits operating at 120°C. Since the lifetimes were expected to be rather long, an accelerated test was used. Thirty-seven units were assigned at random to one of five different temperature stresses, ranging from 175° to 250°. Table 3.1 gives the $\log_{10}$ lifetimes in hours for the test units. Figure 3.1 shows a set of boxplots for these data with a superimposed line. Simple exploratory plotting of this sort is recommended before any formal analysis. For one thing, it shows us how far 120° (where we wish to predict) is from the data we have.

For this experiment, the choice of units was rather clear: integrated circuits with the resin bond of interest. Choice of treatments, however, depended on knowing that temperature stress reduced resin bond lifetime. The actual choice of temperatures probably benefited from knowledge of the results of previous similar experiments. Once again, experimental design is a combination of subject matter knowledge and statistical methods.

3.2 Goals, Models, and Inference

What are you trying to learn from your experiment? Are you trying to predict future observations? Are you trying to determine whether various treatments yield the same results? Are you trying to extrapolate to unobserved treatment levels? These are a few of the goals you might have based on your experiment. The kind of inference you do can differ depending on the goals, but all of these inferential goals share the need to choose an appropriate statistical model for the data. Comparing models and choosing the right model is the common theme in inference for experimental data.

3.2.1 Models

A *model* for data is a specification of the statistical distribution for the data. Generally, this specification is incomplete in the sense that it depends on
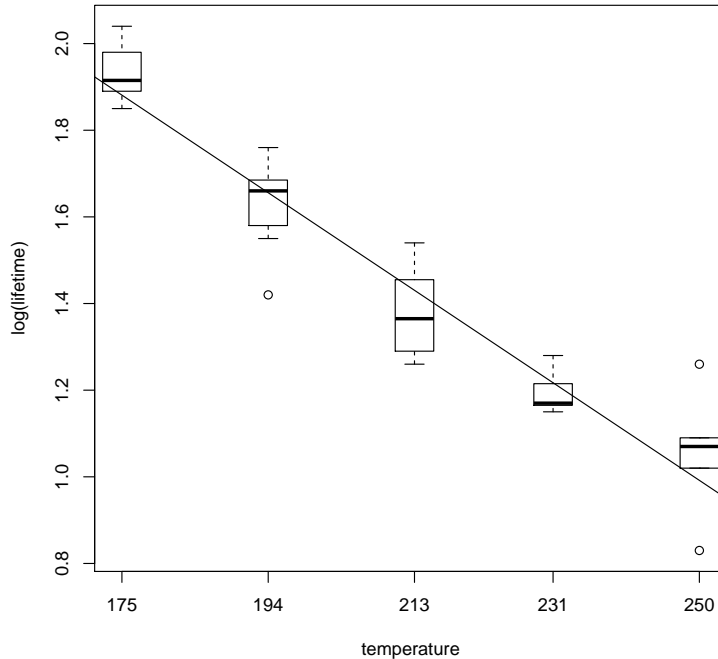
Figure 3.1: Boxplots of $\log_{10}$ times until failure of a resin under five different temperature stresses.

some unknown values called *parameters*. For example, the number of heads in ten tosses of a coin could be modeled as Binomial(10,p), where p is the unknown probability that the coin comes up heads. In the analysis of experimental data, we may posit several different models for the data, all with unknown parameters. The objectives of the experiment can often be described as deciding which model is the best description of the data, and, potentially, making inferences about the parameters in the models.

There are many ways to think about models, but it is common to consider the model to consist of two components: the model for the means (average or expected values), and the model for the variability (experimental error). For example, consider the birch tree weights from Example 3.1. We might assume that all the treatments have the same mean response, or that each treatment has its own mean, or that the means in the treatments are a straight line function of the treatment pH. Each one of these models for the means has its own parameters, namely the common mean, the five separate treatment means, and the slope and intercept of the linear relationship, respectively.

Model for the means

The second basic part of our data models is a description of how the data vary around the treatment means. This is the model for the variability

Model for the variability

or model for the errors. As with the model for the means, there are often several choices for how we model the variability. To begin, we will assume that the variability is normally distributed, with constant variance σ^2, and independent from observation to observation. Assuming normality, constant variance, and independence does not make those assumptions true, and we will eventually need to check, and potentially relax, all of those assumptions. But we begin with this model for the variability because it is the simplest, easiest to understand situation.

We will denote a response value by y, with subscripts indicating the particular value. Thus y_{ij} is the response for the jth unit in treatment i. We have Basic notation
g treatments, so i can run from 1 to g, and we have n_i units in the ith treatment group, so j can run from 1 to n_i. In some situations, we will also have a numerical value z_i associated with each treatment. For the tree seedlings, z_i indicates the pH in treatment i; for the resin lifetimes, z_i indicates the temperature in treatment i. We will generically call z_i a *dose*, but in any particular setting we could be more specific.

Here are a few potential models for the mean structure for y_{ij}: Basic mean structures

0 This is the *zero-mean model*. It is rarely used, and typically only arises when the responses y_{ij} are themselves differences of other values, which differences might reasonably have mean zero. More generally, if you have a model wherein all data should have mean δ, then you can use the zero mean model for $y_{ij} - \delta$.

μ This is the *single-mean model* wherein all responses are assumed to have the same mean, but the mean is unknown and would need to be estimated.

$\theta_0 + \theta_1 z_i$ This is the *linear-in-dose model*, which only makes sense when there is a "dose" in the treatments.

$\theta_0 + \theta_1 z_i + \theta_2 z_i^2$ This is the *quadratic-in-dose model*. Clearly, one can potentially go to higher powers (but no higher than $g - 1$).

$f(z_i; \theta)$ This is a generic, functional *dose-response model* included here merely to indicate that one does not need to use polynomials. However, you do need to specify what form of function you will use if you choose not to use polynomials.

μ_i This is the *separate-means model* or *treatment-means model*. Here all responses in a given treatment have the same mean or expected value, but different treatments can have different means.

μ_{ij} This is the *saturated model,* with every response having its own mean. One does not typically use this model in data analysis, but the saturated model does appear in the definition of some inferential quantities (such as the deviance, which we will see later when we consider models that do not use the normal distribution for variability).

Our basic model is then

$$y_{ij} = E(y_{ij}) + \epsilon_{ij}$$

where the expected value $E(y_{ij})$ is described by a model for the mean structure, and the experimental errors ϵ_{ij} are assumed to be independent, normal, with mean 0 and variance σ^2.

The standard analysis for comparative experiments is concerned with the structure of the means. We are trying to learn whether the means are all the same, or if some differ from the others, and the nature of any differences that might be present. The error structure is assumed to be of lesser interest, and we generally deal with the structure of the variability in service of learning about the means.

Standard analysis explores means

Let me emphasize that the mean structure comparisons in the standard analysis may not be the only models of interest, even though they are often an appropriate place to begin. For example, the structure of the variability is the key in Example 3.3.

Standard analysis is not always appropriate

Example 3.3 Luria, Delbrück, and variances

In the 1940s it was known that some strains of bacteria were sensitive to a particular virus and would be killed if exposed. Nonetheless, some members of those strains did not die when exposed to the virus and happily proceeded to reproduce. What caused this phenomenon? Was it spontaneous mutation, or was it an adaptation that occurred after exposure to the virus? These two competing theories for the phenomenon led to the same average numbers of resistant bacteria, but to different variances in the numbers of resistant bacteria—with the mutation theory leading to a much higher variance. Experiments showed that the variances were high, as predicted by the mutation theory. This was an experiment where all the important information was in the variance, not in the mean. It was also the beginning of a research collaboration that eventually led to the 1969 Nobel Prize for Luria and Delbrück.

3.2.2   Selecting a Model

You select models differently depending on your goals. The two most common goals are determining whether treatments have any effect (that is, deciding between a single-mean model and a more comples means model) and predicting future values. Note that even though the goals may be different, the models you consider are typically the same.

If your goal is to determine whether treatments have any effect, then you are working in the realm of model comparison via Analysis of Variance or Bayes Factor. The philosophy is *parsimony*. A dictionary definition of parsimony is an unwillingness to spend resources. In our context, resources are *degrees of freedom* in the model (roughly the number of parameters in the model), so we choose to use no more parameters than we absolutely need to use. This approach is widely seen in science. "Occam's Razor" holds that when there are two satisfactory explanations, the simpler explanation is better. "Einstein's Blade" holds that things should be made as simple as pos-

Parsimony

sible, but not simpler. In other words, use as many parameters as you need to fit the data well, but do not use more than that.

If your goal is to predict future values, selecting the model that minimizes AICc or LOOCV is recommended.

3.3 Frequentist Model Comparison

Analysis of Variance, or ANOVA for short, is the standard frequentist method for comparing models when the variability is assumed to be independent, normal, and constant variance and the models are fit by using least squares. Let $\hat{y}_{ij}$ be the fitted or predicted value from a model. Least squares fitting chooses from among all potential fitted values that the model could produce to use that set of fitted values that minimizes $\sum_{ij}(y_{ij} - \hat{y}_{ij})^2$, called the sum of squared errors or the sum of squared residuals. Minimizing the sum of squared errors is the Maximum Likelihood approach for independent normal errors with constant variance. When people speak of ANOVA they mean both an algorithm for partitioning the variability in the data (into sums of squares) that can be applied to most any data set as well as an inferential framework that is appropriate for comparing models when the errors from the model are independent and normally distributed with constant variance.

ANOVA comparisons, least squares

3.3.1 Fitting the models

Before comparing models we need to fit models.

Example 3.4 Frequentist model fitting for resin lifetimes

The displays below show excerpts from an **R** session fitting some of these models.

```
1 > options(contrasts=c("contr.sum","contr.poly"))
2 > head(ResinLifetimes,3)
   temp.z logTime temp       Time
1    175    2.04  175 109.64782
2    175    1.91  175  81.28305
3    175    2.00  175 100.00000
3 > myRL <- within(ResinLifetimes,
    {temp.z2 <- temp.z^2;temp.z3<-temp.z^3;temp.z4<-temp.z^4})
4 > head(myRL,3)
   temp.z logTime temp       Time    temp.z4 temp.z3 temp.z2
1    175    2.04  175 109.64782 937890625 5359375   30625
2    175    1.91  175  81.28305 937890625 5359375   30625
3    175    2.00  175 100.00000 937890625 5359375   30625
```

The gray line numbers are not part of the session and have been added for clarity. The greater than sign is the **R** prompt, and the user types commands to the right of the prompt. We will return to line 1 in section 3.5, but for now know that this command is executed silently when you load the `cfcdae` package. Line 2 simply prints the first three rows of the `ResinLifetimes`

data set, which includes the data from Table 3.1. Line 3 creates a new copy of `ResinLifetimes` that includes second, third, and fourth powers of temperature (`temp.z`), and line 4 shows the first three rows of the new data set. Note that `temp` is a factor version of temperature; factors indicate *groups* rather than *quantities*.

```
 5 > fit.single <- lm(logTime ~ 1,data=myRL)
 6 > fit.separate <- lm(logTime ~ temp,data=myRL)
 7 > fit.linear <- lm(logTime ~ temp.z,data=myRL)
 8 > fit.quad <- lm(logTime ~ temp.z + temp.z2,data=myRL)
 9 > fit.quad2 <- lm(logTime ~ temp.z + I(temp.z^2),data=myRL)
10 > fit.cube <- lm(logTime ~ temp.z + temp.z2 + temp.z3,data=myRL)
11 > fit.quart <- lm(logTime ~ temp.z + temp.z2 + temp.z3 + temp.z4,data=myRL)
12 > fit.poly1 <- lm(logTime ~ poly(temp.z,1),data=myRL)
13 > fit.poly2 <- lm(logTime ~ poly(temp.z,2),data=myRL)
14 > fit.poly3 <- lm(logTime ~ poly(temp.z,3),data=myRL)
15 > fit.poly4 <- lm(logTime ~ poly(temp.z,4),data=myRL)
```

Lines 5 and 6 fit the single-mean and separate-means models. The `lm` function fits "linear models" via least squares; this is the standard frequentist approach. The principal argument to `lm` is a formula consisting of a response variable (here `logTime` for the logarithm of the lifetimes), followed by a tilde (interpreted as "is modeled by"), followed by explanatory variables. The term `1` indicates the overall mean or intercept; it is included in the model unless specifically removed from the model. Line 6 uses a factor version of temperature as the explanatory variable, so a separate mean will be fit for each level of temperature. Lines 7–15 fit polynomial models using different approaches. Lines 7–11 fit the polynomial models by explicitly including powers of temperature (line 9 shows that it is possible to create the powers within the model formula). Lines 12–15 show how you can let **R** create the power terms itself. It does this by creating *orthogonal polynomials*. These orthogonal polynomials are mathematically equivalent to the ordinary polynomials we used in lines 7–11 in the sense that you can go back and forth between ordinary and orthogonal polynomials and, again mathematically, the two versions will give the same fitted values. However, the orthogonal polynomials yield much more stable computations; this is especially true when the z values are all far from 0, as is the case with these data.

```
16 > with(myRL,boxplot(logTime ~ temp,ylab="log(lifetime)",xlab="temperature",
         at=unique(temp.z),boxwex=5))
17 > abline(fit.linear)
```

Lines 16 and 17 produced Figure 3.1. Using `with` on line 16 tells **R** to look for variables within that data set. Note: we will see that several of these models produce equivalent fits to the data but use different parameterizations to produce those equivalent fits.

3.3.2 The Analysis of Variance

Using ANOVA to compare two models only works when one model is a special case, or restricted version, of another model. The smaller model is said to

be nested in the larger model. For example, you can produce the single-mean model from the separate-means model by setting all of the separate means to be equal to each other. Likewise, you can get any of the polynomial models from the separate-means model by restricting the separate means to lie on a polynomial curve. More obviously, you can get a lower order polynomial from a higher order polynomial by setting the higher order θ coefficients to 0. Thus we can compare these models using ANOVA.

Strictly speaking, the ANOVA decomposition is just an application of the Pythagorean Theorem. The process of minimizing the sum of squared errors produces a right triangle: in N dimensional space, the vector of residuals $(y_{ij} - \hat{y}_{ij})$ is perpendicular to the vector of fitted values $(\hat{y}_{ij})$. Thus squared lengths (sums of squares) will add up appropriately. If we further try to approximate the fit from a large model by the fit from a smaller, nested model, then we will get another right triangle and a further partitioning of the sums of squares.

Let model 1 be nested in model 2. For example, model 1 could be the single-mean model, and model 2 could be the separate-means model. Let RSS_1 be the residual sum of squares for model 1, and let RSS_2 be the residual sum of squares for model 2. The improvement sum of squares for going from the nested model 1 to the enclosing model 2 is $RSS_1 - RSS_2$. This will always be nonnegative, because the enclosing model can always fit at least as well as the nested model.

Consider a sequence of models: model 1 nested in model 2 nested in model 3 with residual sums of squares RSS_1, RSS_2, and RSS_3. This might be the single-mean model nested in the quadratic model nested in the separate-means model. Model i uses k_i parameters to describe the means. In our example, $k_1 = 1$ (just a single-mean), $k_2 = 3$ (an intercept, a slope, and a quadratic coefficient), and $k_3 = 5$ (five group means). When we go from model 1 to model 2, we spend $k_2 - k_1$ parameters and gain a reduction in residual sum of squares of $RSS_1 - RSS_2$. When we go from model 2 to model 3, we spend an additional $k_3 - k_2$ parameters to gain a reduction in residual sum of squares of $RSS_2 - RSS_3$. For a sequence of models, ANOVA produces a sequence of incremental improvement sums of squares for going to larger and larger models, and finally is left with the RSS for the largest model.

There are several ways to summarize this information. The one most closely related to the last paragraph is

Model	Residual DF	Residual SS	Incremental DF	Incremental SS
1	$N - k_1$	RSS_1		
2	$N - k_2$	RSS_2	$k_2 - k_1$	$RSS_1 - RSS_2$
3	$N - k_3$	RSS_3	$k_3 - k_2$	$RSS_2 - RSS_3$

If you ask **R** to compare several linear models, its output will look something like this table.

The usual version of an ANOVA table hides most of the information in the residual columns of the preceding version and displays incremental information and the "leftovers" (residuals) in the last line:

Source	DF	SS	MS
Model 1 to Model 2	$k_2 - k_1$	$RSS_1 - RSS_2$	$(RSS_1 - RSS_2)/(k_2 - k_1)$
Model 2 to Model 3	$k_3 - k_2$	$RSS_2 - RSS_3$	$(RSS_2 - RSS_3)/(k_3 - k_2)$
Residuals (to model 3)	$N - k_3$	RSS_3	$RSS_3/(N - k_3)$

Here MS abbreviates "mean square," which is a sum of squares divided by its degrees of freedom. A mean square is variability explained per degree of freedom used.

Finally, there are some shortcut, or abbreviated, names for many of the elements of this table. Using these shortcuts, we get what is considered the standard ANOVA table. Typically, "Model i to Model $i + 1$" is written as "Model $i + 1$" with the fact that it is actually an improvement suppressed. For example, when comparing the separate-means model to the single-mean model, the line is often labeled "Treatments," or in the case of our example might be labeled "Temperature." Similarly "$RSS_i - RSS_{i+1}$" is written as SS_{i+1} with the fact that it is a difference suppressed, and "$k_{i+1} - k_i$" is written as df_{i+1}, again with the difference suppressed. This gives us

Standard ANOVA table

Source	DF	SS	MS	F
Model 2	df_2	SS_2	$MS_2 = SS_2/df_2$	MS_2/MS_E
Model 3	df_3	SS_3	$MS_3 = SS_3/df_3$	MS_3/MS_E
Residuals	df_E	RSS_3	$MS_E = RSS_3/df_E$	

Note that we have added a column labeled "F," which gives the ratio of the MS for a line to the MS for error. This compares the variability explained per degree of freedom for increasing the size of the model to the variability per degree of freedom in the residuals.

F-ratio

When the mean structure in our model is large enough that it encompasses the mean structure in the data, then the expected value of MS_E is σ^2, the variance of the experimental errors. If model i describes the complete mean structure for the data, and model i is nested in model $i + 1$, then the (improvement) mean square for moving from model i to model $i+1$ also has expected value σ^2, the same as for MS_E. On the other hand, when the larger model is needed to explain the mean structure, then the expected value of its mean square is larger than σ^2. If the MS for a model term is larger than MS_E, then it is explaining more variability per degree of freedom that we would expect from random variation. If that F-ratio is large enough, then we would conclude that we need that term in the model.

MS_E estimates σ^2

Under the null hypothesis that model i completely explains the mean structure, then the F statistic for going from model i to (enclosing) model $i+1$ follows an F distribution with degrees of freedom the same as its numerator

F test and p value

and denominator mean squares. For example, in the ANOVA table above, the degrees of freedom for testing model 3 against model 2 would be df_3 and df_E. The p value for the model comparison is the area under an F-curve (with those degrees of freedom) to the right of the observed F. A small p-value indicates that we should prefer the larger model.

Example 3.5 ANOVA for resin lifetime models

On line 1, we use the `anova` command and give as arguments a set of nested models fit earlier.

```
1 > anova(fit.single,fit.quad,fit.separate)
  Analysis of Variance Table

  Model 1: logTime ~ 1
  Model 2: logTime ~ temp.z + temp.z2
  Model 3: logTime ~ temp
    Res.Df    RSS Df Sum of Sq        F Pr(>F)
  1     36 3.8313
  2     34 0.2937  2    3.5376 192.7245 <2e-16 ***
  3     32 0.2937  2    0.0000   0.0015 0.9985
  ---
  Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

In this case, **R** displays the information in the first form described above, showing the incremental improvements moving from model to model. The F-test for going from the single-mean model to the quadratic model has a tiny p-value, providing strong evidence that the linear and quadratic terms are needed. On the other hand, the test for going from the quadratic model to the separate-means model has a large p-value, indicating that the additional two degrees of freedom are not explaining enough variability to justify their inclusion after the quadratic model.

```
2 > anova(fit.quart)
  Analysis of Variance Table

  Response: logTime
            Df Sum Sq Mean Sq  F value     Pr(>F)
  temp.z     1 3.4593  3.4593 376.9128 < 2.2e-16 ***
  temp.z2    1 0.0783  0.0783   8.5361  0.006338 **
  temp.z3    1 0.0000  0.0000   0.0020  0.964399
  temp.z4    1 0.0000  0.0000   0.0009  0.976258
  Residuals 32 0.2937  0.0092
  ---
  Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

On line 2, we ask for the ANOVA of a single fitted model. In this form, **R** creates an internal sequence of models beginning with the single-mean model and then adding in each term in the model sequentially. We see the improvement SS for adding in linear, then quadratic after linear, then cubic and linear and quadratic, and so forth. Note that the incremental SS for linear and quadratic in the second command sum to the improvement SS for going from the single-mean model to the quadratic model in the first command. We

see that the p-value for going from linear to quadratic is small (.006). This is significant by traditional standards, but not by the more rigorous standards we advocated earlier.

```
3 > anova(fit.poly4)
  Analysis of Variance Table

  Response: logTime
                 Df Sum Sq Mean Sq F value    Pr(>F)
  poly(temp.z, 4)  4 3.5376 0.88441  96.363 < 2.2e-16 ***
  Residuals       32 0.2937 0.00918
  ---
  Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Line 3 illustrates that while the form `poly(temp.z,4)` generates a fourth power polynomial, ANOVA in **R** treats it as a single bundle of four degrees of freedom (that is, a single term) and does not separate the powers sequentially as it did when we entered the powers ourselves.

```
4 > anova(fit.separate)
  Analysis of Variance Table

  Response: logTime
           Df Sum Sq Mean Sq F value    Pr(>F)
  temp       4 3.5376 0.88441  96.363 < 2.2e-16 ***
  Residuals 32 0.2937 0.00918
  ---
  Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Finally, line 4 shows the ANOVA for the separate-means model, with a single line for the four degrees of freedom between the five treatments. Note that the residual SS and df are the same in all four of these ANOVAs. The separate-means model and the fourth order polynomial (either using `poly()` or done by hand) exhaust the four degrees of freedom between five groups and give the same fitted values in each treatment, and thus the same residual SS. If we compare the single-mean model to the separate-means model, we would conclude that we needed the separate-means model, even though it uses all four degrees of freedom between treatments, and we have seen that two degrees of freedom suffice.

Note: order matters in these models. If you fit third and fourth powers first and then see whether you need first and second powers, you will find that you do not need first and second powers if third and fourth powers are in the model. However, we usually try to maintain *hierarchy*, where the presence of a term implies the presence of lower order terms (so cubic present would imply linear and quadratic present).

Order of terms matters

3.3.3 ANOVA Computations

In general we will use computer software to calculate the sums of squares in an ANOVA. However, there are simple formulae for SS for certain special situations. Understanding these special situations can help you understand

$$
\begin{aligned}
\text{SS}_{\text{Trt}} &= \sum_{i=1}^{g} n_i (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 \\[1ex]
\text{SS}_{\text{E}} &= \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\bullet})^2 \\[1ex]
\text{df}_{\text{Trt}} &= g - 1 \\[1ex]
\text{df}_{\text{E}} &= N - g
\end{aligned}
$$

Display 3.1: ANOVA quantities in the separate-means model.

what is going on in ANOVA.

Let's establish some notation for sample averages and the like. The sum of the observations in the ith treatment group is

$$
y_{i\bullet} = \sum_{j=1}^{n_i} y_{ij} \ .
$$

The mean of the observations in the ith treatment group is

Treatment means

$$
\overline{y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = y_{i\bullet}/n_i \ .
$$

The overbar indicates averaging, and the dot ($\bullet$) indicates that we have averaged (or summed) over the indicated subscript. The sum of all observations is

$$
y_{\bullet\bullet} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^{g} y_{i\bullet} \ ,
$$

and the grand mean of all observations is

Grand mean

$$
\overline{y}_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij} = y_{\bullet\bullet}/N \ .
$$

We want the ANOVA for the separate-means model. The fitted values for that model for data in treatment i will be the mean response from that treatment $\overline{y}_{i\bullet}$. The sum of squared deviations of the data from the group means is

Error SS

$$
\text{SS}_{\text{E}} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\bullet})^2 \ .
$$

We also need the improvement SS for going from a single mean to separate

Treatment or groups SS

means, denoted SS_{Trt}. This is

$$SS_{Trt} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 = \sum_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 \ .$$

It is a bit tedious to compute these SS by hand, but it is eminently doable (indeed, old guys such as myself remember computing this by hand!).

It also helps to see where these formulae come from. Consider the following:

$$y_{ij} = \overline{y}_{\bullet\bullet} + (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}) + (y_{ij} - \overline{y}_{i\bullet})$$

That is easy enough, we have just added and subtracted overall and group means on the right hand side. Now square both sides and add up.

Pythagorean theorem and ANOVA

$$
\begin{aligned}
\sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij}^2 &= N\overline{y}_{\bullet\bullet}^2 + \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 + \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\bullet})^2 \\
&= N\overline{y}_{\bullet\bullet}^2 + \sum_{i=1}^{g} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 + \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\bullet})^2 \\
&= SS_{Mean} + SS_{Trt} + SS_E
\end{aligned}
$$

Wait, time out! What happened to all of the cross-product terms that should appear when we square that sum on the right hand side? It turns out that the cross products add to zero; that is an algebraic expression of the fact this decomposition is forming right angles in N dimensional space, and the Pythagorean Theorem is going to work.

3.4 Predictive Model Comparison

Predictive model comparison is fairly straightforward: find the model with the lowest AIC or AICc value.

Example 3.6 ANOVA for resin lifetime models

Here we see the AICc computations in **R**.

```
1 > AICc(fit.single)
  [1] 25.44922
2 > AICc(fit.linear)
  [1] -58.4569
3 > AICc(fit.quad)
  [1] -64.68237
4 > AICc(fit.cube)
  [1] -61.99923
5 > AICc(fit.quart)
  [1] -59.13575
6 > AICc(fit.separate)
  [1] -59.13575
```

For these data, AICc model selection agrees with ANOVA: the separate-means model (line 6) is greatly preferred to the single-mean model (line 1), but the quadratic model (line 3) is preferred to the separate-means model, because the separate-means model uses more degrees of freedom than are strictly necessary.

3.5   Parameters

Statistical models contain parameters that control means, variances, and, potentially, other aspects of how the data are distributed. As straightforward as that sounds, mean parameters in particular are somewhat slippery beasts to latch onto. The main issues are alternatives and redundancy. Consider these two quadratic models for a mean response:

Alternatives and redundancy

$$\theta_0 + \theta_1 z_{ij} + \theta_2 z_{ij}^2$$

and

$$\tilde{\theta}_0 + \tilde{\theta}_1(z_{ij} - 2) + \tilde{\theta}_2(z_{ij} - 2)^2$$

These two models are both second order (quadratic) models, they will yield the same ANOVA or Bayes factors, the same fitted values, and the same residuals. But $\theta_0 \neq \tilde{\theta}_0$ and $\theta_1 \neq \tilde{\theta}_1$. Thus you and I can be talking about quadratic models and yet be talking about different parameters.

It is yet more interesting when there is redundancy. Suppose we wish to model a response using two predictors, w and z, via

$$\theta_0 + \theta_1 z_{ij} + \theta_2 w_{ij}$$

and unbeknownst to us, these two predictors are related via $w = 2z$. In this case, any combination of θ_1 and θ_2 that has the same value of $\theta_1 + 2\theta_2$ will yield exactly the same fitted values. Thus the pair (4,5) cannot be distinguished from the pair (6,4) or any of an infinite number of other combinations. You cannot say which is the "correct" pair, because all the pairs that follow that relationship are equally valid.

The redundancy example seems extreme, but it is very close to home. As we move further into our study, we will find that it will be convenient to write treatment means as a central value plus a deviation from the central value:

Many models have redundant parameters

$$\mu_i = \mu + \alpha_i$$

where μ is the central value and α_i is the deviation from the central value called the *treatment effect*. But we have $g+1$ parameters to describe g means, meaning there is redundancy. If we add 10 to μ and subtract 10 from each α_i, we get the same totals $\mu + \alpha_i$. The parameters μ and α_i are not *estimable* in and of themselves.

The way to move forward is to restrict or constrain the coefficients in our model so that we remove the redundancy. In the z, w example, we might

assume that $\theta_2 = 0$ and move forward with just z. In the μ, α_i example, we might assume that $\mu = \sum_i \mu_i / g$ (this is the same thing as assuming that $0 = \sum_i \alpha_i$) making μ be the average of the means. **R**, on the other hand, by default assumes $\alpha_1 = 0$ (or, equivalently, $\mu_1 = \mu$), although it lets you change that. For hand computation, assuming $\mu = \overline{y}_{\bullet\bullet}$ has some advantages. Each of these assumptions is equally valid in a mathematical sense.

The point of all this is not to scare you away from parameters but rather to impress on you the importance of knowing exactly how your parameters are defined and constrained. Using parameters computed under constraint A as if they were computed under constraint B will lead to confusion and chaos.

> In this book, unless otherwise noted, we will assume that for grouping factors the treatment effects sum to 0:
>
> $$\sum_{i=1}^{g} \alpha_i = 0$$

Combined with our assumptions of normality and constant variance, we can write this more completely as

$$
\begin{aligned}
y_{ij} &= \mu + \alpha_i + \epsilon_{ij} \quad \text{where} \\
\epsilon_{ij} &\sim \text{independent } N(0, \sigma^2) \quad \text{and} \\
0 &= \sum_{i=1}^{g} \alpha_i
\end{aligned}
$$

Fortunately, even though we had to make a choice of constraint, *the important things don't depend on which set of constraints we use.* Important things are treatment means, differences of treatment means (or, equivalently, differences of α_i's), and comparisons of models.

Our constraint that the treatment effects α_i add to zero implies that the treatment effects are not completely free to vary. We can set $g-1$ of them however we wish, but the remaining treatment effect is then determined because it must be whatever value makes the zero sum true. We express this by saying that the treatment effects have $g-1$ degrees of freedom. This is exactly the same as the increase in the number of free parameters going from the single-mean model to the separate-means model.

3.5.1 Estimating Parameters

Most data analysis these days is done using a computer. Few of us sit down and crunch through the necessary calculations by hand. Nonetheless, knowing the basic formulae and ideas behind our analysis helps us understand and interpret the quantities that come out of the software black box. If we don't understand the quantities printed by the software, we cannot possibly use them to understand the data and answer our questions.

The first thing to understand is that standard frequentist estimates will *not* equal the corresponding Bayesian estimates, and the estimates from different inference schools have different desirable properties. For example, the standard frequentist estimates are *unbiased.* Unbiased means that when you average the values of the estimates across all repeated experimental outcomes, you get the true parameter values. Bayesian estimates are generally at least a little biased. On the other hand, if you look at how far the estimate is from its target value, Bayesian estimates have a lower mean squared error than standard frequentist estimates.

The second thing to understand is that there are no explicit formulae for Bayesian estimates, only an algorithm to derive them, and the simple, explicit formulae for frequentist estimates only work in special circumstances. Thus what we are trying to accomplish in this section is to gain some insights, not learn a set of steps for general use.

3.5.2 Frequentist estimates

It is convenient to introduce a notation to indicate the estimator of a parameter. The usual notation in statistics is to put a "hat" over the parameter to indicate the estimator; thus $\widehat{\mu}$ is an estimator of μ. Because we have parameters that satisfy $\mu_i = \mu + \alpha_i$, our unbiased estimators will satisfy $\widehat{\mu}_i = \widehat{\mu} + \widehat{\alpha}_i$.

Consider first the separate-means model, with each treatment group having its own mean μ_i. The natural estimator of μ_i is $\overline{y}_{i\bullet}$, the average of the observations in that treatment group. We estimate the expected (or average) response in the ith treatment group by the observed average in the ith treatment group responses. Thus we have

$$\widehat{\mu}_i = \overline{y}_{i\bullet} .$$

The sample average is an unbiased estimator of the population average, so $\widehat{\mu}_i$ is an unbiased estimator of μ_i.

The treatment effects α_i are the differences of the treatment mean and the central value:

$$\alpha_i = \mu_i - \mu \;;$$

and the same will be true of the estimates

$$\widehat{\alpha}_i = \widehat{\mu}_i - \widehat{\mu} \;.$$

To go beyond this, we must make explicit what we mean by μ.

Use the notation $\overline{\mu}_\bullet$ for the mean of the group means:

$$\overline{\mu}_\bullet = \frac{1}{g} \sum_{i=1}^{g} \mu_i$$

We estimate $\overline{\mu}_\bullet$ via

$$\widehat{\overline{\mu}}_\bullet = \frac{1}{g} \sum_{i=1}^{g} \widehat{\mu}_i = \frac{1}{g} \sum_{i=1}^{g} \overline{y}_{i\bullet}$$

Because our standard, default constraint is that the sum of the treatment effects is 0 (or, equivalently, $\mu = \overline{\mu}_\bullet$), we have

$$\alpha_i = \mu_i - \overline{\mu}_\bullet$$

leading to estimates

$$\widehat{\alpha}_i = \widehat{\mu}_i - \widehat{\overline{\mu}}_\bullet = \overline{y}_{i\bullet} - \frac{1}{g}\sum_{i=1}^{g} \overline{y}_{i\bullet} \ .$$

The overall grand mean of the data is

$$\overline{y}_{\bullet\bullet} = \frac{1}{N}\sum_{i-1}^{g}\sum_{j=1}^{n_i} y_{ij} \ .$$

If the group sample sizes are all the same, called balanced data, that is, if $n = n_1 = \cdots = n_g$, then $\widehat{\overline{\mu}}_\bullet = \overline{y}_{\bullet\bullet}$:

$$\widehat{\overline{\mu}}_\bullet = \frac{1}{N}\sum_{i-1}^{g}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{g}\sum_{i-1}^{g}\frac{1}{n}\sum_{j=1}^{n} y_{ij} = \frac{1}{g}\sum_{i-1}^{g} \overline{y}_{i\bullet} = \overline{y}_{\bullet\bullet} \ .$$

We know that

$$\text{SS}_{\text{Trt}} = \sum_{i=1}^{g} n_i (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 \ ,$$

so *if all of the group sample sizes are the same*, we also have

$$\text{SS}_{\text{Trt}} = \sum_{i=1}^{g} n\widehat{\alpha}_i^2 \ .$$

This pattern of take an effect, square it, multiply by the number of units receiving the effect, and then add over the levels of the effect to get the sum of squares for the term in the model is generic even in more complex, but still balanced, models.

The remaining parameter in the separate-means model is the error variance σ^2. We estimate that variance by the mean square for error in the model, also denoted s^2:

$$\begin{aligned}
\widehat{\sigma}^2 = s^2 = \text{MS}_{\text{E}} \ &= \ \frac{1}{N-g}\sum_{i=1}^{g}\sum_{j=1}^{n} r_{ij}^2 \\
&= \ \frac{1}{N-g}\sum_{i=1}^{g}\sum_{j=1}^{n} [y_{ij} - (\widehat{\mu} + \widehat{\alpha}_i)]^2 \\
&= \ \frac{1}{N-g}\sum_{i=1}^{g}\sum_{j=1}^{n} [y_{ij} - \overline{y}_{i\bullet}]^2
\end{aligned}$$

Model	Parameter	Estimator
Separate means	μ_i	$\overline{y}_{i\bullet}$
	σ^2	$\frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet})^2}{N-g}$
	μ	$\sum_{i=1}^{g}\overline{y}_{i\bullet}/g$
	α_i	$\overline{y}_{i\bullet} - \sum_{i=1}^{g}\overline{y}_{i\bullet}/g$
(if balanced)	μ	$\overline{y}_{\bullet\bullet}$
(if balanced)	α_i	$\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}$

Display 3.2: Parameter estimators in the separate-means model.

where r_{ij} is the *residual* for the i,j point, equal to the response minus the fitted value. This is an unbiased estimate of σ^2. The formulae for these estimators are collected in Display 3.2.

The deviations from the group mean $y_{ij}-\overline{y}_{i\bullet}$ add to zero in any treatment group, so that any $n_i - 1$ of them determine the remaining one. Put another way, there are $n_i - 1$ degrees of freedom for error in each group, or $N - g = \sum_i(n_i - 1)$ degrees of freedom for error for the experiment. There are thus $N - g$ degrees of freedom for our estimate $\widehat{\sigma}^2$. This is analogous to the formula n_1+n_2-2 for the degrees of freedom in a two-sample t-test. Another way to think of $N - g$ is the number of data values minus the number of mean parameters estimated.

Error degrees of freedom

A point estimate gives our best guess as to the value of a parameter. A confidence interval gives a plausible range for the parameter, that is, a set of parameter values that are consistent with the data. Confidence intervals for the μ_i's are useful and straightforward to compute. Confidence intervals for μ and the α_i's are only slightly more trouble to compute, but are perhaps less useful, because there are several potential ways to define the α's based on different constraints. Differences between μ_i's, or equivalently, differences between α_i's, are extremely useful; these will be considered in depth in Chapter 4. Confidence intervals for the error variance σ^2 will be considered in Chapter 10.

Confidence intervals for means and effects

Confidence intervals for parameters in the mean structure have the general form:

unbiased estimate $\pm$ multiplier $\times$ standard error of estimate.

Generic confidence interval for mean parameter

The standard deviation for the average $\overline{y}_{i\bullet}$ is $\sigma/\sqrt{n_i}$. We do not know σ, so we use $\widehat{\sigma} = s = \sqrt{\mathrm{MS_E}}$ as an estimate and obtain $s/\sqrt{n_i}$ as the standard errors for $\overline{y}_{i\bullet}$. The standard error of an estimated treatment effect $\widehat{\alpha}_i$ in the balanced (equal replication) case is $\sigma\sqrt{1/n - 1/N}$. Again, we must use an

Parameter	Estimator	Standard Error
μ	$\overline{y}_{\bullet\bullet}$	$s/\sqrt{N}$
μ_i	$\overline{y}_{i\bullet}$	$s/\sqrt{n}$
α_i	$\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}$	$s\sqrt{1/n - 1/N}$

Display 3.3: Standard errors of point estimators in the separate-means model with balanced data.

estimate of σ, yielding $s\sqrt{1/n - 1/N}$ for the standard error. Keep in mind that the treatment effects $\widehat{\alpha}_i$ are negatively correlated, because they must add to zero. These standard errors appear in Display 3.3.

For an interval with coverage $1 - \mathcal{E}$, we use the upper $\mathcal{E}/2$ percent point of the t-distribution with $N - g$ degrees of freedom as the multipler. This is denoted $t_{\mathcal{E}/2, N-g}$. We use the $\mathcal{E}/2$ percent point because we are constructing a two-sided confidence interval, and we are allowing error rates of $\mathcal{E}/2$ on both the low and high ends. For example, we use the upper 2.5% point (or 97.5% cumulative point) of t for 95% coverage. The degrees of freedom for the t-distribution come from $\widehat{\sigma}^2$, our estimate of the error variance. For the separate-means model, the degrees of freedom are $N - g$.

Use t multiplier when error is estimated

Example 3.7 Frequentist estimates for resin lifetimes

Back in Example 3.4 there was a line 1 that set the `contrasts` option in **R** to `contr.sum`. That told **R** to use the parameterization where treatment effects sum to zero; you must set this option before fitting your model, although you only need to set it once. When you load the **R** package `cfcdae` (Companion to a First Course in Design and Analysis of Experiments, an **R** package to accompany this book), it automatically sets contrasts to `contr.sum`.

```
1 > summary(fit.separate)

Call:
lm.default(formula = logTime ~ temp)

Residuals:
     Min       1Q    Median       3Q      Max
-0.22667 -0.03667  0.00250  0.03125  0.20333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.43794    0.01585  90.708  < 2e-16 ***
temp1        0.49456    0.03065  16.134  < 2e-16 ***
temp2        0.19081    0.03065   6.225 5.67e-07 ***
temp3       -0.06044    0.03065  -1.972   0.0573 .
temp4       -0.24365    0.03222  -7.563 1.30e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.0958 on 32 degrees of freedom
Multiple R-squared:  0.9233,Adjusted R-squared:  0.9138
F-statistic: 96.36 on 4 and 32 DF,  p-value: < 2.2e-16
```

Line 1 shows the basic summary of the fitted separate-means model. We are particularly interested in the coefficients. For these, **R** gives us estimates and standard errors as well as results of a test that the coefficient is 0. The (Intercept) is what we were calling μ, so $\widehat{\mu} = 1.44$. Note: the sample sizes in the resin lifetime data set are *not* equal, so $\widehat{\mu} \neq \overline{y}_{\bullet\bullet} = 1.47$. Unequal sample sizes also causes the standard errors for the treatment effects to be unequal.

The interpretation of the t-tests in the output of line 1 is a test of the null hypothesis that a particular α_i is zero, allowing all the other effects in the model to be non-zero. We don't usually test that individual α_is are zero; we generally include all the levels in a factor or none of them. In fact, we can make any particular α_i zero by changing our definition of μ and thus changing our parameterization, so testing individual treatment effects is problematic. However, in the next chapter we will consider tests of linear combinations of treatment effects.

R only prints four treatment effects $\widehat{\alpha}_i$, one for each degree of freedom between treatments. The final treatment effect is set by the constraint that they must all add to 0.

```
2 > confint(fit.separate,level=.995)
                   0.25 %      99.75 %
(Intercept)    1.39014624   1.48573471
temp1          0.40214062   0.58697842
temp2          0.09839062   0.28322842
temp3         -0.15285938   0.03197842
temp4         -0.34078923  -0.14652029
```

Line 2 shows that it is straightforward to get confidence intervals for the coefficients, and you can also set the level of coverage.

```
3 > model.effects(fit.separate,"temp")
          175         194         213         231         250
   0.49455952  0.19080952 -0.06044048 -0.24365476 -0.38127381
```

Line 3 uses the `model.effects` function from the `cfcdae` library; it can compute that left out treatment effect for you. That is not challenging in this example, but it is helpful in more complicated situations.

```
4 > summary(fit.quad)
  ...
  Coefficients:
                Estimate Std. Error t value Pr(>|t|)
  (Intercept)    7.4179987  1.1564331   6.415 2.51e-07 ***
  temp.z        -0.0450981  0.0110542  -4.080 0.000258 ***
  temp.z2        0.0000786  0.0000261   3.011 0.004879 **


  ...
```

Now we see coefficient results for the quadratic model,

```
5 > summary(fit.quart)
  ...
  Coefficients:
                Estimate Std. Error t value Pr(>|t|)
  (Intercept)   9.699e-01  1.957e+02    0.005    0.996
  temp.z        7.573e-02  3.750e+00    0.020    0.984
  temp.z2      -7.649e-04  2.679e-02   -0.029    0.977
  temp.z3       2.600e-06  8.459e-05    0.031    0.976
  temp.z4      -2.988e-09  9.962e-08   -0.030    0.976

  Residual standard error: 0.0958 on 32 degrees of freedom
  Multiple R-squared:  0.9233,Adjusted R-squared:  0.9138
  F-statistic: 96.36 on 4 and 32 DF,  p-value: < 2.2e-16
```

the quartic model with ordinary polynomials,

```
6 > summary(fit.poly4)
  ...
  Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
  (Intercept)      1.465135   0.015750  93.027  < 2e-16 ***
  poly(temp.z, 4)1 -1.859909   0.095801 -19.414  < 2e-16 ***
  poly(temp.z, 4)2  0.279899   0.095801   2.922  0.00634 **
  poly(temp.z, 4)3  0.004310   0.095801   0.045  0.96440
  poly(temp.z, 4)4 -0.002873   0.095801  -0.030  0.97626
  ...
```

and the quartic model with orthogonal polynomials. Here are a few important lessons. First, the first and second order coefficients (linear and quadratic) are different in every model. In particular, you cannot simply drop the third and fourth order terms from the model in line 2 and assume the first and second order coefficients remaining will work on their own. In general, term coefficients depend on what other terms are in the model.

Second, standard errors of polynomial terms depend on how you parameterize the polynomial and, in general, what other polynomial terms are

in the model. For example, compare the standard errors for `temp.z` and `temp.z2` in the models from lines 1 and 2; they are orders of magnitude different! The problem is that the raw polynomial terms are highly correlated with each other, and that correlation increases the standard error of the estimates.

Third, think back to the meaning of the t-tests, which test whether a particular θ_i is zero allowing the other coefficients to be non-zero. The results from line 5 show that we can get rid of any single term if we keep the other three. However, it is highly recommended that you maintain *hierarchy*, which means that if you have a term of order k in your model, you also have terms of all lower orders in your model. Maintaining hierarchy, there is really only one test that matters, that of the fourth order term, which has a very large p-value and is not needed.

Fourth, the t-tests in the results of line 5 make it appear that the terms have no predictive ability. However, the overall F-test for the model shows that it is highly significant. Be sure to inspect the overall F as well as the individual coefficient ts.

3.6   Bayesian Analysis

We must specify the prior distributions in the Bayesian model before doing model fitting, model comparison, or inference on parameters. A bona fide Bayesian analysis would elicit these prior distributions from subject matter experts, but for ease of presentation, we will typically make use of generic, weakly informative priors created by our modeling tools.

In principle, Bayesian inference is straightforward. For model comparison, we compute the marginal likelihood of the data under each model and then compare these marginal likelihoods by taking their ratio as the Bayes factor. The model with the highest marginal likelihood (highest Bayes Factor as the numerator of the factor) is the preferred model. Alternatively, we can compute LOOCV for each potential model, and select the model with lowest value. Inference on parameters can be made by taking a sample from the posterior distribution and then computing appropriate summaries on that sample. For example, we take the average of the posterior samples as our estimate of the mean of the posterior distribution. We use quantiles (percent points) of the posterior samples to estimate quantiles of the posterior distribution. We can thus easily compute interval estimates by, for example, computing the 2.5% and 97.5% quantiles of the posterior samples. We can even get fancy and make inference about functions and combinations of parameters by computing the corresponding functions of the posterior samples, and then use means, quantiles, and so forth of the computed values to make the inference.

Inference based on posterior distribution

In practice, things are not quite so simple. For example, we can take samples from the (approximate) posterior distribution of the parameters using an approach called Markov chain Monte Carlo. These samples are *not* independent samples. When the correlation among these samples is too high, we say

that the chain is not mixing well, and we must either take additional samples to compensate for the correlation or we must reframe the model to reduce the correlation. When we take more samples, we might "thin" the results by retaining only a subset (say, every fifth sample) of those actually generated. The only purpose of this is to save space in the fitted results; it actually makes the estimates a little worse. One example of reframing the model is to use orthogonal polynomials in place of ordinary polynomials.

There are diagnostics that help us detect when a chain is not mixing well. For a well-mixing chain:

- A plot of the chain values against time (a trace plot) should look like a horizontal blur with no visible trends. Plots of multiple chains should overlap.

- The autocorrelation of the values of the chain should decay to zero quickly.

- The "effective sample size" of the chain should be fairly large, ideally nearly as large as the length of the chain.

If the chain is not mixing well, our inferences may not be accurate. A well-mixing chain makes life easy, and a poorly-mixing chain makes life hard.

The good news is that you can do Bayesian inference in **R**. The bad news is that there is no one-size-fits-all approach to Bayesian inference in **R**, but instead there are several packages that approach these computations in different ways, each with its own advantages and disadvantages. We will use the tools in `cfcdae`, because they are reasonably easy to use and set up priors in a way that matches this book. Note, however, that packages `brms` (Bürkner 2017) and `rstanarm` (Stan Development Team 2016) are more comprehensive than `cfcdae`. All three of these packages are wrappers around the `rstan` (Stan Development Team 2018) package, which is itself an interface to the Stan Bayesian computation system (Stan Development Team 2017). There are also packages that provide interfaces to the JAGS Bayesian computation system (Plummer 2003).

We will also occasionally use the `BayesFactor` package in **R** (Morey, Rouder, and Jamil 2015). This package only fits linear mixed models with normally distributed errors and a specific type of prior, but it is *very* fast, and we will need that speed in some situations.

Function `bglmm` in package `cfcdae` fits a variety of Bayesian models (the "b"), including linear models (the "lm"), linear mixed models (the second "m"), and generalized linear models (the "g"). `bglmm` assumes that coefficients for different terms in a model are independent *a priori*. Coefficients within a single model are not necessarily independent; for example, effects α_i for a categorical/factor predictor are assumed to sum to zero. The prior mean for all coefficients is assumed to be zero.

`bglmm` gives model coefficients a two-stage prior. The first stage of the prior for term A is constructed assuming that we know a standard deviation σ_A. Regression-like coefficients in term A are assumed to be normally distributed with mean 0 and standard deviation σ_A. The prior for categorical

factor effects is a bit more complicated. Each α_i is assumed to be normal with mean 0 and variance $\sigma_A^2(g-1)/g$, and the treatment effects add to zero. This looks a bit odd, but it is the distribution you get when you start with g independent outcomes with mean 0 and standard deviation σ_A and then subtract out their mean so that they sum to zero. In other words, the first stage prior for factor effects contains negative covariances that guarantee the zero sum constraint.

At the second stage, σ_A is assumed to follow a gamma distribution with a prior mean and shape parameter that the user can set. Letting σ_A vary yields an overall (marginal) prior for the coefficients with longer tails than normal. As the shape parameter gets bigger and bigger, the distribution for σ_A concentrates more and more around the prior mean. Thus you can turn the prior for coefficients into an simple, one-stage normal prior by choosing a very large shape parameter.

Two-stage prior for means

The final part of the overall prior distribution is a gamma prior on the standard deviation of the residuals.

`bglmm` lets you choose the shape and mean parameters for the standard deviations of all terms and the residuals. If you leave them unspecified, it tries to select reasonable values based on the data. Note: using the data to specify the prior is not a strictly Bayesian way of doing things. It is a particularly questionable idea if you use Bayes factors.

Example 3.8 Resin lifetimes, Bayesian model fitting with `bglmm`.

We now fit Bayesian analogues to many of the frequentist models we fit to the resin lifetime data.

```
1 > bfit.single <- bglmm(logTime ~ 1, data=myRL, quiet=TRUE)
  Need to compile the model. Compiling now.
2 > bfit.separate <- bglmm(logTime ~ temp, data=myRL, quiet=TRUE)
3 > bfit.linear <- bglmm(logTime ~ poly(temp.z,1), data=myRL,quiet=TRUE)
4 > bfit.quad <- bglmm(logTime ~ poly(temp.z,2), data=myRL, quiet=TRUE)
5 > bfit.quart <- bglmm(logTime ~ poly(temp.z,4), data=myRL, quiet=TRUE)
6 > bfit.quad.trouble <- bglmm(logTime ~ temp.z+temp.z2, data=myRL, quiet=TRUE)
  Warning messages:
  1: There were 3983 transitions after warmup that exceeded the maximum treedepth.
  ...
7 > myRL <- within(myRL, stemp <- (temp.z - mean(temp.z))/sd(temp.z))
8 > myRL <- within(myRL, stemp.2 <- stemp^2)
9 > bfit.quad.lesstrouble <- bglmm(logTime ~ stemp+stemp.2, data=myRL, quiet=TRUE)
```

The model formula for `bglmm` is the same as that for `lm`; the `quiet=TRUE` suppresses progress information printed during the MCMC sampling. By default, `bglmm` uses four chains with each having 1,000 warm up steps and 1,000 sampling steps for a total of 4,000 samples. One thing to note is the "Need to compile the model" message after line 1. Once in every **R** session, `bglmm` will need to compile a model for every family of models used; this takes a minute or so. The default family is independent, normally distributed errors with mean 0 and constant variance. All of the models we are fitting at this point fall into that category, so we only need one compilation for all of

the commands shown here.

The quadratic model is fit three times: once with orthogonal polynomials (line 4), once with ordinary polynomials (line 6), and once with ordinary polynomials (line 9) after `temp.z` was standardized (line 7). Fitting the model with raw polynomials (line 6) produces an ominous warning about exceeding maximum tree depth. This means that the algorithm is trying very hard at each iteration, but it recognizes that it is not working very well. This is also apparent in the execution times. The quadratic model with raw polynomials took 58 seconds to fit, whereas the other two quadratic models finished in under one second. Increasing the tree depth, for example, by using `max_treedepth=15` as an argument to `bglmm` will help the raw quadratic model, but the real fix is to use predictors that are all on approximately the same scale.

Before we go any further, it is important to check whether the Markov chains we are using are working well.

```
10 > summary(bfit.separate)[,9:10]
                    n_eff  Rhat
   (Intercept)      4000 1.000
   temp1            4000 1.000
   temp2            4000 1.000
   temp3            4000 0.999
   temp4            4000 1.000
   sigma.Intercept  4000 1.000
   sigma.temp       4000 1.000
   sigma0           4000 1.000
```

The summary of `bglmm` output contains diagnostic information in two of its columns. The diagnostics are the effective sample size `n_eff` (bigger is better, and close to the actual number of Markov chain samples is great), and the potential scale reduction factor `Rhat` (near 1 is good, under 1.01 is good enough). Things look fine for the separate-means model,

```
11 > summary(bfit.quad)[,9:10]
                        n_eff  Rhat
   (Intercept)          4000 0.999
   poly(temp.z, 2)1     4000 1.000
   poly(temp.z, 2)2     4000 1.000
   sigma.Intercept      4000 1.000
   sigma.poly(temp.z, 2) 4000 1.000
   sigma0               4000 1.000
```

and for the orthogonal polynomials model. However, the results are strikingly different for the raw polynomials model.
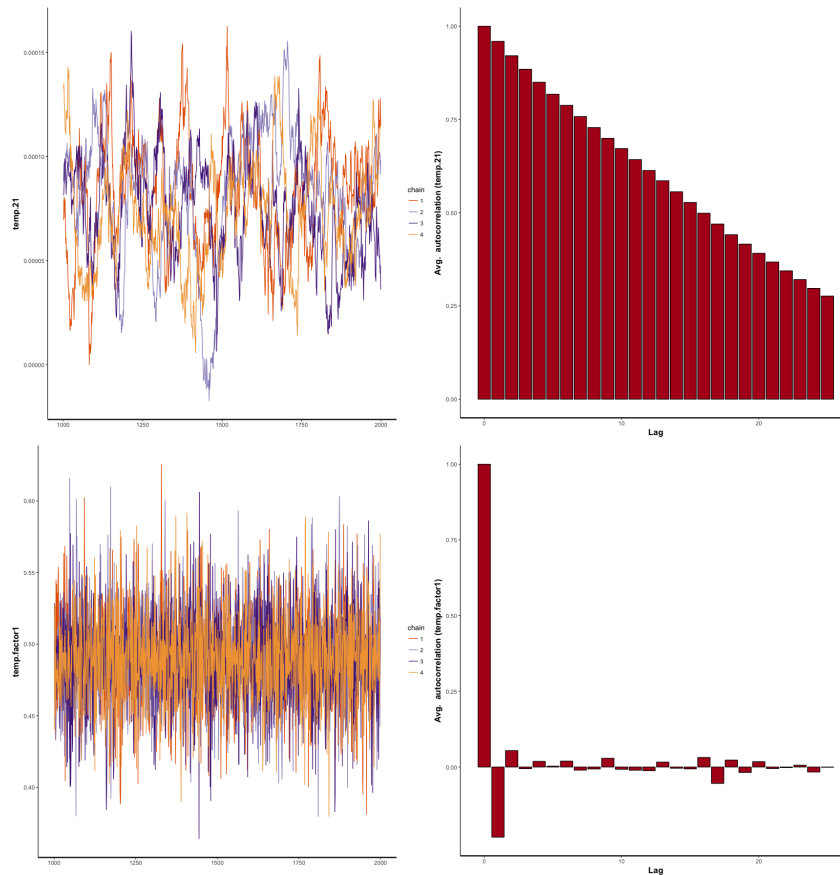
Figure 3.2: Trace and autocorrelation diagnostic plots of the quadratic coefficient in `bfit.quad.trouble` (top row) and the first group effect in `bfit.separate` (bottom row).

```
2 > summary(bfit.quad.trouble)[,9:10]
                 n_eff Rhat
(Intercept)      83.10 1.07
temp.z1          81.20 1.07
temp.z21         81.10 1.07
sigma.Intercept   7.22 1.35
sigma.temp.z      3.77 1.94
sigma.temp.z2     2.11 4.68
sigma0           83.20 1.01
```

The Markov chain for the `bfit.quad.trouble` model is performing very poorly, and we should be skeptical of results that it produces.

We should also examine the variable traces and autocorrelations.

```
13 plot(bfit.quad.trouble,plottype="trace",pars="temp.z21")
14 plot(bfit.quad.trouble,plottype="autocor",pars="temp.z21")
15 plot(bfit.separate,plottype="trace",pars="temp1")
16 plot(bfit.separate,plottype="autocor",pars="temp1")
```

Figure 3.2 shows diagnostic plots for the quadratic coefficient in the raw quadratic model and for α_1 in the separate-means model, using the commands on lines 13–16. The trace plots show the values of the four chains plotted against their index. For the separate-means coefficient, we see horizontal blurs, with no discernible trends; that is what we hope to see. Contrast that with the plot for the quadratic coefficient, where we can see clear internal trends and no horizontal blur. Autocorrelation plots show the correlation between a value in the Markov chain, and the following value, or the value two steps later, and so on. These should drop to zero as quickly as possible. In the plot we see that they effectively disappear after one lag for α_1, but they remain high for many lags for the quadratic coefficient. The separate-means model Markov chain is performing much better.

After assessing the Markov chains, we can compare models.

```
17 > loo::loo(bfit.separate)
   [1] -59.46428
   Warning message:
   Some Pareto k diagnostic values are slightly high.
   See help('pareto-k-diagnostic') for details.
18 > loo::loo(bfit.separate,verbose=TRUE)
   Computed from 4000 by 37 log-likelihood matrix

            Estimate   SE
   elpd_loo     29.7  5.3
   p_loo         6.2  1.8
   looic       -59.5 10.5

   Pareto k diagnostic values:
                             Count  Pct
   (-Inf, 0.5]   (good)        34   91.9%
    (0.5, 0.7]   (ok)           3    8.1%
      (0.7, 1]   (bad)          0    0.0%
      (1, Inf)   (very bad)     0    0.0%

   All Pareto k estimates are ok (k < 0.7)
   See help('pareto-k-diagnostic') for details.
```

The LOOCV values are computed with the `loo` function (apologies to readers in the UK), which uses something called Pareto tail smoothing to improve the computations. We see a warning after line 17 (and there are others, although it will be suppressed in the output) that things might not be working prefectly. Calling `loo` again with the `verbose=TRUE` option (line 18) provides more diagnostic information. Here we see that there is not too much to worry about, so we proceed with other models.

```
9 > loo::loo(bfit.single)
 [1] 24.69686
0 > loo::loo(bfit.linear)
 [1] -58.28534
1 > loo::loo(bfit.quad)
 [1] -64.45412
2 > loo::loo(bfit.quart)
 [1] -60.51446
3 > loo::loo(bfit.quad.trouble)
 [1] -63.97302
4 > loo::loo(bfit.quad.lesstrouble)
 [1] -64.29662
```

There is variability among LOOCV values for supposedly equivalent models, but that can be eliminated by taking more samples from the posterior (increase `n.samples` in `bglmm`). Curiously, even though the Markov chain performed poorly for the raw quadratic model, its LOOCV value is just about where it should be.

The LOOCV values support our previous conclusions from ANOVA or AICc: the quadratic model is the best choice. This is also reinforced by the Bayes factor calculation comparing the quadratic and separate-means models

```
5 > brms::bayes_factor(bfit.quad,bfit.separate)
 Iteration: 1
 ...
 The estimated Bayes factor in favor of x1 over x2 is equal to: 159.1455
```

(although, the standard warning about the Bayes factor depending strongly on the prior still holds).

`bglmm` computes samples from the posterior as part of its call, and you can review the estimation results either by printing the output or using the `summary` function. Here are the results for the separate-means model applied to the resin lifetime data.

```
5 > summary(bfit.separate)[,1:8]
                   mean   se_mean      sd    2.5%     25%     50%     75%    97.5%
 (Intercept)      1.4400 0.000272 0.0172  1.4000  1.4300  1.4400  1.4500  1.4700
 temp1            0.4900 0.000525 0.0332  0.4220  0.4680  0.4900  0.5110  0.5550
 temp2            0.1890 0.000522 0.0330  0.1240  0.1670  0.1890  0.2110  0.2520
 temp3           -0.0609 0.000514 0.0325 -0.1250 -0.0825 -0.0605 -0.0392  0.0024
 temp4           -0.2410 0.000559 0.0354 -0.3110 -0.2640 -0.2420 -0.2170 -0.1710
 sigma.Intercept  2.1100 0.019200 1.2200  0.7170  1.2500  1.7800  2.6200  5.3800
 sigma.temp       0.4780 0.003470 0.2200  0.2240  0.3340  0.4260  0.5620  1.0500
 sigma0           0.1010 0.000214 0.0135  0.0783  0.0914  0.0992  0.1080  0.1320
```

The results include for each parameter the estimated posterior mean, a measure of how precisely that mean was estimated (`se_mean`), the posterior standard deviation, and several quantiles of the posterior distribution. For $\widehat{\alpha}_1$ (`temp1`), the estimated mean is .49, the posterior standard deviation is .0332, and a 95% posterior interval runs from .422 to .555. These results are effectively identical to the frequentist estimates and confidence intervals.

You can also visualize the results in a variety of ways.
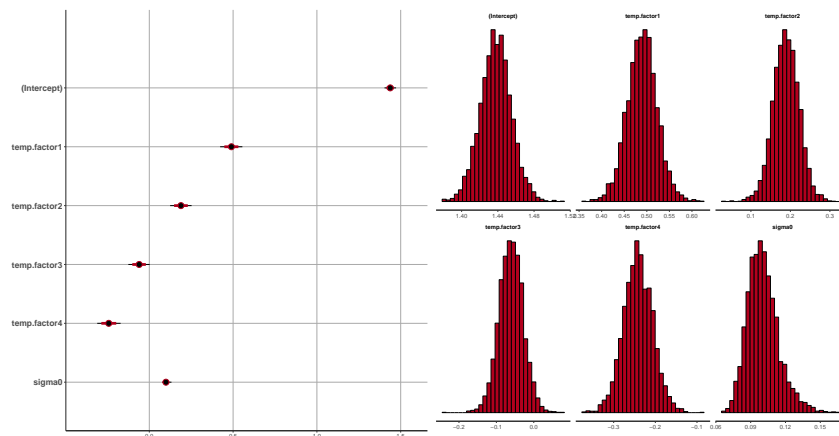
Figure 3.3: Posterior credible intervals and histograms for mean parameters and error standard deviation in the separate-means model for the resin lifetime data.

```
27 > plot(bfit.separate,plottype="interval",pars=c("internal_FullBeta","sigma0"))
   ci_level: 0.8 (80% intervals)
   outer_level: 0.95 (95% intervals)
28 > plot(bfit.separate,plottype="histogram",pars=c("internal_FullBeta","sigma0"))
```

For example, Figure 3.3 shows posterior credible intervals and histograms for the mean parameters and error standard deviation (`internal_FullBeta` is a shortcut denoting all parameters in the mean structure).

Suppose that your colleague is pretty darn sure that the standard deviation of the errors should be about 2. You can incorporate that information by using the arguments `sigma.scale0=2, gamma.shape0=100` (sigma0 being the error standard deviation, and a shape of 100 indicating a strong prior belief).

```
29 > bfit.separate.alt <- bglmm(logTime~temp,quiet=TRUE,
     sigma.scale0 =2,gamma.shape0 =100,data=myRL)
   Warning messages:
   1: There were 37 divergent transitions after warmup.
   Increasing adapt_delta above 0.8 may help
```

When we do that in line 29, we get a warning message about divergent transitions. Divergent transitions are a bad thing, and we need to try to eliminate them. The principal way to do that is by increasing `adapt_delta` from its default value of .8 toward 1. Divergent transitions can also indicate that the model should be reparameterized (as we did before by using scaled predictors or orthogonal polynomials), or they might also indicate that the model just doesn't fit very well (which is the case here).

```
) > bfit.separate.alt <- bglmm(logTime~temp,quiet=TRUE,
    sigma.scale0 =2,gamma.shape0 =100,data=myRL,adapt_delta=.995)
1 > summary(bfit.separate.alt)[,1:8]
                     mean se_mean     sd     2.5%       25%       50%     75% 97.5%
 (Intercept)       1.4200 0.00398 0.221   0.9900   1.28000   1.4200 1.5700 1.840
 temp1             0.1570 0.00730 0.259  -0.2650  -0.00965   0.1050 0.2890 0.792
 temp2             0.0513 0.00477 0.241  -0.4150  -0.07890   0.0271 0.1750 0.586
 temp3            -0.0266 0.00459 0.238  -0.5550  -0.14800  -0.0161 0.0988 0.466
 temp4            -0.0713 0.00473 0.242  -0.6150  -0.19500  -0.0399 0.0630 0.385
 sigma.Intercept  2.1300 0.02490 1.270   0.6370   1.19000   1.8000 2.6900 5.520
 sigma.temp        0.3480 0.01060 0.262   0.0379   0.15900   0.2840 0.4680 1.020
 sigma0           1.3300 0.00300 0.158   1.0400   1.22000   1.3200 1.4300 1.650
) > loo::loo(bfit.separate.alt)
 [1] 92.53691
```

After refitting with an increased `adapt_delta` in line 29, the summary after line 30 shows results very different than those of the default prior in line 1. Even though we changed the prior on the error variance and not on the mean effects, the mean effects are shrunk back toward 0. This is because our prior says that there is a lot of error variability, and as a consequence, some of the difference between treatment means is being attributed to error variability rather than treatment mean differences. However, the LOOCV shows that the model with this modified prior is a very poor predictor of the data relative to the other models we have investigated (it's even worse than the single-mean model).

We can visualize the problem with pointwise posterior predictive plots

```
2 > plot(bfit.separate.alt,plottype="pointwise")
3 > plot(bfit.separate,plottype="pointwise",ylim=c(-5,9))
```

as seen in Figure 3.4. The two plots are on the same scale, and one can easily see that the default prior produces accurate, yet much more precise, predictions.

```
34 > BayesFactor::lmBF(logTime~temp,myRL)
   Bayes factor analysis
   --------------
   [1] temp : 1.561215e+14 0%

   Against denominator:
     Intercept only
   ---
   Bayes factor type: BFlinearModel, JZS

35 > BayesFactor::lmBF(logTime~temp.z+temp.z2,myRL)
   Bayes factor analysis
   --------------
   [1] temp.z + temp.z2 : 3.15165e+16 0%

   Against denominator:
     Intercept only
   ---
   Bayes factor type: BFlinearModel, JZS

36 > 3.1516e16/1.561215e+14
   [1] 201.8684
```

Finally, consider computing the Bayes Factor using the `lmBF` function in package `BayesFactor`. By default, this function computes the Bayes factor of the fitted model relative to the single-mean model. We do this on lines 34–35 for the separate-means model and the quadratic model. Taking the ratio of those two Bayes factors on line 36 gives us the Bayes factor for the quadratic model relative to the separate-means model. The value here is not the same as we computed on line 25, but the prior distributions used by the two packages are different, so we cannot expect the Bayes factors to be the same.

Bayes Factors, LOOCV values, and estimates can all be affected by the choice of prior for the variances in the models. These effects are smallest when the priors do not provide much information about the parameters; such priors are called diffuse, vague, non-informative, and similar names. As priors become more informative and/or less congruent with the data, the effect of the prior will be seen in our inferential quantities. In general, estimated values are much less sensitive to priors as long as the priors are reasonably non-informative.

3.7 Side-by-Side Plots

Hoaglin, Mosteller, and Tukey (1991) introduce the *side-by-side* plot as a method for visualizing treatment effects and residuals. For each term in the model (including residuals but usually excluding the overall mean), we plot the values for that term in horizontal rows. Figure 3.5 shows a side-by-side plot for the resin lifetime data of Example 3.2. We plot the estimated treatment effects $\widehat{\alpha}_i$ in one row and the residuals r_{ij} in a second row. (There will be more rows in more complicated models we will see later.) The horizontal

Side-by-side plots show effects and residuals
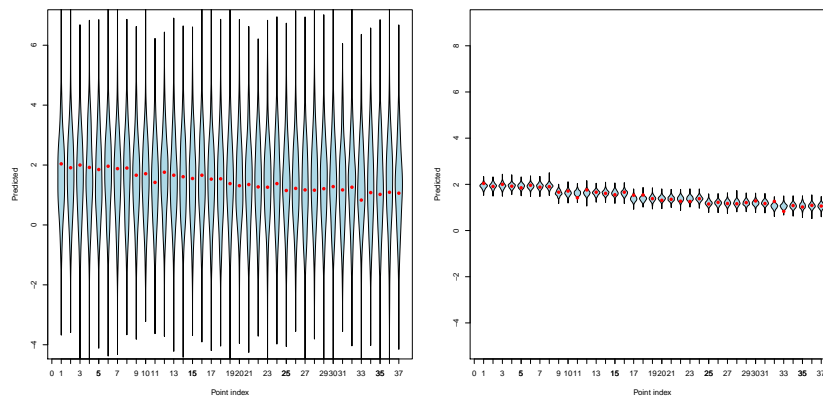
Figure 3.4: Pointwise posterior predictive intervals for each response in the resin lifetime data under the alternative prior and the default prior, plotted on the same vertical scale.

scale is in the same units as the response. In this plot, we have used a boxplot for the residuals rather than plot them individually; this will usually be more understandable when there are relatively many points to be put in a single row.

Example 3.9 Side-by-side plot for effects in the (non-Bayesian) separate-means model for resin lifetimes.

We can produce a side-by-side plot with one simple command.

```
1 > sidebyside(fit.separate)
```

What we see from the side-by-side plot is that the treatment effects are large compared to the size of the residuals. We were also able to see this in the parallel boxplots in the exploratory analysis, but the side-by-side plots will generalize better to more complicated models.

3.8   Wrap Up

The design aspects of a Completely Randomized Design are almost trivial. Once you have the treatments, units, and sample sizes, randomly assign treatments to units so that all possible assignments of treatments to units with the prescribed sample sizes are equally likely. It's just like drawing tickets from a hat. It's dead easy, but it is incredibly effective and robust. When creating an experiment, always consider the CRD first. You may, in the end, decide to use another design, but the CRD is your starting point and your reference

Side by Side Plot for Resin Lifetime Data
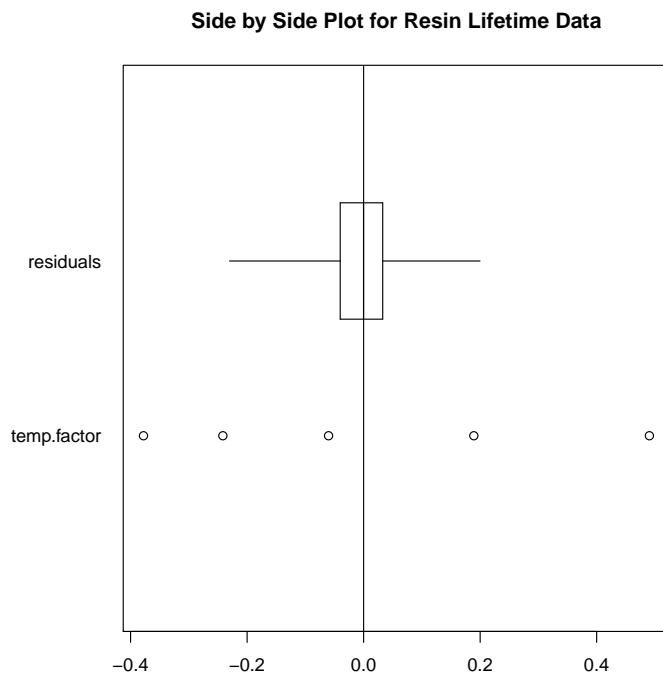
Figure 3.5: Side-by-side plot for the (non-Bayesian) separate-means model fit to the resin lifetime data.

for how good some other design might be.

Most of this chapter has been about analysis of a CRD. Some people incorrectly conflate the design and analysis aspects of an experiment; for example, I have heard someone say "I ran an ANOVA design." Do not be that person. Understand the difference between design and analysis and understand your options for analysis.

We have rather explicitly separated the idea of choosing a model from the idea of inference on parameters. At the complexity of the models of this chapter, that separation is a distinction without much of a difference. However, in later chapters we will consider more complex designs and models where there is a difference between choosing a model and testing some parameters. It is best to get used to the idea now.

3.9   Problems

Rats were given one of four different diets at random, and the response   **Exercise 3.1**

measure was liver weight as a percentage of body weight. The responses were (data set `RatLiverWeight`).

	Treatment		
1	2	3	4
3.52	3.47	3.54	3.74
3.36	3.73	3.52	3.83
3.57	3.38	3.61	3.87
4.19	3.87	3.76	4.08
3.88	3.69	3.65	4.31
3.76	3.51	3.51	3.98
3.94	3.35		3.86
	3.64		3.71

(a) Compute the overall mean and treatment effects.

(b) Compute the Analysis of Variance table for these data. What would you conclude about the four diets?

Exercise 3.2

An experimenter randomly allocated 125 male turkeys to five treatment groups: control and treatments A, B, C, and D. There were 25 birds in each group, and the mean results were 2.16, 2.45, 2.91, 3.00, and 2.71, respectively. The sum of squares for experimental error was 153.4. Test the null hypothesis that the five group means are the same against the alternative that one or more of the treatments differs from the control.

Exercise 3.3

Twelve orange pulp silage samples were divided at random into four groups of three. One of the groups was left as an untreated control, while the other three groups were treated with formic acid, beet pulp, and sodium chloride, respectively. One of the responses was the moisture content of the silage. The observed moisture contents of the silage are shown below (data from Caro *et al.* 1990, data set `OrangePulpSilage`):

	NaCl	Formic acid	Beet pulp	Control
	80.5	89.1	77.8	76.7
	79.3	75.7	79.5	77.2
	79.0	81.2	77.0	78.6
Means	79.6	82.0	78.1	77.5
Grand mean	79.3			

Compute an analysis of variance table for these data and test the null hypothesis that all four treatments yield the same average moisture contents.

Exercise 3.4

We have five groups and three observations per group. The group means are 6.5, 4.5, 5.7, 5.7, and 5.1, and the mean square for error is .75. Compute an ANOVA table for these data.

Exercise 3.5

The leaves of certain plants in the genus *Albizia* will fold and unfold in various light conditions. We have taken fifteen different leaves and subjected them to red light for 3 minutes. The leaves were divided into three groups

of five at random. The leaflet angles were then measured 30, 45, and 60 minutes after light exposure in the three groups. Data from W. Hughes, data set `Albizia`.

Delay (minutes)	Angle (degrees)				
30	140	138	140	138	142
45	140	150	120	128	130
60	118	130	128	118	118

Analyze these data to test the null hypothesis that delay after exposure does not affect leaflet angle.

Exercise 3.6

Suppose that we have a completely randomized design that has five treatments, with six units assigned to each treatment, and two measurements on each unit for a total of 60 responses. What are the degrees of freedom of the F-ratio for testing the null hypothesis that there is no treatment effect?

Problem 3.1

Cardiac pacemakers contain electrical connections that are platinum pins soldered onto a substrate. The question of interest is whether different operators produce solder joints with the same strength. Twelve substrates are randomly assigned to four operators. Each operator solders four pins on each substrate, and then these solder joints are assessed by measuring the shear strength of the pins. Data from T. Kerkow, data set `PacemakerPins`.

Strength (lb)

Operator	Substrate 1	Substrate 2	Substrate 3
1	5.60 6.80 8.32 8.70	7.64 7.44 7.48 7.80	7.72 8.40 6.98 8.00
2	5.04 7.38 5.56 6.96	8.30 6.86 5.62 7.22	5.72 6.40 7.54 7.50
3	8.36 7.04 6.92 8.18	6.20 6.10 2.75 8.14	9.00 8.64 6.60 8.18
4	8.30 8.54 7.68 8.92	8.46 7.38 8.08 8.12	8.68 8.24 8.09 8.06

Analyze these data to determine if there is any evidence that the operators produce different mean shear strengths. (Hint: what are the experimental units?)

Problem 3.2

Scientists are interested in whether the energy costs involved in reproduction affect longevity. In this experiment, 125 male fruit flies were divided at random into five sets of 25. In one group, the males were kept by themselves. In two groups, the males were supplied with one or eight receptive virgin female fruit flies per day. In the final two groups, the males were supplied with one or eight unreceptive (pregnant) female fruit flies per day. Other than the number and type of companions, the males were treated identically. The longevity of the flies was observed. Data from Hanley and Shapiro (1994), data set `FruitFlyLifespan`.

Companions	Longevity (days)												
None	35	37	49	46	63	39	46	56	63	65	56	65	70
	63	65	70	77	81	86	70	70	77	77	81	77	
1 pregnant	40	37	44	47	47	47	68	47	54	61	71	75	89
	58	59	62	79	96	58	62	70	72	75	96	75	
1 virgin	46	42	65	46	58	42	48	58	50	80	63	65	70
	70	72	97	46	56	70	70	72	76	90	76	92	
8 pregnant	21	40	44	54	36	40	56	60	48	53	60	60	65
	68	60	81	81	48	48	56	68	75	81	48	68	
8 virgin	16	19	19	32	33	33	30	42	42	33	26	30	40
	54	34	34	47	47	42	47	54	54	56	60	44	

Analyze these data to test the null hypothesis that reproductive activity does not affect longevity. Write a report on your analysis. Be sure to describe the experiment as well as your results.

Problem 3.3

Park managers need to know how resistant different vegetative types are to trampling so that the number of visitors can be controlled in sensitive areas. The experiment deals with alpine meadows in the White Mountains of New Hampshire. Twenty lanes were established, each .5 m wide and 1.5 m long. These twenty lanes were randomly assigned to five treatments: 0, 25, 75, 200, or 500 walking passes. Each pass consists of a 70-kg individual wearing lug-soled boots walking in a natural gait down the lane. The response measured is the average height of the vegetation along the lane one year after trampling. Data based on Table 16 of Cole (1993), data set `TrampledPlants`.

Number of passes	Height (cm)			
0	20.7	15.9	17.8	17.6
25	12.9	13.4	12.7	9.0
75	11.8	12.6	11.4	12.1
200	7.6	9.5	9.9	9.0
500	7.8	9.0	8.5	6.7

Analyze these data to determine if trampling has an effect after one year, and if so, describe that effect.

Problem 3.4

Caffeine is a common drug that affects the central nervous system. Among the issues involved with caffeine are how does it get from the blood to the brain, and does the presence of caffeine alter the ability of similar compounds to move across the blood-brain barrier? In this experiment, 43 lab rats were randomly assigned to one of eight treatments. Each treatment consisted of an arterial injection of C^{14}-labeled adenine together with a concentration of caffeine (0 to 50 mM). Shortly after injection, the concentration of labeled

adenine in the rat brains is measured as the response (data from McCall, Millington, and Wurtman 1982, data set `CaffeineAdenine`).

Caffeine (mM)			Adenine			
0.0	5.74	6.90	3.86	6.94	6.49	1.87
0.1	2.91	4.14	6.29	4.40	3.77	
0.5	5.80	5.84	3.18	3.18		
1	3.49	2.16	7.36	1.98	5.51	
5	5.92	3.66	4.62	3.47	1.33	
10	3.05	1.94	1.23	3.45	1.61	4.32
25	1.27	.69	.85	.71	1.04	.84
50	.93	1.47	1.27	1.13	1.25	.55

The main issues in this experiment are whether the amount of caffeine present affects the amount of adenine that can move from the blood to the brain, and if so, what is the dose response relationship. Analyze these data.

Problem 3.5

I am curious about the role of the First Year Experience course (required of all freshmen in our college) on student retention. The 2450 incoming freshmen self select into 100 groups (half with 24 students and half with 25 students). The 100 sections are divided into 4 groups of 25 at random. These four groups are assigned to the factor level combinations of medium (online versus face to face) and freedom (student choice about which units to do or no student choice). Two years later, when students would be entering their third year of college, we determine which of the 2450 students have returned for their third year (that is, are retained into the third year).

How many error degrees of freedom does this design have? Justify your answer.

Problem 3.6

Parkinson's disease appears to be caused by reduced transmission of GABA (gamma-aminobutyric acid) to the sub thalamic region of the brain. GABA is an inhibitor, so that area becomes over excited leading to the symptoms. Researchers test an experimental gene therapy. Twenty-two advanced stage Parkinson's patients receive the gene therapy, and 23 other patients receive a placebo therapy. From a pre-treatment average score of 25, the gene therapy group improved by 8 points, and the control group improved by 4 points. The four point difference between the gene therapy and the control group is statistically significant (using an ordinary two-sample t-test).

(a) Is this a randomized experiment?

(b) Is the four point improvement by the control group a statistically significant improvement from baseline? Why or why not?

(c) This experiment is an example of what well known experimental phenomenon?

Problem 3.7

Engineers wish to know the effect of polypropylene fibers on the com-

pressive strength of concrete. Fifteen concrete cubes are produced and randomly assigned to five levels of fiber content (0, .25, .50, .75, and 1%). Data from Figure 2 of Paskova and Meyer (1997), data set `ConcreteStrength`.

Fiber content (%)	Strength (ksi)		
0	7.8	7.4	7.2
.25	7.9	7.5	7.3
.50	7.4	6.9	6.3
.75	7.0	6.7	6.4
1	5.9	5.8	5.6

Analyze these data to determine if fiber content has an effect on concrete strength, and if so, describe that effect.

Problem 3.8

Under the right conditions, cells in a fibrin solution will deposit on a mold and create a bio-artificial vascular graft (a manufactured patch for an artery). If the cells are subjected to mechanical stress during deposition, they respond by producing more extra-cellular matrix, which increases the mechanical strength of the graft. One sensible way to produce the stress is to sleeve the graft over a latex tube and cyclically distend the tube with air. The problem is that the apparatus needs to be sterile, the convenient method for sterilizing the latex tube is autoclaving, and the heat of autoclaving will modify the elastic properties of the latex.

This experiment examines the effects of autoclaving on the elastic modulus (in kPa) of the latex. The adjustable factors for the autoclave are the temperature (121 C or 135 C) and the time (10 minutes or 20 minutes). We look at five treatments: Control (no autoclaving), and the factorial combinations of time and temperature, each at the two levels given above. Fifteen latex tube samples are randomly assigned to the five treatments, and the modulus is then observed after treatment. Large values of the modulus are good. Data follow (from Z. Syedain, data set `Autoclaving`).

Treatment	modulus (kPa)		
Control	1117.5	1076.2	951.1
121°, 10 min	732.8	750.3	707.8
121°, 20 min	596.0	648.6	713.9
135°, 10 min	565.4	623.4	608.0
135°, 20 min	510.9	664.6	484.9

Analyze these data to determine if autoclaving affects the modulus of elasticity.

Problem 3.9

Ninety-three student volunteers are told that they will be having a conversation with a member of their same gender. The students are randomly assigned to three treatments, 31 students per treatment. One group of students will be told that their conversation partner is is an extrovert, a second

group is told that their partner is an introvert, and the third group is given no information. Prior to beginning the conversations, subjects fill out a questionaire.

In fact, there is no conversation; the quantities derived from the pre-conversation questionaire are the responses. The response shown here is "perception of power" in the upcoming conversation, a scale derived from 18 items in the questionaire. Higher responses on this scale indicate perceptions of greater power. (Synthetic data, data set `PowerPerception`.)

Information	Perceived power							
Extrovert	3.41	3.11	3.06	3.95	4.23	2.62	3.42	4.35
	3.77	2.65	3.62	2.73	4.25	3.47	3.14	4.54
	3.44	3.83	4.67	3.84	3.47	2.80	4.75	4.00
	3.70	3.59	4.61	4.98	4.02	2.65	4.99	
Introvert	4.60	3.79	4.14	4.39	4.34	3.91	3.51	3.17
	3.93	5.68	3.01	3.26	5.13	3.85	2.98	4.03
	4.28	6.41	2.54	4.93	4.78	3.60	3.67	3.92
	3.73	5.13	4.09	3.89	4.38	4.20	5.06	
None	2.22	2.36	1.52	3.08	2.50	4.68	2.16	3.29
	3.23	2.14	3.79	3.53	3.14	2.50	2.25	4.97
	2.24	3.63	5.54	3.91	3.71	3.81	3.50	2.74
	4.07	2.47	2.69	2.49	3.65	2.71	3.43	

Analyze these data to determine if the expectation of certain personality characteristics in the conversation partner affects the perception of power.

Prove that $\mu^\star = \sum_{i=1}^g \mu_i/g$ is equivalent to $\sum_{i=1}^g \alpha_i = 0$.   **Question 3.1**

Prove that   **Question 3.2**

$$0 = \sum_{i=1}^g \sum_{j=1}^{n_i} \widehat{\alpha}_i r_{ij} \quad .$$

Chapter 4

Looking for Specific Differences—Contrasts

An Analysis of Variance can give us an indication that not all the treatment groups have the same mean response, but an ANOVA does not, by itself, tell us which treatments are different or in what ways they differ. To do this, we need to look at the treatment means, or equivalently, at the treatment effects. One method to examine treatment effects is called a *contrast*.

ANOVA is like background lighting that dimly illuminates all of our data, but not giving enough light to see details. Using a contrast is like using a spotlight; it enables us to focus in on a specific, narrow feature of the data. But the contrast has such a narrow focus that it does not give the overall picture. By using several contrasts, we can move our focus around and see more features. Intelligent use of contrasts involves choosing our contrasts so that they highlight interesting features in our data.

4.1 Contrast Basics

Contrasts take the form of a difference between means or averages of means. For example, here are two contrasts:

$$(\mu + \alpha_6) - (\mu + \alpha_3)$$

and

$$\frac{\mu + \alpha_2 + \mu + \alpha_4}{2} - \frac{\mu + \alpha_1 + \mu + \alpha_3 + \mu + \alpha_5}{3} \quad .$$

The first compares the means of treatments 6 and 3, while the second compares the mean response in groups 2 and 4 with the mean response in groups 1, 3, and 5.

Formally, a *contrast* is a linear combination of treatment means or effects $\sum_{i=1}^{g} w_i \mu_i = w(\{\mu_i\})$ or $\sum_{i=1}^{g} w_i \alpha_i = w(\{\alpha_i\})$, where the coefficients w_i satisfy $\sum_{i=1}^{g} w_i = 0$.

> Contrast coefficients add to zero.

Less formally, we sometimes speak of the set of contrast coefficients $\{w_i\}$ as being a contrast; we will try to avoid ambiguity. Notice that because the sum of the coefficients is zero, we have that

$$
\begin{aligned}
w(\{\alpha_i\}) &= \sum_{i=1}^{g} w_i \alpha_i = \mu \sum_{i=1}^{g} w_i + \sum_{i=1}^{g} w_i \alpha_i \\
&= \sum_{i=1}^{g} w_i(\mu + \alpha_i) = w(\{\mu_i\}) \ .
\end{aligned}
$$

(You can replace μ with any constant, but μ gives us the link to treatment means.) We may also make contrasts in the observed data:

$$
w(\{\overline{y}_{i\bullet}\}) = \sum_{i=1}^{g} w_i \overline{y}_{i\bullet} = \sum_{i=1}^{g} w_i(\overline{y}_{i\bullet} - \sum_{i=1}^{g} \overline{y}_{i\bullet}/g) = \sum_{i=1}^{g} w_i \widehat{\alpha}_i = w(\{\widehat{\alpha}_i\}) \ .
$$

The $\sum \overline{y}_{i\bullet}/g$ can be replaced with any other definition of $\widehat{\mu}$, and we find that the contrast in the treatment effects equals the contrast in the treatment means, regardless of how $\widehat{\mu}$ is defined, and thus regardless of how the treatment effects are constrained. Put another way, a contrast depends on the differences between the values being contrasted, but not on the overall level of the values. Recall that with respect to restrictions on the treatment effects, we said that "the important things don't depend on which set of restrictions we use." In particular, contrasts don't depend on the restrictions.

Contrasts do not depend on α-restrictions

> Contrasts do not depend on how treatment effects are parameterized.

We may use several different kinds of contrasts in any one analysis. The trick is to use contrasts that focus on interesting questions. These questions will differ from situation to situation.

Probably the most common contrasts are *pairwise comparisons*, where we contrast the mean response in one treatment with the mean response in a second treatment. For a pairwise comparison, one contrast coefficient is 1, a second contrast coefficient is -1, and all other contrast coefficients are 0. For example, in an experiment with $g = 4$ treatments, the coefficients $(0, 1, -1, 0)$ compare the means of treatments 2 and 3, and the coefficients $(-1, 0, 1, 0)$ compare the means of treatments 1 and 3. For g treatments, there are $g(g-1)/2$ different pairwise comparisons. We will consider simultaneous inference for pairwise comparisons in Section 5.4.

Pairwise comparisons

A second classic example of contrasts occurs in an experiment with a control and two or more new treatments. Suppose that treatment 1 is a control, and treatments 2 and 3 are new treatments. We might wish to compare the average response in the new treatments to the average response in the control; that is, on average do the new treatments have the same response as

the control? Here we could use coefficients $(-1, .5, .5)$, which would subtract the average control response from the average of treatments 2 and 3's average responses. As discussed below, this contrast applied to the observed treatment means $((\overline{y}_{2\bullet} + \overline{y}_{3\bullet})/2 - \overline{y}_{1\bullet})$ would estimate the contrast in the treatment effects $((\alpha_2 + \alpha_3)/2 - \alpha_1)$. Note that we would get the same kind of information from contrasts with coefficients $(1, -.5, -.5)$ or $(-6, 3, 3)$; we've just rescaled the result with no essential loss of information. We might also be interested in the pairwise comparisons, including a comparison of the new treatments to each other $(0, 1, -1)$ and comparisons of each of the new treatments to control $(1, -1, 0)$ and $(1, 0, -1)$.

<div align="right">Control versus
other treatments</div>

Another common form of contrast is useful in situations where the treatments are naturally grouped. Consider next an experiment with four treatments examining the growth rate of lambs. The treatments are four different food supplements. Treatment 1 is soy meal and ground corn, treatment 2 is soy meal and ground oats, treatment 3 is fish meal and ground corn, and treatment 4 is fish meal and ground oats. Again, there are many potential contrasts of interest. A contrast with coefficients $(.5, .5, -.5, -.5)$ would take the average response for fish meal treatments and subtract it from the average response for soy meal treatments. This could tell us about how the protein source affects the response. Similarly, a contrast with coefficients $(.5, -.5, .5, -.5)$ would take the average response for ground oats and subtract it from the average response for ground corn, telling us about the effect of the carbohydrate source.

<div align="right">Compare related
groups of
treatments</div>

Finally, consider an experiment with three treatments examining the effect of development time on the number of defects in computer chips produced using photolithography. The three treatments are 30, 45, and 60 seconds of developing. If we think of the responses as lying on a straight line function of development time, then the contrast with coefficients $(-1/30, 0, 1/30)$ will estimate the slope of the line relating response and time. If instead we think that the responses lie on a quadratic function of development time, then the contrast with coefficients $(1/450, -2/450, 1/450)$ will estimate the quadratic term in the response function. The use of contrasts to estimate or test polynomial terms is now mostly historical, because modern software allows us to easily refit with polynomial predictors rather than extract the information from a separate means model using contrasts.

<div align="right">Polynomial
contrasts for
quantitative
doses</div>

Two contrasts $\{w\}$ and $\{w^\star\}$ are said to be *orthogonal* if

$$\sum_{i=1}^{g} w_i w_i^\star / n_i = 0 \ \ .$$

If there are g treatments, you can find a set of $g-1$ contrasts that are mutually orthogonal, that is, each one is orthogonal to all of the others. However, there are infinitely many sets of $g - 1$ mutually orthogonal contrasts, and there are no mutually orthogonal sets with more than $g-1$ contrasts. There is an analogy from geometry. In a plane, you can have two lines that are perpendicular (orthogonal), but you *can't* find a third line that is perpendicular to both of the others. On the other hand, there are infinitely many pairs of perpendicular lines.

<div align="right">$g - 1$ orthogonal
contrasts</div>

Orthogonal contrasts have two properties that can provide modest advantages. First, orthogonal contrasts applied to observed means are independent (as random variables). Thus, the random error of one contrast is not correlated with the random error of an orthogonal contrast. Second, a complete set of orthogonal contrasts partitions the between groups sum of squares. That is, if you compute the sums of squares for a full set of orthogonal contrasts ($g-1$ contrasts for g groups), then adding up those $g-1$ sums of squares will give you exactly the between groups sum of squares (which also has $g-1$ degrees of freedom).

Orthogonal contrasts are independent and partition variation

In any experimental situation, you should use the contrasts that address meaningful questions. If they do not happen to be orthogonal, that is alright. It is *much* more important to address the correct questions than it is to be orthogonal.

> Use contrasts that address the questions you are trying to answer.

4.2 Standard Frequentist Inference for Contrasts

We use contrasts in observed treatment means or effects to make inference about the corresponding contrasts in the true treatment means or effects. The kinds of inference we work with here are point estimates, confidence intervals, and tests of significance. The procedures we use for contrasts are similar to the procedures we use when estimating or testing means.

The observed treatment mean $\overline{y}_{i\bullet}$ is an unbiased estimate of $\mu_i = \mu + \alpha_i$, so a sum or other linear combination of observed treatment means is an unbiased estimate of the corresponding combination of the μ_i's. In particular, a contrast in the observed treatment means is an unbiased estimate of the corresponding contrast in the true treatment means. Thus we have:

$w(\{\overline{y}_{i\bullet}\})$ estimates $w(\{\mu_i\})$

$$E[w(\{\overline{y}_{i\bullet}\})] = E[w(\{\widehat{\alpha}_i\})] = w(\{\mu_i\}) = w(\{\alpha_i\}) \ .$$

The variance of $\overline{y}_{i\bullet}$ is σ^2/n_i, and the treatment means are independent, so the variance of a contrast in the observed means is

$$\mathrm{Var}\left[w(\{\overline{y}_{i\bullet}\})\right] = \sigma^2 \sum_{i=1}^{g} \frac{w_i^2}{n_i} \ .$$

We will usually not know σ^2, so we estimate it by the mean square for error from the ANOVA.

We compute a confidence interval for a mean parameter with the general form: *unbiased estimate $\pm$ t-multiplier $\times$ estimated standard error*. Contrasts are linear combinations of mean parameters, so we use the same basic form. We have already seen how to compute an estimate and standard error, so

Confidence interval for $w(\{\mu_i\})$

$$w(\{\overline{y}_{i\bullet}\}) \pm \ t_{\mathcal{E}/2, N-g} \ \sqrt{\mathrm{MS_E}} \ \sqrt{\sum_{i=1}^{g} \frac{w_i^2}{n_i}}$$

forms a $1 - \mathcal{E}$ confidence interval for $w(\{\mu_i\})$. As usual, the degrees of freedom for our t-percent point come from the degrees of freedom for our estimate of error variance, here $N - g$. We use the $\mathcal{E}/2$ percent point because we are forming a two-sided confidence interval, with $\mathcal{E}/2$ error on each side.

The usual t-test statistic for a mean parameter takes the form

$$\frac{\textit{unbiased estimate} - \textit{null hypothesis value}}{\textit{estimated standard error of estimate}} .$$

This form also works for contrasts. If we have the null hypothesis $H_0\colon w(\{\mu_i\}) = \delta$, then we can do a t-test of that null hypothesis by computing the test statistic

$$t = \frac{w(\{\overline{y}_{i\bullet}\}) - \delta}{\sqrt{\mathrm{MS_E}} \sqrt{\sum_{i=1}^{g} \frac{w_i^2}{n_i}}} .$$

Under H_0, this t-statistic will have a t-distribution with $N - g$ degrees of freedom. Again, the degrees of freedom come from our estimate of error variance. The p-value for this t-test is computed by getting the area under the t-distribution with $N - g$ degrees of freedom for the appropriate region: either less or greater than the observed t-statistic for one-sided alternatives, or twice the tail area for a two-sided alternative.

t-test for $w(\{\mu_i\})$

We may also compute a sum of squares for any contrast $w(\{\overline{y}_{i\bullet}\})$:

$$\mathrm{SS_w} = \frac{(\sum_{i=1}^{g} w_i \overline{y}_{i\bullet})^2}{\sum_{i=1}^{g} \frac{w_i^2}{n_i}} .$$

This sum of squares has 1 degree of freedom, so its mean square is $\mathrm{MS_w} = \mathrm{SS_w}/1 = \mathrm{SS_w}$. We may use $\mathrm{MS_w}$ to test the null hypothesis that $w(\{\mu_i\}) = 0$ by forming the F-statistic $\mathrm{MS_w}/\mathrm{MS_E}$. If H_0 is true, this F-statistic will have an F-distribution with 1 and $N - g$ degrees of freedom ($N - g$ from the $\mathrm{MS_E}$). It is not too hard to see that this F is exactly equal to the square of the t-statistic computed for same null hypothesis $\delta = 0$. Thus the F-test and two-sided t-tests are equivalent for the null hypothesis of zero contrast mean. It is also not too hard to see that if you multiply the contrast coefficients by a nonzero constant (for example, change from $(-1, .5, .5)$ to $(2, -1, -1)$), then the contrast sum of squares is unchanged. The squared constant cancels from the numerator and denominator of the formula.

SS and F-test for $w(\{\mu_i\})$

Example 4.1 Rat liver weights

Exercise 3.1 provided data on the weight of rat livers as a percentage of body weight for four different diets. Summary statistics from those data follow:

i	1	2	3	4	
$\overline{y}_{i\bullet}$	3.75	3.58	3.60	3.92	
n_i	7	8	6	8	$\mathrm{MS_E} = .04138$

If diets 1, 2, and 3 are rations made by one manufacturer, and diet 4 is a

ration made by a second manufacturer, then it may be of interest to compare the responses from the diets of the two manufacturers to see if there is any difference. Further, if ration 1 is an "premium" ration, and rations 2 and3 are standard, then it may be of interest to compare the premium ration from the first manufacturer to the standard rations from the first manufacturer.

The contrast with coefficients $(1/3, 1/3, 1/3, -1)$ will compare the mean response in the first three diets (manufacturer 1) with the mean response in the last diet (manufacturer 2). Note that we intend "the mean response in the first three diets" to denote the average of the treatment averages, not the simple average of all the data from those three treatments. The simple average will not be the same as the average of the averages because the sample sizes are different.

Our point estimate of this contrast is

$$w(\{\bar{y}_{i\bullet}\}) = \frac{1}{3}3.75 + \frac{1}{3}3.58 + \frac{1}{3}3.60 + (-1)3.92 = -.277$$

with standard error

$$SE(w(\{\bar{y}_{i\bullet}\})) = \sqrt{.04138}\sqrt{\frac{(\frac{1}{3})^2}{7} + \frac{(\frac{1}{3})^2}{8} + \frac{(\frac{1}{3})^2}{6} + \frac{(-1)^2}{8}} = .0847 \quad .$$

The mean square for error has $29 - 4 = 25$ degrees of freedom. To construct a 95% confidence interval for $w(\{\mu_i\})$, we need the upper 2.5% point of a t-distribution with 25 degrees of freedom; this is 2.06, as can be found in Appendix Table C.3 or using software. Thus our 95% confidence interval is

$$-.277 \pm 2.06 \times .0847 = -.277 \pm .174 = (-.451, -.103) \quad .$$

Suppose that we wish to test the null hypothesis $H_0\colon w(\{\mu_i\}) = \delta$. Here we will use the t-test and F-test to test $H_0\colon w(\{\mu_i\}) = \delta = 0$, but the t-test can test other values of δ. Our t-test is

$$\frac{-.277 - 0}{.0847} = -3.27 \quad,$$

with 25 degrees of freedom. For a two-sided alternative, we compute the p-value by finding the tail area under the t-curve and doubling it. Here we get twice .00156 or about .003. This is fairly strong evidence against the null hypothesis.

Because our null hypothesis value is zero with a two-sided alternative, we can also test our null hypothesis by computing a mean square for the contrast and forming an F-statistic. The sum of squares for our contrast is

$$\frac{(\frac{1}{3}3.75 + \frac{1}{3}3.58 + \frac{1}{3}3.60 + (-1)3.92)^2}{\frac{(1/3)^2}{7} + \frac{(1/3)^2}{8} + \frac{(1/3)^2}{6} + \frac{(-1)^2}{8}} = \frac{(-.277)^2}{.1733} = .443 \quad .$$

The mean square is also .443, so the F-statistic is .443/.04138 = 10.7. We compute a p-value by finding the area to the right of 10.7 under the F-distribution with 1 and 25 degrees of freedom, getting .003 as for the t-test.

Of course, we are not likely to be doing these contrast calculations by hand. The `linear.contrast` function does the work for us.

```
1 > fit1 <- lm(weight ~ diet, data=RatLiverWeight)
2 > linear.contrast(fit1,diet,c(1/3,1/3,1/3,-1))
     estimates         se  t-value    p-value   lower-ci   upper-ci
 1 -0.2811508 0.08467399 -3.320392 0.00276231 -0.4555401 -0.1067615
3 > linear.contrast(fit1,diet,c(1,-.5,-.5,0))
   estimates          se  t-value   p-value    lower-ci  upper-ci
 1 0.1565476 0.09448756 1.656807 0.1100575 -0.03805315 0.3511484
4 > linear.contrast(fit1,diet,cbind(c(1/3,1/3,1/3,-1),c(1,-.5,-.5,0)))
     estimates          se  t-value    p-value    lower-ci   upper-ci
 1 -0.2811508 0.08467399 -3.320392 0.00276231 -0.45554013 -0.1067615
 2  0.1565476 0.09448756  1.656807 0.11005747 -0.03805315  0.3511484
```

Line 1 fits the linear model. Line 2 uses `linear.contrast` to calculate the inferential quantities for the contrast with coefficients $(1/3, 1/3, 1/3, -1)$. There is more roundoff error in our hand calculations than is typical, so perhaps we should have carried more digits. The inference remains the same, however. Note that you can change the coverage of the interval estimate using the `confidence` argument.

Line 3 uses the contrast with coefficients $(1, -.5, -.5, 0)$ to compare the premium to standard rations within manufacturer 1. In this case, the p-value is .11, so there is little evidence against the null hypothesis that the premium ration delivers the same response as the standard rations. Line 4 merely shows that one can do both contrasts in a single command by creating a matrix with a column for each set of contrast coefficients.

4.3   Bayesian Inference for Contrasts

Usual Bayesian inference for a contrast would be the posterior mean, posterior standard deviation, and a posterior interval estimate. These are easily computed from the MCMC samples from the posterior by applying the contrast to each sample.

In addition, one can do model comparison. In this case, the models being compared would be the unrestricted model and a model where we constrain the estimated treatment effects in such a way that the contrast value will be exactly zero. This constrained model plays the role of the "null" model. The constrained and unconstrained models can then be compared via LOOCV or Bayes factor. Unfortunately, this means that we need to refit a constrained Bayesian model for every contrast we wish to consider.

Example 4.2 Bayesian contrast analysis of rat liver weights

We begin by fitting the Bayesian model in line 1, and then use the `linear.contrast` function in line 2.

```
 1 > fit2 <- bglmm(weight~diet,data=RatLiverWeight,adapt_delta = .99)
 2 > linear.contrast(fit2,diet,cbind(c(1/3,1/3,1/3,-1),c(1,-.5,-.5,0)))
     post. mean    post. sd        lower        upper approx BF
   1 -0.2299550 0.08994509 -0.40264604 -0.04355175 12.224385
   2  0.1258293 0.09543917 -0.05953985  0.31158403  1.110057
 3 > fit2c1 <- bglmm(weight~diet,data=RatLiverWeight,adapt_delta = .99,
     add.constraints = list(diet=c(1/3,1/3,1/3,-1)), show.redundant.effects = TRUE)
 4 > fit2c2 <- bglmm(weight~diet,data=RatLiverWeight,adapt_delta = .99,
     add.constraints = list(diet=c(1,-.5,-.5,0)),show.redundant.effects = TRUE)
 5 > summary(fit2c2)[,c(1,3)]
                       mean     sd
   (Intercept)      3.7100 0.0419
   diet1           -0.0587 0.0249
   diet2           -0.0728 0.0600
   diet3           -0.0446 0.0594
   diet4            0.1760 0.0747
   sigma.Intercept  4.9100 2.7600
   sigma.diet       0.2950 0.2170
   sigma0           0.2220 0.0322
 6 > loo(fit2c1)
   [1] 3.845055
 7 > loo(fit2c2)
   [1] -2.940697
 8 > loo(fit2)
   [1] -3.602275
 9 > bayes_factor(fit2,fit2c1)
   The estimated Bayes factor in favor of x1 over x2 is equal to: 11.7724
10 > bayes_factor(fit2,fit2c2)
   The estimated Bayes factor in favor of x1 over x2 is equal to: 0.9501
11 > linear.contrast(fit2,diet,cbind(c(1/3,1/3,1/3,-1),c(1,-.5,-.5,0)),exactBF=TRUE)
     post. mean    post. sd       lower       upper         BF
   1 -0.2261716 0.09495031 -0.4050476 -0.0322537 12.0078787
   2  0.1240304 0.09312277 -0.0528093  0.3054965  0.9773381
```

For a Bayesian model, the output of `linear.contrast` includes the posterior mean, posterior standard deviation, and a posterior interval estimate. It also includes an *approximate* Bayes factor for comparing the unrestricted model to the model with the coefficients constrained to have contrast value zero. In this case, the (approximate) Bayes factor shows a postive preference for the unconstrained model over the model that requires the mean of the first three treatments to equal the mean of the third treatment. For the second contrast, the (approximate) Bayes factor is close enough to 1 that it is difficult to have much of a preference for full model over the constrained model (or vice versa).

Lines 3 and 4 fit two constrained models. In line 3, we add constraints so that the coefficients of the term `diet` satisfy the contrast (1/3, 1/3, 1/3, −1), that is, this contrast will always be 0. Line 4 is analogous for the contrast (1,−.5,−.5,0). Note that in addition to adding the constraint to the function call, we also ask `bglmm` to return all the coefficients, even the redundant (unneeded mathematically) coefficients. We use this in line 5, where we ask for a summary of model results for the second constrained model. Because we asked to see redundant coefficients, all the coefficients show up in the summary. In particular, we can see that the average of the posterior means for diets 2 and 3 is the same as the posterior mean for diet 1.

Lines 6, 7, and 8 compute LOOCV values for our three models, and lines 9 and 10 compute Bayes factors comparing the unconstrained model to each of the constrained models. The inference from the actual Bayes factors agrees with that of the approximate Bayes factors, although the values are not that similar. The Bayes factor ever so slightly prefers the model with the contrast $(1, -.5, -.5, 0)$ constrained to be zero. LOOCV, however, prefers the full, unconstrained model. The is an example of a predictive criterion like LOOCV preferring larger models than a more stringent criterion such as the Bayes factor.

The approximation used in `linear.contrast` to compute the Bayes factor is asymptotic, meaning that it works better for larger data sets. The `linear.contrast` function also has an option `exactBF=TRUE` (as shown in line 11), in which case it will refit the model constraining each contrast and then compute the Bayes factor. This is, of course, much slower than the approximate method.

4.4 Further Reading and Extensions

Contrasts are a special case of *estimable functions*, which are described in some detail in Appendix Section A.6. Treatment means and averages of treatment means are other estimable functions. Estimable functions are those features of the data that do not depend on how we choose to restrict the treatment effects.

4.5 Problems

Use the data from Exercise 3.3. Compute a 99% confidence interval for the difference in response between the average of the three treatment groups (acid, pulp, and salt) and the control group.

Exercise 4.1

Refer to the data in Problem 3.1. Workers 1 and 2 were experienced, whereas workers 3 and 4 were novices. Find a contrast to compare the experienced and novice workers and test the null hypothesis that experienced and novice works produce the same average shear strength.

Exercise 4.2

Consider an experiment taste-testing six types of chocolate chip cookies: 1 (brand A, chewy, expensive), 2 (brand A, crispy, expensive), 3 (brand B, chewy, inexpensive), 4 (brand B, crispy, inexpensive), 5 (brand C, chewy, expensive), and 6 (brand D, crispy, inexpensive). We will use twenty different raters randomly assigned to each type (120 total raters).

Exercise 4.3

(a) Design contrasts to compare chewy with crispy, and expensive with inexpensive.

(b) Are your contrasts in part (a) orthogonal? Why or why not?

The resistance of a wood product to the flow of electricity depends on

Exercise 4.4

the moisture content of the product. Thus a measure of resistance is sometimes used to measure moisture content. Consider an experiment with six treatments: particle board at low moisture, particle board at medium moisture, particle board at high moisture, plywood at low moisture, plywood at medium moisture, and solid fir at low moisture. Resistance is measured several times for each treatment.

Construct a contrast comparing low moisture to medium moisture, and describe why your contrast is good.

Exercise 4.5

I have a completely randomized design with 40 observations, 10 in each of four treatment groups. The treatments are different temperatures: 90, 100, 110, and 120 degrees C. The sum of squares between treatments is 250. The sum of squares for the contrast with coefficients $(-3, 1, 1, 1)$ is 125, and the sum of squares for the contrast with coefficients $(0, -2, 1, 1)$ is 80. What is the sum of squares for the contrast with coefficients $(0, 0, 1, -1)$? Justify your answer.

Problem 4.1

A consumer testing agency obtains four cars from each of six makes: Ford, Chevrolet, Nissan, Lincoln, Cadillac, and Mercedes. Makes 3 and 6 are imported while the others are domestic; makes 4, 5, and 6 are expensive while 1, 2, and 3 are less expensive; 1 and 4 are Ford products, while 2 and 5 are GM products. We wish to compare the six makes on their oil use per 100,000 miles driven. The mean responses by make of car were 4.6, 4.3, 4.4, 4.7, 4.8, and 6.2, and the sum of squares for error was 2.25.

Design a set of contrasts that seem meaningful. For each contrast, outline its purpose.

Problem 4.2

Consider the data in Problem 3.2. Design a set of contrasts that seem meaningful. For each contrast, outline its purpose and test the null hypothesis that the contrast has expected value zero.

Problem 4.3

One-hundred thirteen people were randomly assigned into five groups. Each group will receive some information about a political campaign, and then make a determination about whether the campaign has been using positive or negative advertising. They make that rating on a 1 to 7 scale, with 1 being most positive and 7 being most negative. Group one receives the transcript of a television ad that attacks an opposing candidate. Group two receives the transcript of the ad plus an editorial describing the campaign as generally positive. Group three receives the transcript of the ad plus an editorial describing the campaign as generally negative. Groups four and five receive only the positive and negative editorials, respectively. Data follow (data set `cfcdae::PoliticalAds`).

Trans. only	4	2	6	4	6	3	4	5	4	4	6	6	3	4
	4	4	4	2	4	6	4	5	4					
Trans. & pos. ed.	4	5	2	4	6	5	5	2	4	5	5	5	5	4
	5	6	4	5	5	3	3	4	6	2	5	4	3	
Trans. & neg. ed.	7	5	6	4	3	5	7	7	6	5	7	5	5	7
	6	5	7	7	6	4	5	6	5					
Pos. ed. only	7	3	3	5	5	4	4	4	4	3	2	4	3	5
	2	4	3	2	4	1	4							
Neg. ed. only	3	5	7	5	6	3	5	6	5	5	4	6	4	7
	7	7	6	6	6									

Does media coverage affect perception? Does the actual transcript change the response to media reporting? Assuming the editorials affect preception, does positive reporting improve perception as much as negative reporting decreases it?

Problem 4.4

Cakes can be baked either one at a time on the top or bottom oven rack, or two at a time, with one on the top rack and one on the bottom rack. We bake eight cakes, with two cake mix boxes randomly assigned to each of the four treatments: top rack single, bottom rack single, top rack double, bottom rack double. After the cake is baked and allowed to cool for 1 hour, its height (mm) is measured at five locations, with the average of the five measurements taken as the height for the cake. Data follow (data from D. Schendel, data set `cfcdae:CakeHeights`).

Brand	Height (mm)	
Top, single	50.6	49.2
Bottom, single	46.0	46.4
Top, double	48.4	47.3
Bottom, double	45.1	46.1

Use contrasts to compare the top rack heights to the bottom rack heights, and the single cake heights to the double cake heights.

Problem 4.5

Everyone likes a strong, durable paper towel (absorbency is another desirable property, but this experiment is about durability). Five rolls of paper towels are purchased, one each from five brands (B, V, S, S2, and T). The first three brands are name brands, and the last two are store brands. From each roll, we randomly select three towels. In random order, each towel is dipped in water until wet, gently squeezed by hand to remove absorbed water, then spread across the top of a bowl and held in place by clothespins. Then, pennies are gently placed on the suspended towel until the towel breaks. The number of pennies until the towel is the measure of strength. Data follow (data from A. Frosch, data set `cfcdae::TowelStrength`).

Brand	Pennies		
B	145	119	162
V	159	170	133
S	73	80	74
S2	84	89	94
T	138	125	140

Use contrasts to (a) test the null hypothesis that name brands have the same strength as store brands, and (b) test the null hypothesis that brands coded with an S have the same strength as brands not coded with an S.

In an experiment with three treatments and samples sizes $n_1 = 15, n_2 = 10, n_3 = 10$, the sum of squares for treatments is 100 with 2 degrees of freedom. We also know that the sum of squares for the contrast with coefficients $(2, -1, -1)$ is 80.

Problem 4.6

Consider the contrast with coefficients $(0, 1, -1)$. Can you determine the sum of squares for this contrast? If so, compute the sum of squares and explain why your computation is correct. If not, explain why the information given is inadequate.

Show that under our assumptions orthogonal contrasts in the observed treatment means are uncorrelated random variables.

Question 4.1

Chapter 5

Multiple Comparisons

When we make several related tests or interval estimates at the same time, frequentist analysis must compensate. This is called making *multiple comparisons* or doing *simultaneous inference.* The issue of multiple comparisons is one of error rates. Each of the individual tests or confidence intervals has a Type I error rate $\mathcal{E}_i$ that can be controlled by the experimenter. If we consider the tests together as a *family,* then we can also compute a combined Type I error rate for the family of tests or intervals. When a family contains more and more true null hypotheses, the probability that one or more of these true null hypotheses is rejected increases, and the probability of any Type I errors in the family can become quite large. Multiple comparisons procedures deal with Type I error rates for families of tests.

Multiple
comparisons,
simultaneous
inference, families
of hypotheses

Example 5.1 Carcinogenic mixtures

We are considering a new cleaning solvent that is a mixture of 100 chemicals. Suppose that regulations state that a mixture is safe if all of its constituents are safe (pretending we can ignore chemical interaction). We test the 100 chemicals for causing cancer, running each test at the 5% level. This is the individual error rate that we can control.

What happens if all 100 chemicals are harmless and safe? Because we are testing at the 5% level, we expect 5% of the nulls to be rejected even when all the nulls are true. Thus, on average, 5 of the 100 chemicals will be declared to be carcinogenic, even when all are safe. Moreover, if the tests are independent, then one or more of the chemicals will be declared unsafe in 99.4% of all sets of experiments we run, even if all the chemicals are safe. This 99.4% is a combined Type I error rate; clearly we have a problem.

A particularly insidious version of multiple testing arises when data arrive sequentially and the same hypothesis is tested multiple times as data accumulate. The problem arises when you stop taking data as soon as a "significant" result is obtained. For an extreme example, suppose that you are accumulating data from a normal with mean μ and variance 1, and you wish to test $H_0 \colon \mu = 0$. You can take up to 100 data points, but you will take 10
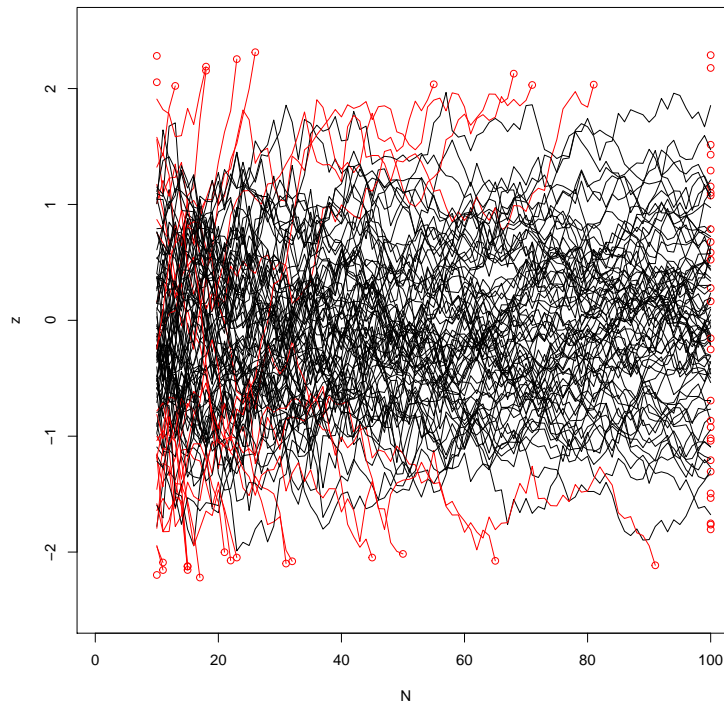
Figure 5.1: Traces of the z statistic from 10 to 100 samples, stopping when $|z| > 2$. Twenty-eight of 100 traces stopped; these are marked in red. Final z values of the stopped traces are shown at $n = 100$.

to begin and keep taking data one point at a time until you get a z statistic ($z = \overline{x}\sqrt{n}$) that is greater than 2 in absolute value or until you reach $n = 100$. Even though we are testing the same null value for the same parameter, this is a family of tests indexed by the sample size.

Figure 5.1 shows the trajectories of z for 100 simulated experiments. Of the 100 simulations, 28 reached $|z| > 2$. The nominal 5% error rate was nearly 30% when following this scheme. If you look just at $n = 100$, only 2 of the 100 simulations are significant (this is within sampling variability of the expected 5 significant). The problem is not the test itself; the problem is repeated peeking over time and the fact that the z test can reject, but it will never stop in favor of the null. There are sequential methods that allow you to test as the sample size increases, but the boundaries for stopping the test are much wider than the boundaries computed for fixed sample sizes.

Testing as data accumulate inflates Type I error

Bayesians have a different point of view on multiple testing

Before moving on to a closer study of the frequentist approach to multi-

ple comparisons, let us note here that the situation is quite different for the Bayesian statistician. Fundamentally, Bayesians are not concerned with Type I error rates, which are just not part of the Bayesian paradigm. Bayesians are concerned with posterior distributions, and these depend on prior distributions and data. The posterior does not depend on a stopping rule for collecting data so long as the stopping rule is *ignorable*, which means that the stopping rule only depends on unknown parameters through the data. This has led to claims that Bayesians never need to worry about stopping rules for data collection. This is, perhaps, an overstatement, as it assumes that our prior and likelihood are exact representations of nature. For example, Rosenbaum and Rubin (1984) show that stopping rules can increase the sensitivity of Bayesian analysis to the specification of the prior.

<div style="text-align: right;">Posteriors unaffected by ignorable stopping rules</div>

Still, there is evidence that Bayesian methods are less affected by stopping rules than frequentist methods. This evidence is often in the context of looking at the frequentist properties of a Bayesian technique, something a fervent Bayesian might not do. For example, consider the sampling situation described above, but instead of stopping when $|z| > 2$, stop when the Bayes factor is greater than 3 (favor alternative, which has a prior for μ that is normal with mean 0 and variance 1) or less than 1/3 (favor null). Recall that 3 is the smallest Bayes factor we are considering positive evidence for a hypothesis. Of 100 simulations, four stopped in favor of the alternative and the other 96 stopped in favor of the null. If we don't stop sampling, of the Bayes factors at $n = 100$, 83 were in favor of the null, and two were in favor of the alternative. That is certainly better than what we saw with the z test.

This chapter will spend a lot of time on pairwise comparisons, where we consider the possibility that $\mu_i = \mu_j$. Consider, however, the prior distributions we have been using for the single mean model and the separate means model. In the former, all treatments have the same mean. In the latter, all treatments have different means with probability 1. That is, under the separate means model with our prior, the posterior will never assign positive probability to a situation with two or more equal means.

One can create Bayesian procedures that behave in an analogous way to certain frequentist multiple comparison procedures by creating loss functions that lead to decisions that include declaring some treatment means to be grouped or clustered, using a different prior for treatment means that *does* allow means to be equal, or augmenting the data description with many potential models that assume various subsets of treatments have the same mean and then doing model selection. However, error rates are still not a part of any of these. Read more about multiple comparisons in a Bayesian context in Section 5.9.

5.1 Error Rates

When we have more than one test or interval to consider, there are several ways to define a combined Type I error rate for the family of tests. This variety of combined Type I error rates is the source of much confusion in the use

<div style="text-align: right;">Determine error rate to control</div>

of multiple comparisons, as different error rates lead to different procedures. People sometimes ask "Which procedure should I use?" when the real question is "Which error rate do I want to control?". As data analyst, you need to decide which error rate is appropriate for your situation and then choose a method of analysis appropriate for that error rate. This choice of error rate is not so much a statistical decision as a scientific decision in the particular area under consideration.

Data snooping is a practice related to having many tests. Data snooping occurs when we first look over the data and then choose the null hypotheses to be tested based on "interesting" features in the data. What we tend to do is consider many potential features of the data and discard those with uninteresting or null behavior. When we data snoop and then perform a test, we tend to see the smallest p-value from the ill-defined family of tests that we considered when we were snooping; we have not really performed just one test. Some multiple comparisons procedures can actually control for data snooping.

Data snooping performs many implicit tests

> *Simultaneous inference* is deciding which error rate we wish to control, and then using a procedure that controls the desired error rate.

Let's set up some notation for our problem. We have a set of K null hypotheses H_{01}, H_{02}, ..., H_{0K}. We also have the "combined," "overall," or "intersection" null hypotheses H_0 which is true if *all* of the H_{0i} are true. In formula,

$$H_0 = H_{01} \cap H_{02} \cap \cdots \cap H_{0K}.$$

Individual and combined null hypotheses

The collection H_{01}, H_{02}, ..., H_{0K} is sometimes called a family of null hypotheses. We reject H_0 if any of null hypotheses H_{0i} is rejected. In Example 5.1, $K = 100$, H_{0i} is the null hypothesis that chemical i is safe, and H_0 is the null hypothesis that all chemicals are safe so that the mixture is safe.

We now define five combined Type I error rates. The definitions of these error rates depend on numbers or fractions of falsely rejected null hypotheses H_{0i}, which will never be known in practice. We set up the error rates here and later give procedures that can be shown mathematically to control the error rates.

The *per comparison error rate* or *comparisonwise error rate* is the probability of rejecting a particular H_{0i} in a single test when that H_{0i} is true. Controlling the per comparison error rate at $\mathcal{E}$ means that the expected fraction of individual tests that reject H_{0i} when H_0 is true is $\mathcal{E}$. This is just the usual error rate for a t-test or F-test; it makes no correction for multiple comparisons. The tests in Example 5.1 controlled the per comparison error rate at 5%.

Comparisonwise error rate

The *per experiment error rate* or *experimentwise error rate* or *familywise error rate* is the probability of rejecting one or more of the H_{0i} (and thus rejecting H_0) in a series of tests when all of the H_{0i} are true. Controlling the experimentwise error rate at $\mathcal{E}$ means that the expected fraction of experiments in which we would reject one or more of the H_{0i} when H_0 is true is $\mathcal{E}$. In Example 5.1, the per experiment error rate is the fraction of times

Experimentwise error rate

we would declare one or more of the chemicals unsafe when in fact all were safe. Controlling the experimentwise error rate at $\mathcal{E}$ necessarily controls the comparisonwise error rate at no more than $\mathcal{E}$. The experimentwise error rate considers all individual null hypotheses that were rejected; if any one of them was correctly rejected, then there is no penalty for any false rejections that may have occurred.

A statistical discovery is the rejection of an H_{0i}. The false discovery fraction is 0 if there are no rejections; otherwise it is the number of false discoveries (Type I errors) divided by the total number of discoveries. The *false discovery rate* (FDR) is the expected value of the false discovery fraction. If H_0 is true, then all discoveries are false and the FDR is just the experimentwise error rate. Thus controlling the FDR at $\mathcal{E}$ also controls the experimentwise error at $\mathcal{E}$. However, the FDR also controls at $\mathcal{E}$ the average fraction of rejections that are Type I errors when some H_{0i} are true and some are false, a control that the experimentwise error rate does not provide. With the FDR, we are allowed more incorrect rejections as the number of true rejections increases, but the ratio is limited. For example, with FDR at .05, we are allowed just one incorrect rejection with 19 correct rejections.

The *strong familywise error rate* is the probability of making any false discoveries, that is, the probability that the false discovery fraction is greater than zero. Controlling the strong familywise error rate at $\mathcal{E}$ means that the probability of making any false rejections is $\mathcal{E}$ or less, regardless of how many correct rejections are made. Thus one true rejection cannot make any false rejections more likely. Controlling the strong familywise error rate at $\mathcal{E}$ controls the FDR at no more than $\mathcal{E}$. In Example 5.1, a strong familywise error rate of $\mathcal{E}$ would imply that in a situation where 2 of the chemicals were carcinogenic, the probability of declaring one of the other 98 to be carcinogenic would be no more than $\mathcal{E}$.

Finally, suppose that each null hypothesis relates to some parameter (for example, a mean), and we put confidence intervals on all these parameters. An error occurs when one of our confidence intervals fails to cover the true parameter value. If this true parameter value is also the null hypothesis value, then an error is a false rejection. The *simultaneous confidence intervals* criterion states that all of our confidence intervals must cover their true parameters simultaneously with confidence $1 - \mathcal{E}$. Simultaneous $1 - \mathcal{E}$ confidence intervals also control the strong familywise error rate at no more than $\mathcal{E}$. (In effect, the strong familywise criterion only requires simultaneous intervals for the null parameters.) In Example 5.1, we could construct simultaneous confidence intervals for the cancer rates of each of the 100 chemicals. Note that a single confidence interval in a collection of intervals with simultaneous coverage $1 - \mathcal{E}$ will have coverage greater than $1 - \mathcal{E}$.

There is a trade-off between Type I error and Type II error (failing to reject a null when it is false). As we go to more and more stringent Type I error rates, we become more confident in the rejections that we do make, but it also becomes more difficult to make rejections. Thus, when using the more stringent Type I error controls, we are more likely to fail to reject some null hypotheses that should be rejected than when using the less stringent rates. In

False discovery rate

Strong familywise error rate

Simultaneous confidence intervals

More stringent procedures are less powerful

simultaneous inference, controlling stronger error rates leads to less powerful tests.

Example 5.2 Functional magnetic resonance imaging

Many functional Magnetic Resonance Imaging (fMRI) studies are interested in determining which areas of the brain are "activated" when a subject is engaged in some task. Any one image slice of the brain may contain 5000 voxels (individual locations to be studied), and one analysis method produces a t-test for each of the 5000 voxels. Null hypothesis H_{0i} is that voxel i is not activated. Which error rate should we use?

If we are studying a small, narrowly defined brain region and are unconcerned with other brain regions, then we would want to test individually the voxels in the brain regions of interest. The fact that there are 4999 other voxels is unimportant, so we would use a per comparison method.

Suppose instead that we are interested in determining if there are any activations in the image. We recognize that by making many tests we are likely to find one that is "significant", even when all nulls are true; we want to protect ourselves against that possibility, but otherwise need no stronger control. Here we would use a per experiment error rate.

Suppose that we believe that there will be many activations, so that H_0 is not true. We don't want some correct discoveries to open the flood gates for many false discoveries, but we are willing to live with some false discoveries as long as they are a controlled fraction of the total made. This is acceptable because we are going to investigate several subjects; the truly activated rejections should be rejections in most subjects, and the false rejections will be scattered. Here we would use the FDR.

Suppose that in addition to expecting true activations, we are also only looking at a single subject, so that we can't use multiple subjects to determine which activations are real. Here we don't want false activations to cloud our picture, so we use the strong familywise error rate.

Finally, we might want to be able to estimate the amount of activation in every voxel, with simultaneous accuracy for all voxels. Here we would use simultaneous confidence intervals.

> A *multiple comparisons procedure* is a method for controlling a Type I error rate other than the per comparison error rate.

The literature on multiple comparisons is vast, and despite the length of this Chapter, we will only touch the highlights. I have seen quite a bit of nonsense regarding these methods, so I will try to set out rather carefully what the methods are doing. We begin with a discussion of Bonferroni-based methods for combining generic tests. Next we consider the Scheffé procedure, which is useful for contrasts suggested by data (data snooping). Then we turn our attention to pairwise comparisons, for which there are dozens of methods. Finally, we consider comparing treatments to a control or to the best response.

5.2 Bonferroni-Style Methods

The Bonferroni technique is the simplest, most widely applicable multiple comparisons procedure. The Bonferroni procedure works for a fixed set of K null hypotheses to test or parameters to estimate. Let p_i be the p-value for testing H_{0i}. The Bonferroni procedure says to obtain simultaneous $1 - \mathcal{E}$ confidence intervals by constructing individual confidence intervals with coverage $1 - \mathcal{E}/K$, or reject H_{0i} (and thus H_0) if

$$p_i < \mathcal{E}/K \ .$$

<div align="right">Ordinary Bonferroni</div>

That is, simply run each test at level $\mathcal{E}/K$. The testing version controls the strong familywise error rate, and the confidence intervals are simultaneous. The tests and/or intervals need not be independent, of the same type, or related in any way.

The *Holm* procedure is a modification of Bonferroni that controls the strong familywise error rate, but does not produce simultaneous confidence intervals (Holm 1979). Let $p_{(1)}, \ldots, p_{(K)}$ be the p-values for the K tests sorted into increasing order, and let $H_{0(i)}$ be the null hypotheses sorted along with the p-values. Then reject $H_{0(i)}$ if

<div align="right">Holm</div>

$$p_{(j)} \leq \mathcal{E}/(K - j + 1) \text{ for all } j = 1, \ldots, i.$$

Thus we start with the smallest p-value; if it is rejected we consider the next smallest, and so on. We stop when we reach the first nonsignificant p-value. This is a little more complicated, but we gain some power since only the smallest p-value is compared to $\mathcal{E}/K$.

The Benjamini and Hochberg method (Benjamini and Hochberg 1995, abbreviated BH) controls the False Discovery Rate. Once again, sort the p-values and the hypotheses. For BH, start with the largest p-value and work down. Reject $H_{0(i)}$ if

<div align="right">Benjamini and Hochberg requires independent tests</div>

$$p_{(j)} \leq \mathcal{E}(j/K) \text{ for some } j \geq i.$$

This procedure is correct when the tests are statistically independent. It controls the FDR, but not the strong familywise error rate.

Benjamini and Yekutieli (Benjamini and Yekutieli 2001) show that the BH approach also controls the FDR for some types of dependent tests. They also provide a modification to the BH approach that controls FDR for tests with any form of dependence. This modification will be more conservative (make fewer rejections) than BH when the tests are independent. For Benjamini and Yekutieli (abbreviated BY), reject $H_{0(i)}$ if

$$p_{(j)} \leq \mathcal{E}j/\sum_{k=1}^{K} K/k \text{ for some } j \geq i.$$

The expression in the denominator can be reasonably approximated via:

$$\sum_{k=1}^{K} K/k \approx K(\ln(K) + .5)$$

Reject $H_{0(i)}$ if	Method	Control
$p_{(i)} < \mathcal{E}/K$	Bonferroni	Simultaneous confidence intervals
$p_{(j)} < \mathcal{E}/(K-j+1)$ for all $j = 1,\ldots,i$	Holm	Strong familywise error rate
$p_{(j)} \leq \mathcal{E}j/K$ for some $j \geq i$	Benjamini & Hochberg	False discovery rate; needs independent tests
$p_{(j)} \leq \mathcal{E}j/\sum_{k=1}^{K} K/k$ for some $j \geq i$	Benjamini & Yekutieli	False discovery rate; general tests

Display 5.1: Summary of Bonferroni-style methods for K comparisons.

The four Bonferroni methods are summarized in Display 5.1. Example 5.3 illustrates their use.

Example 5.3 Sensory characteristics of cottage cheeses

Table 5.1 shows the results of an experiment comparing the sensory characteristics of nonfat, 2% fat, and 4% fat cottage cheese (Michicich 1995). The table shows the characteristics grouped by type and p-values for testing the null hypothesis that there was no difference between the three cheeses in the various sensory characteristics. There are 21 characteristics in three groups of sizes 7, 6, and 8.

How do we do multiple comparisons here? First we need to know:

1. Which error rate is of interest?

2. If we do choose an error rate other than the per comparison error rate, what is the appropriate "family" of tests? Is it all 21 characteristics, or separately within group of characteristic?

There is no automatic answer to either of these questions. The answers depend on the goals of the study, the tolerance of the investigator to Type I error, how the results of the study will be used, whether the investigator views the three groups of characteristics as distinct, and so on.

The last two columns of Table 5.1 give the results of the Bonferroni, Holm, and BH procedures applied at the 5% level to all 21 comparisons and within each group. The p-values are compared to the criteria in Display 5.1 using $K = 21$ for the overall family and K of 7, 6, or 8 for by group comparisons.

Consider the characteristic "cheesy flavor" with a .01 p-value. If we use the overall family, this is the tenth smallest p-value out of 21 p-values. The

Table 5.1: Sensory attributes of three cottage cheeses: p-values and 5% significant results overall and familywise by type of attribute using the Bonferroni ($\bullet$), Holm ($\circ$), and BH methods($\star$).

Appearance

Characteristic	p-value	Overall	By group
White	.004	$\star$	$\bullet\circ\star$
Yellow	.002	$\bullet\circ\star$	$\bullet\circ\star$
Gray	.13		
Curd size	.29		
Size uniformity	.73		
Shape uniformity	.08		
Liquid/solid ratio	.02	$\star$	$\star$

Flavor

Characteristic	p-value	Overall	By group
Sour	.40		
Sweet	.24		
Cheesy	.01	$\star$	$\circ\star$
Rancid	.0001	$\bullet\circ\star$	$\bullet\circ\star$
Cardboard	.0001	$\bullet\circ\star$	$\bullet\circ\star$
Storage	.001	$\bullet\circ\star$	$\bullet\circ\star$

Texture

Characteristic	p-value	Overall	By group
Breakdown rate	.001	$\bullet\circ\star$	$\bullet\circ\star$
Firm	.0001	$\bullet\circ\star$	$\bullet\circ\star$
Sticky	.41		
Slippery	.07		
Heavy	.15		
Particle size	.42		
Runny	.002	$\bullet\circ\star$	$\bullet\circ\star$
Rubbery	.006	$\star$	$\bullet\circ\star$

results are

- *Bonferroni* The critical value is $.05/21 = .0024$—not significant.

- *Holm* The critical value is $.05/(21 - 10 + 1) = .0042$—not significant.

- *BH* The critical value is $10 \times .05/21 = .024$—significant.

If we use the flavor family, this is the fourth smallest p-value out of six p-values. Now the results are

- *Bonferroni* The critical value is $.05/6 = .008$—not significant.

- *Holm* The critical value is $.05/(6 - 4 + 1) = .017$ (and all smaller p-values meet their critical values)—significant.

- *BH* The critical value is $4 \times .05/6 = .033$—significant.

R has a function that will adjust p-values for Bonferroni, Holm, BH, and BYF, as well as some additional methods.

```
 1 > p.appearance <- c(.004,.002,.13,.29,.73,.02,.02)
 2 > p.flavor <- c(.4,.24,.01,.0001,.0001,.001)
 3 > p.texture <- c(.001,.0001,.41,.07,.15,.42,.002,.006)
 4 > p.adjust(p.appearance,method='holm')
   [1] 0.024 0.014 0.390 0.580 0.730 0.100 0.100
 5 > p.adjust(p.flavor,method='holm')
   [1] 0.4800 0.4800 0.0300 0.0006 0.0006 0.0040
 6 > p.adjust(p.texture,method='holm')
   [1] 0.0070 0.0008 0.8200 0.2800 0.4500 0.8200 0.0120 0.0300
 7 > p.adjust(c(p.appearance,p.flavor,p.texture),method='holm')
    [1] 0.056 0.032 1.000 1.000 1.000 0.220 0.220 1.000 1.000 0.120 0.002 0.002
   [13] 0.018 0.018 0.002 1.000 0.630 1.000 1.000 0.032 0.078
 8 > p.adjust(p.appearance,method='fdr')
   [1] 0.01400 0.01400 0.18200 0.33833 0.73000 0.03500 0.03500
 9 > p.adjust(p.flavor,method='fdr')
   [1] 0.4000 0.2880 0.0150 0.0003 0.0003 0.0020
10 > p.adjust(p.texture,method='fdr')
   [1] 0.00400 0.00080 0.42000 0.11200 0.20000 0.42000 0.00533 0.01200
11 > p.adjust(c(p.appearance,p.flavor,p.texture),method='fdr')
    [1] 0.0105 0.0060 0.1950 0.3582 0.7300 0.0350 0.0350 0.4410 0.3150 0.0210
   [11] 0.0007 0.0007 0.0042 0.0042 0.0007 0.4410 0.1130 0.2100 0.4410 0.0060
   [21] 0.0140
```

Lines 1–3 simply input the p-values for the three groups of characteristics. Lines 4–6 apply the Holm adjustment to the p-values in each group separately, and line 7 does the same for all p-values considered as a single family. The remaining four lines repeat that process using BH.

We have presented Holm and BH as comparing p-values to the nominal rate $\mathcal{E}$ divided by an adjustment factor. In **R**, `p.adjust` instead multiplies each p-value by the adjustment factor, and you then compare that to the nominal rate $\mathcal{E}$.

"Cheesy" is the third characteristic in the flavor group; the Holm and FDR adjusted p-values are .03 and .015. Both of these are significant at the .05 level. When we combine all of the p-values, cheesy is the tenth characteristic in the list; the Holm and FDR adjusted p-values are .12 and .021. Only the FDR is significant for cheesy in the combined set of characteristics.

These results illustrate that more null hypotheses are rejected considering each group of characteristics to be a family of tests rather than overall (the K is smaller for the individual groups), and fewer rejections are made using the more stringent error rates. Again, the choices of error rate and family of tests are not purely statistical, and controlling an error rate within a group of tests does not control that error rate for all tests.

5.3   The Scheffé Method for *All* Contrasts

The Scheffé method is a multiple comparisons technique for contrasts that produces simultaneous confidence intervals for *any* and *all* contrasts, *including contrasts suggested by the data.* Thus Scheffé is the appropriate technique for assessing contrasts that result from data snooping. This sounds like the ultimate in error rate control—arbitrarily many comparisons, even ones suggested from the data! The downside of this amazing protection is low power, that is, a reduced ability to detect differences that are there. Thus we only use the Scheffé method in those situations where we have a contrast suggested by the data, or many, many contrasts that cannot be handled by other techniques. In addition, pairwise comparison contrasts $\overline{y}_{i\bullet} - \overline{y}_{j\bullet}$, even pairwise comparisons suggested by the data, are better handled by methods specifically designed for pairwise comparisons.

<div align="right">Scheffé protects against data snooping, but has low power</div>

We begin with the Scheffé test of the null hypothesis $H_0\colon w(\{\alpha_i\}) = 0$ against a two-sided alternative. The Scheffé test statistic is the ratio

$$\frac{\mathrm{SS_w}/(g-1)}{\mathrm{MS_E}} \ ;$$

we get a p-value as the area under an F-distribution with $g-1$ and ν degrees of freedom to the right of the test statistic. The degrees of freedom ν are from our denominator $\mathrm{MS_E}$; $\nu = N - g$ for the completely randomized designs we have been considering so far. Reject the null hypothesis if this p-value is less than our Type I error rate $\mathcal{E}$. In effect, the Scheffé procedure treats the mean square for any single contrast as if it were the full $g-1$ degrees of freedom between groups mean square.

<div align="right">Scheffé F-test</div>

There is also a Scheffé t-test for contrasts. Suppose that we are testing the null hypothesis $H_0\colon w(\{\alpha_i\}) = \delta$ against a two-sided alternative. The Scheffé t-test controls the Type I error rate at $\mathcal{E}$ by rejecting the null hypothesis when

<div align="right">Scheffé t-test</div>

$$\frac{|w(\{\overline{y}_{i\bullet}\}) - \delta|}{\sqrt{\mathrm{MS_E} \ \sum_{i=1}^{g} \frac{w_i^2}{n_i}}} > \sqrt{(g-1)F_{\mathcal{E},g-1,\nu}} \ ,$$

where $F_{\mathcal{E},g-1,\nu}$ is the upper $\mathcal{E}$ percent point of an F-distribution with $g-1$ and ν degrees of freedom. Again, ν is the degrees of freedom for $\mathrm{MS_E}$. For the usual null hypothesis value $\delta = 0$, this is equivalent to the ratio-of-mean-squares version given above.

We may also use the Scheffé approach to form simultaneous confidence intervals for any $w(\{\alpha_i\})$:

<div align="right">Scheffé confidence interval</div>

$$w(\{\overline{y}_{i\bullet}\}) \pm \sqrt{(g-1)F_{\mathcal{E},g-1,\nu}} \times \sqrt{\mathrm{MS_E} \ \sum_{i=1}^{g} \frac{w_i^2}{n_i}} \ .$$

These Scheffé intervals have simultaneous $1 - \mathcal{E}$ coverage over any set of contrasts, including contrasts suggested by the data.

Example 5.4 Acid rain and birch seedlings, continued

Example 3.1 introduced an experiment in which birch seedlings were exposed to various levels of artificial acid rain. The following table gives some summaries for the data:

			pH		
	4.7	4.0	3.3	3.0	2.3
avg. weight	.337	.296	.320	.298	.177
n	48	48	48	48	48

The MS_E was .0119 with 235 degrees of freedom.

Inspection of the means shows that most of the response means are about .3, but the response for the pH 2.3 treatment is much lower. This suggests that a contrast comparing the pH 2.3 treatment with the mean of the other treatments would have a large value. The coefficients for this contrast are $(.25, .25, .25, .25, -1)$. This contrast has value

$$\frac{.337 + .296 + .320 + .298}{4} - .177 = .1357$$

and standard error

$$\sqrt{.0119 \left(\frac{.0625}{48} + \frac{.0625}{48} + \frac{.0625}{48} + \frac{.0625}{48} + \frac{1}{48} \right)} = .0176 .$$

We must use the Scheffé procedure to construct a confidence interval or assess the significance of this contrast, because the contrast was suggested by the data. For a 99% confidence interval, the Scheffé multiplier is

$$\sqrt{4 \, F_{.01,4,235}} = 3.688 .$$

Thus the 99% confidence interval for this contrast is $.1357 - 3.688 \times .0176$ up to $.1357 + 3.688 \times .0176$, or $(.0708, .2006)$. Alternatively, the t-statistic for testing the null hypothesis that the mean response in the last group is equal to the average of the mean responses in the other four groups is $.1357/.0176 = 7.71$. The Scheffé critical value for testing the null hypothesis at the $\mathcal{E} = .001$ level is

$$\sqrt{(g-1)F_{\mathcal{E},g-1,N-g}} = \sqrt{4 \, F_{.001,4,235}} = \sqrt{4 \times 4.782} = 4.37 ,$$

so we can reject the null at the .001 level.

> Remember, it is not fair to hunt around through the data for a big contrast, test it, and think that you've only done one comparison.

5.4   Pairwise Comparisons

A *pairwise comparison* is a contrast that examines the difference between two treatment means $\overline{y}_{i\bullet} - \overline{y}_{j\bullet}$. For g treatment groups, there are

$$\binom{g}{2} = \frac{g(g-1)}{2}$$

different pairwise comparisons. Pairwise comparisons procedures control a Type I error rate at $\mathcal{E}$ for all pairwise comparisons. If we data snoop, choose the biggest and smallest $\overline{y}_{i\bullet}$'s and take the difference, we have not made just one comparison; rather we have made all $g(g-1)/2$ pairwise comparisons, and selected the largest. Controlling a Type I error rate for this greatest difference is one way to control the error rate for all differences.

As with many other inference problems, pairwise comparisons can be approached using confidence intervals or tests. That is, we may compute confidence intervals for the differences $\mu_i - \mu_j$ or $\alpha_i - \alpha_j$ or test the null hypotheses $H_{0ij} : \mu_i = \mu_j$ or $H_{0ij} : \alpha_i = \alpha_j$. Confidence regions for the differences of means are generally more informative than tests.

Tests or confidence intervals

A pairwise comparisons procedure can generally be viewed as a critical value (or set of values) for the t-tests of the pairwise comparison contrasts. Thus we would reject the null hypothesis that $\alpha_i - \alpha_j = 0$ if

$$\frac{|\overline{y}_{i\bullet} - \overline{y}_{j\bullet}|}{\sqrt{\mathrm{MS_E}}\sqrt{1/n_i + 1/n_j}} > u \ ,$$

where u is a critical value. Various pairwise comparisons procedures differ in how they define the critical value u, and u may depend on several things, including $\mathcal{E}$, the degrees of freedom for $\mathrm{MS_E}$, the number of treatments, the number of treatments with means between $\overline{y}_{i\bullet}$ and $\overline{y}_{j\bullet}$, and the number of treatment comparisons with larger t-statistics.

Critical values u for t-tests

An equivalent form of the test will reject if

$$|\overline{y}_{i\bullet} - \overline{y}_{j\bullet}| > u \ \sqrt{\mathrm{MS_E}}\sqrt{1/n_i + 1/n_j} = D_{ij} \ .$$

If all sample sizes are equal and the critical value u is constant, then D_{ij} will be the same for all i, j pairs and we would reject the null if any pair of treatments had mean responses that differed by D or more. This quantity D is called a *significant difference*; for example, using a Bonferroni adjustment to the $g(g-1)/2$ pairwise comparisons tests leads to a Bonferroni significant difference (BSD).

Significant differences D_{ij}

Confidence intervals for pairwise differences $\mu_i - \mu_j$ can be formed from the pairwise tests via

$$(\overline{y}_{i\bullet} - \overline{y}_{j\bullet}) \pm u\sqrt{\mathrm{MS_E}}\sqrt{1/n_i + 1/n_j} \ .$$

The remainder of this section presents methods for displaying the results of pairwise comparisons, introduces the Studentized range, discusses several pairwise comparisons methods, and then illustrates the methods with an example.

5.4.1   Displaying the results

Pairwise comparisons generate a lot of tests, so we need convenient and compact ways to present the results. An *underline diagram* is a graphical presentation of pairwise comparison results; construct the underline diagram in the following steps.

<div style="text-align: right">Underline
diagram
summarizes
pairwise
comparisons</div>

1. Sort the treatment means into increasing order and write out treatment labels (numbers or names) along a horizontal axis. The $\overline{y}_{i\bullet}$ values may be added if desired.

2. Draw a line segment under a group of treatments if no pair of treatments in that group is significantly different. Do not include short lines that are implied by long lines. That is, if treatments 4, 5, and 6 are not significantly different, only use one line under all of them—not a line under 4 and 5, and a line under 5 and 6, and a line under 4, 5, and 6.

Here is a sample diagram for three treatments that we label A, B, and C:

$$\underline{\text{C}\quad \underline{\text{A}\quad \text{B}}}$$
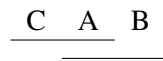
This diagram includes treatment labels, but not treatment means. From this summary we can see that C can be distinguished from B (there is no underline that covers both B and C), but A cannot be distinguished from either B or C (there are underlines under A and C, and under A and B).

Note that there can be some confusion after pairwise comparisons. You must not confuse "is not significantly different from" or "cannot be distinguished from" with "is equal to." Treatment mean A cannot be equal to treatment means B and C and still have treatment means B and C not equal each other. Such a pattern can hold for results of significance tests.

<div style="text-align: right">Insignificant
difference does
not imply equality</div>

There are also several nongraphical methods for displaying pairwise comparisons results. In one method, we sort the treatments into order of increasing means and print the treatment labels. Each treatment label is followed by one or more numbers (letters are sometimes used instead). Any treatments sharing a number (or letter) are not significantly different. Thus treatments sharing common numbers or letters are analogous to treatments being connected by an underline. The grouping letters are often put in parentheses or set as sub- or superscripts. The results in our sample underline diagram might thus be presented as one of the following:

<div style="text-align: right">Letter or number
tags</div>

C (1)	A (12)	B (2)		C (a)	A (ab)	B (b)

$$C^1 \qquad A^{12} \qquad B^2 \qquad\qquad C^a \qquad A^{ab} \qquad B^b$$

There are several other variations on this theme.

A third way to present pairwise comparisons is as a table, with treatments labeling both rows and columns. Table elements can flag significant differences or contain confidence intervals for the differences. Only entries above or below the diagonal of the table are needed.

5.4.2 The Studentized range

The range of a set is the maximum value minus the minimum value, and *Studentization* means dividing a statistic by an estimate of its standard error. Thus the *Studentized range* for a set of treatment means is

$$
\max_i \frac{\overline{y}_{i\bullet}}{\sqrt{\mathrm{MS_E}/n}} - \min_j \frac{\overline{y}_{j\bullet}}{\sqrt{\mathrm{MS_E}/n}} \ .
$$

Note that we have implicitly assumed that all the sample sizes n_i are the same.

If all the treatments have the same mean, that is, if H_0 is true, and all of our distributional assumptions are correct, then the Studentized range statistic follows the Studentized range distribution. Large values of the Studentized range are less likely under H_0 and more likely under the alternative when the means are not all equal, so we may use the Studentized range as a test statistic for H_0, rejecting H_0 when the Studentized range statistic is sufficiently large. This Studentized range test is a legitimate alternative to the ANOVA F-test.

The Studentized range distribution is important for pairwise comparisons because it is the distribution of the biggest (scaled) difference between treatment means when the null hypothesis is true. We will use it as a building block in several pairwise comparisons methods.

The Studentized range distribution depends only on g and ν, the number of groups and the degrees of freedom for the error estimate $\mathrm{MS_E}$. The quantity $q_{\mathcal{E}}(g, \nu)$ is the upper $\mathcal{E}$ percent point of the Studentized range distribution for g groups and ν error degrees of freedom; it is tabulated in Appendix Table C.8.

5.4.3 Simultaneous confidence intervals

The Tukey honest significant difference (HSD) is a pairwise comparisons technique that uses the Studentized range distribution to construct simultaneous confidence intervals for differences of all pairs of means. If we reject the null hypothesis H_{0ij} when the (simultaneous) confidence interval for $\mu_i - \mu_j$ does not include 0, then the HSD also controls the strong familywise error rate.

The HSD uses the critical value

$$
u(\mathcal{E}, \nu, g) = \frac{q_{\mathcal{E}}(g, \nu)}{\sqrt{2}} \ ,
$$

leading to

$$HSD = \frac{q_{\mathcal{E}}(g,\nu)}{\sqrt{2}}\sqrt{MS_E}\sqrt{\frac{1}{n}+\frac{1}{n}} = \frac{q_{\mathcal{E}}(g,\nu)\sqrt{MS_E}}{\sqrt{n}} \; .$$

Form simultaneous $1 - \mathcal{E}$ confidence intervals via

$$\overline{y}_{i\bullet} - \overline{y}_{j\bullet} \pm \frac{q_{\mathcal{E}}(g,\nu)}{\sqrt{2}}\sqrt{MS_E}\sqrt{\frac{1}{n}+\frac{1}{n}} \; .$$

The degrees of freedom ν are the degrees of freedom for the error estimate MS_E.

Strictly speaking, the HSD is only applicable to the equal sample size situation. For the unequal sample size case, the approximate HSD is

$$HSD_{ij} = q_{\mathcal{E}}(g,\nu)\sqrt{MS_E}\sqrt{\frac{1}{2n_in_j/(n_i+n_j)}}$$

or, equivalently,

$$HSD_{ij} = \frac{q_{\mathcal{E}}(g,\nu)}{\sqrt{2}}\sqrt{MS_E}\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \; .$$

This approximate HSD, often called the Tukey-Kramer form, tends to be slightly conservative (that is, the true error rate is slightly less than $\mathcal{E}$).

The Bonferroni significant difference (BSD) is simply the application of the Bonferroni technique to the pairwise comparisons problem to obtain

$$\begin{aligned} u = u(\mathcal{E},\nu,K) &= t_{\mathcal{E}/(2K),\nu} \; , \\ BSD_{ij} &= t_{\mathcal{E}/(2K),\nu}\sqrt{MS_E}\sqrt{1/n_i + 1/n_j} \; , \end{aligned}$$

where K is the number of pairwise comparisons. We have $K = g(g-1)/2$ for all pairwise comparisons between g groups. BSD produces simultaneous confidence intervals and controls the strong familywise error rate.

When making all pairwise comparisons, the HSD is less than the BSD. Thus we prefer the HSD to the BSD for all pairwise comparisons, because the HSD will produce shorter confidence intervals that are still simultaneous. When only a preplanned subset of all the pairs is being considered, the BSD may be less than, and thus preferable, to the HSD.

Example 5.5 Free amino acids in cheese

Cheese is produced by bacterial fermentation of milk. Some bacteria in cheese are added by the cheese producer. Other bacteria are present but were not added deliberately; these are called nonstarter bacteria. Nonstarter

Table 5.2: Total free amino acids in cheeses after 168 days of ripening, data set `CheeseAminoAcid`.

	Strain added		
None	A	B	A&B
4.195	4.125	4.865	6.155
4.175	4.735	5.745	6.488

bacteria vary from facility to facility and are believed to influence the quality of cheese.

Two strains (A and B) of nonstarter bacteria were isolated at a premium cheese facility. These strains will be added experimentally to cheese to determine their effects. Eight cheeses are made. These cheeses all get a standard starter bacteria. In addition, two cheeses will be randomly selected for each of the following four treatments: control, add strain A, add strain B, or add both strains A and B. Table 5.2 gives the total free amino acids in the cheeses after 168 days of ripening. (Free amino acids are thought to contribute to flavor.)

In this example we will make HSD comparisons (tests and confidence intervals) using $\mathcal{E} = .1$. No one would use such a high $\mathcal{E}$, but it permits a nice comparison of the techniques in this example. HSD is appropriate if we want simultaneous confidence intervals on the pairwise differences. The HSD is

$$\frac{q_{\mathcal{E}}(g,\nu)}{\sqrt{2}}\ \sqrt{\mathrm{MS_E}}\ \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\ =\ \frac{q_{.1}(4,4)}{\sqrt{2}}\ \sqrt{.1572}\ \sqrt{\frac{1}{2} + \frac{1}{2}}$$
$$=\ 4.586 \times .3965/1.414 = 1.286\ \ .$$

We form confidence intervals as the observed difference in treatment means, plus or minus 1.286; so for A&B minus control, we have

$$6.322 - 4.185 \pm 1.286 \text{ or } (.851, 3.423)\ \ .$$

In fact, only two confidence intervals for pairwise differences do not include zero. The underline diagram is:

$$
\begin{array}{cccc}
\text{C} & \text{A} & \text{B} & \text{A\&B} \\
4.19 & 4.43 & 5.31 & 6.32 \\
\end{array}
$$

We can use the `pairwise` function from `cfcdae` to automate this process.

```
1 > amino.acid.fit <- lm(freeAminoAcid~strain,data=CheeseAminoAcid)
2 > pairs.out <- pairwise(amino.acid.fit,strain,confidence=.9,type="hsd")
3 > pairs.out
  Pairwise comparisons ( hsd ) of strain
              estimate signif diff      lower      upper
  * A - AB     -1.8915    1.285889 -3.1773889 -0.6056111
    A - B      -0.8750    1.285889 -2.1608889  0.4108889
    A - none    0.2450    1.285889 -1.0408889  1.5308889
    AB - B      1.0165    1.285889 -0.2693889  2.3023889
  * AB - none   2.1365    1.285889  0.8506111  3.4223889
    B - none    1.1200    1.285889 -0.1658889  2.4058889
4 > sidelines(pairs.out)
  none -0.875 |
  A    -0.630 |
  B     0.245 | |
  AB    1.261   |
```

`pairwise` works on a model as fit in line 1. In line 2 we do all pairwise comparisons, selecting the HSD method and confidence .9. Line 3 prints the results, showing a line for every comparison with a point estimate and an interval estimate. Comparisons that are significant are tagged with a $*$. Line 4 shows that we can visualize the results with an "underline" diagram, although it is shown on its side.

5.4.4 Strong familywise error rate

A *step-down method* is a procedure for organizing pairwise comparisons starting with the most extreme pair and then working in. Relabel the groups so that the sample means are in increasing order with $\overline{y}_{(1)\bullet}$ smallest and $\overline{y}_{(g)\bullet}$ largest. (The relabeled estimated effects $\widehat{\alpha}_{(i)}$ will also be in increasing order, but the relabeled true effects $\alpha_{[i]}$ may or may not be in increasing order.) With this ordering, $\overline{y}_{(1)\bullet}$ to $\overline{y}_{(g)\bullet}$ is a stretch of g means, $\overline{y}_{(1)\bullet}$ to $\overline{y}_{(g-1)\bullet}$ is a stretch of $g-1$ means, and $\overline{y}_{(i)\bullet}$ to $\overline{y}_{(j)\bullet}$ is a stretch of $j-i+1$ means. In a step-down procedure, all comparisons for stretches of k means use the same critical value, but we may use different critical values for different k. This has the advantage that we can use larger critical values for long stretches and smaller critical values for short stretches.

Step-down methods work inward from the outside comparisons

Begin with the most extreme pair (1) and (g). Test the null hypothesis that all the means for (1) up through (g) are equal. If you fail to reject, declare all means equal and stop. If you reject, declare (1) different from (g) and go on to the next step. At the next step, we consider the stretches (1) through $(g-1)$ and (2) through (g). If one of these rejects, we declare its ends to be different and then look at shorter stretches within it. If we fail to reject for a stretch, we do not consider any substretches within the stretch. We repeat this subdivision till there are no more rejections. In other words, we declare that means (i) and (j) are different if the stretch from (i) to (j) rejects its null hypothesis and all stretches containing (i) to (j) also reject their null hypotheses.

(i) and (j) are different if their stretch and all containing stretches reject

The REGWR procedure is a step-down range method that controls the

strong familywise error rate without producing simultaneous confidence intervals. The awkward name REGWR abbreviates the Ryan-Einot-Gabriel-Welsch range test, named for the authors who worked on it. The REGWR critical value for testing a stretch of length k depends on $\mathcal{E}$, ν, k, and g. Specifically, we use

$$u = u(\mathcal{E}, \nu, k, g) = q_{\mathcal{E}}(k, \nu)/\sqrt{2} \quad k = g, g-1,$$

and

$$u = u(\mathcal{E}, \nu, k, g) = q_{k\mathcal{E}/g}(k, \nu)/\sqrt{2} \quad k = g-2, g-3, \ldots, 2.$$

This critical value derives from a Studentized range with k groups, and we use percent points with smaller tail areas as we move in to smaller stretches.

As with the HSD, REGWR error rate control is approximate when the sample sizes are not equal. Unlike the situation where we want simultaneous confidence intervals, we could also use the Holm method to control the strong familywise error rate.

Example 5.6 Free amino acids in cheese, continued

Suppose that we only wished to control the strong familywise error rate instead of producing simultaneous confidence intervals. Then we could use REGWR or Holm instead of HSD and could potentially see additional significant differences.

REGWR is a step-down method that begins like the HSD. Comparing C and A&B, we conclude as in the HSD that they are different. We may now compare C with B and A with A&B. These are comparisons that involve stretches of $k = 3$ means; since $k = g - 1$, we still use $\mathcal{E}$ as the error rate. The significant difference for these comparisons is

$$\frac{q_{\mathcal{E}}(k, \nu)}{\sqrt{2}} \sqrt{\text{MS}_{\text{E}}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = \frac{q_{.1}(3, 4)}{\sqrt{2}} \sqrt{.1572} \sqrt{\frac{1}{2} + \frac{1}{2}} = 1.115 \ .$$

Both the B-C and A&B-A differences (1.12 and 1.89) exceed this cutoff, so REGWR concludes that B differs from C, and A differs from A&B. Recall that the HSD did not distinguish C from B.

Having concluded that there are B-C and A&B-A differences, we can now compare stretches of means within them, namely C to A, A to B, and B to A&B. These are stretches of $k = 2$ means, so for REGWR we use the error rate $k\mathcal{E}/g = .05$. The significant difference for these comparisons is

$$\frac{q_{\mathcal{E}/2}(k, \nu)}{\sqrt{2}} \sqrt{\text{MS}_{\text{E}}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = \frac{q_{.05}(2, 4)}{\sqrt{2}} \sqrt{.1572} \sqrt{\frac{1}{2} + \frac{1}{2}} = 1.101 \ .$$

None of the three differences exceeds this cutoff, so we fail to conclude that those treatments differ and finish. The underline diagram is:

$$\begin{array}{cccc} \text{C} & \text{A} & \text{B} & \text{A\&B} \\ 4.19 & 4.43 & 5.31 & 6.32 \end{array}$$

We can automate all this in **R**.

```
1 > pairs.out.regwr <- pairwise(amino.acid.fit,strain,confidence=.9,type="regwr")
2 > pairs.out.regwr
  Pairwise comparisons ( regwr ) of strain
            estimate signif diff        lower       upper
  * A - AB    -1.8915    1.114670 -3.006169508 -0.7768305
    A - B     -0.8750    1.100905 -1.975905084  0.2259051
    A - none   0.2450    1.100905 -0.855905084  1.3459051
    AB - B     1.0165    1.100905 -0.084405084  2.1174051
  * AB - none  2.1365    1.285889  0.850611111  3.4223889
  * B - none   1.1200    1.114670  0.005330492  2.2346695
3 > sidelines(pairs.out.regwr)
  none -0.875 |
  A    -0.630 | |
  B     0.245   | |
  AB    1.261     |
4 > with(CheeseAminoAcid,pairwise.t.test(freeAminoAcid,strain,
        p.adjust.method="holm"))
   Pairwise comparisons using t tests with pooled SD
  data:  freeAminoAcid and strain

       A     AB    B
  AB   0.044 -     -
  B    0.190 0.190 -
  none 0.570 0.034 0.190

  P value adjustment method: holm
```

In `pairwise`, all we need to do is change the type of comparison we wish to make and proceed as before. Note that the computed significant differences are not all the same, unlike the case with HSD.

R also has a builtin function `pairwise.t.test` that works on a response vector and a grouping vector rather than a fitted model. It then does p-value adjustments, in this case using the Holm method. Note that the Holm approach does *not* find the significant difference between the control treatment and treatment B (it is just barely significant using REGWR). Holm is a general purpose method to control the strong familywise error rate, and it is not quite as sensitive as the REGWR method, which was created expressly for pairwise comparisons.

5.4.5   False discovery rate

We can apply the BY method to pairwise comparisons and control the FDR. In addition to BY, there are two methods for which we have evidence of FDR control in pairwise comparisons, but for which we have no definitive proof or counterexample at this time. Benjamini and Yekutieli (2001) propose, and provides some evidence, that BH controls FDR in pairwise comparisons. Shaffer 2007 proposes, and provides some evidence, that Student-Newman-

Keuls also controls the FDR. SNK and BH are each somewhat more sensitive to some differences and less sensitive to others. Neither BY nor BH translates easily to a "significant difference."

The Student-Newman-Keuls (SNK) procedure is a step-down method that uses the Studentized range test with critical value

<div align="right">SNK</div>

$$u = u(\mathcal{E}, \nu, k, g) = q_\mathcal{E}(k, \nu)/\sqrt{2}$$

for a stretch of k means. This is similar to REGWR, except that we keep the percent point of the Studentized range constant as we go to shorter stretches. SNK does not control the strong familywise error rate.

Example 5.7 Free amino acids in cheese, continued

Suppose that we only wished to control the false discovery rate; now we would use SNK instead of the more stringent HSD or REGWR.

SNK is identical to REGWR in the first two stages, so SNK will also get to the point of making the comparisons of the three pairs C to A, A to B, and B to A&B. However, the SNK significant difference for these pairs is less than that used in REGWR:

$$\frac{q_\mathcal{E}(k, \nu)}{\sqrt{2}} \sqrt{\mathrm{MS_E}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = \frac{q_{.1}(2, 4)}{\sqrt{2}} \sqrt{.1572} \sqrt{\frac{1}{2} + \frac{1}{2}} = .845 \ \ .$$

Both the B-A and A&B-B differences (1.02 and .98) exceed the cutoff, but the A-C difference (.14) does not. The underline diagram for SNK is:

$$\begin{array}{cccc} \mathrm{C} & \mathrm{A} & \mathrm{B} & \mathrm{A\&B} \\ \underline{4.19 \quad 4.43} & 5.31 & 6.32 \end{array}$$

In **R**, we have

```
1 > pairs.out.snk <- pairwise(amino.acid.fit,strain,confidence=.9,type="snk")
2 > pairs.out.snk
  Pairwise comparisons ( snk ) of treatment
             estimate signif diff        lower        upper
* A - AB     -1.8915   1.1146695 -3.006169508 -0.77683049
* A - B      -0.8750   0.8453095 -1.720309491 -0.02969051
  A - none    0.2450   0.8453095 -0.600309491  1.09030949
* AB - B      1.0165   0.8453095  0.171190509  1.86180949
* AB - none   2.1365   1.2858889  0.850611111  3.42238889
* B - none    1.1200   1.1146695  0.005330492  2.23466951
3 > sidelines(pairs.out.snk)
none -0.875 |
A    -0.630 |
B     0.245
AB    1.261
```

We can also look at the BH and BY adjusted p-values.

```
4 > with(CheeseAminoAcid,pairwise.t.test(freeAminoAcid,strain,p.adjust.method="BH"))

  Pairwise comparisons using t tests with pooled SD

  data:  freeAminoAcid and strain

       A     AB    B
  AB   0.027 -     -
  B    0.110 0.094 -
  none 0.570 0.027 0.094

  P value adjustment method: BH
```

The output from line 2 shows that every comparison except A–none is significant (at the .1 level) according to SNK. In contrast, we see in the output from line 4 that BH also declares the A–B difference to be non-significant.

```
5 > with(CheeseAminoAcid,pairwise.t.test(freeAminoAcid,strain,p.adjust.method="BY"))

  Pairwise comparisons using t tests with pooled SD

  data:  freeAminoAcid and strain

       A     AB    B
  AB   0.065 -     -
  B    0.270 0.229 -
  none 1.000 0.065 0.229

  P value adjustment method: BY
```

BY is much more conservative, but it guaranteed to control FDR. In this small example, the BY p-values are actually larger than the Holm p-values. That will not always be true when there are more comparisons.

5.4.6   Experimentwise error rate

The Analysis of Variance F-test for equality of means controls the experimentwise error rate. Thus investigating pairwise differences only when the F-test has a p-value less than $\mathcal{E}$ will control the experimentwise error rate. This is the basis for the Protected least significant difference, or Protected LSD. If the F-test rejects at level $\mathcal{E}$, then do simple t-tests at level $\mathcal{E}$ among the different treatments.

Protected LSD uses F-test to control experimentwise error rate

The critical values are from a t-distribution:

$$u(\mathcal{E},\nu) = t_{\mathcal{E}/2,\nu} \ ,$$

leading to the significant difference

$$LSD = t_{\mathcal{E}/2,\nu}\sqrt{\mathrm{MS_E}}\sqrt{1/n_i + 1/n_j} \ .$$

As usual, ν is the degrees of freedom for $\mathrm{MS_E}$, and $t_{\mathcal{E}/2,\nu}$ is the upper $\mathcal{E}/2$ percent point of a t-curve with ν degrees of freedom.

Confidence intervals produced from the protected LSD do not have the anticipated $1 - \mathcal{E}$ coverage rate, either individually or simultaneously. See Section 5.7.

Example 5.8 Free amino acids in cheese, continued

Finally, suppose that we only wish to control the experimentwise error rate. Protected LSD will work here. LSD uses the same significant difference for all pairs:

$$ t_{\mathcal{E}/2,\nu} \sqrt{\mathrm{MS_E}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = t_{.05,4} \sqrt{.1572} \sqrt{\frac{1}{2} + \frac{1}{2}} = .845 \ . $$

This is the same as the SNK comparison for a stretch of length 2. All differences except A-C exceed the cutoff, so the underline diagram for LSD is:

$$
\begin{array}{cccc}
\mathrm{C} & \mathrm{A} & \mathrm{B} & \mathrm{A\&B} \\
\underline{4.19} & \underline{4.43} & 5.31 & 6.32
\end{array}
$$

In **R**, we have

```
> anova(amino.acid.fit)
          Df Sum Sq Mean Sq F value Pr(>F)
strain     3 5.6279 1.87595  11.932 0.0183 *
Residuals  4 0.6289 0.15722
> pairs.out.lsd <- pairwise(amino.acid.fit,strain,confidence=.9,type="lsd")
> pairs.out.lsd

Pairwise comparisons ( lsd ) of treatment
             estimate signif diff       lower       upper
* A - AB      -1.8915   0.8453078 -2.7368078 -1.04619217
* A - B       -0.8750   0.8453078 -1.7203078 -0.02969217
  A - none     0.2450   0.8453078 -0.6003078  1.09030783
* AB - B       1.0165   0.8453078  0.1711922  1.86180783
* AB - none    2.1365   0.8453078  1.2911922  2.98180783
* B - none     1.1200   0.8453078  0.2746922  1.96530783
> sidelines(pairs.out.lsd)
none -0.875 |
A    -0.630 |
B     0.245
AB    1.261
> with(CheeseAminoAcid,pairwise.t.test(freeAminoAcid,treatment,
      p.adjust.method = "none"))
Pairwise comparisons using t tests with pooled SD
data:  freeAminoAcid and strain

      A      AB     B
AB    0.0088 -      -
B     0.0920 0.0624 -
none  0.5701 0.0057 0.0476

P value adjustment method: none
```

First we check to make sure that the F-test rejects at the selected error rate,

Error rate	Method
Simultaneous confidence intervals	HSD
Strong familywise	REGWR
False discovery rate	SNK or BH
Experimentwise	Protected LSD
Comparisonwise	LSD

Display 5.2: Pairwise comparison methods.

then we use `pairwise`. Alternatively, we could use `pairwise.t.test` with adjustment method `none`.

5.4.7 Comparisonwise error rate

Ordinary t-tests and confidence intervals without any adjustment control the comparisonwise error rate. In the context of pairwise comparisons, this is called the least significant difference (LSD) method.

LSD

The critical values are the same as for the protected LSD:

$$u(\mathcal{E}, \nu) = t_{\mathcal{E}/2,\nu},$$

and

$$LSD = t_{\mathcal{E}/2,\nu} \sqrt{\mathrm{MS_E}} \sqrt{1/n_i + 1/n_j} \ .$$

5.4.8 Pairwise testing reprise

It is easy to get overwhelmed by the abundance of methods, and there are still many more that we haven't discussed. Your anchor in all this is your error rate. Once you have determined your error rate, the choice of method is reasonably automatic, as summarized in Display 5.2. Your choice of error rate is determined by the needs of your study, bearing in mind that the more stringent error rates have fewer false rejections, but also fewer correct rejections.

Choose your error rate, not your method

5.4.9 Pairwise comparisons methods that do *not* control combined Type I error rates

There are many other pairwise comparisons methods beyond those already mentioned. In this Section we discuss two methods that are motivated by

completely different criteria than controlling a combined Type I error rate. These two techniques do *not* control the experimentwise error rate or any of the more stringent error rates, and you should not use them with the expectation that they do. You should only use them when the situation and assumptions under which they were developed are appropriate for your experimental analysis.

Suppose that you believe *a priori* that the overall null hypothesis H_0 is less and less likely to be true as the number of treatments increases. Then the strength of evidence required to reject H_0 should decrease as the number of groups increases. Alternatively, suppose that there is a quantifiable penalty for each incorrect (pairwise comparison) decision we make, and that the total loss for the overall test is the sum of the losses from the individual decisions. Under either of these assumptions, the Duncan multiple range (given below) or something like it is appropriate. Note by comparison that the procedures that control combined Type I error rates require more evidence to reject H_0 as the number of groups increases, while Duncan's method requires less. Also, a procedure that controls the experimentwise error rate has a penalty of 1 if there are any rejections when H_0 is true and a penalty of 0 otherwise; this is very different from the summed loss that leads to Duncan's multiple range.

Duncan's multiple range (sometimes called Duncan's test or Duncan's new multiple range) is a step-down Studentized range method. You specify a "protection level" $\mathcal{E}$ and proceed in step-down fashion using

$$u = u(\mathcal{E}, \nu, k, g) = q_{1-(1-\mathcal{E})^{k-1}}(k, \nu)/\sqrt{2}$$

for the critical values. Notice that $\mathcal{E}$ is the comparisonwise error rate for testing a stretch of length 2, and the experimentwise error rate will be $1 - (1 - \mathcal{E})^{g-1}$, which can be considerably more than $\mathcal{E}$. Thus *fixing Duncan's protection level at $\mathcal{E}$ does **not** control the experimentwise error rate or any more stringent rate*. Do not use Duncan's procedure if you are interested in controlling any of the combined Type I error rates.

As a second alternative to combined Type I error rates, suppose that our interest is in predicting future observations from the treatment groups, and that we would like to have a prediction method that makes the average squared prediction error small. One way to do this prediction is to first partition the g treatments into p classes, $1 \le p \le g$; second, find the average response in each of these p classes; and third, predict a future observation from a treatment by the observed mean response of the class for the treatment. We thus look for partitions that will lead to good predictions.

One way to choose among the partitions is to use AICc. Partitions with low values of AICc should give better predictions.

This predictive approach makes no attempt to control any Type I error rate; in fact, the Type I error rate is .15 or greater even for $g = 2$ groups! This approach is useful when prediction is the goal, but can be quite misleading if interpreted as a test of H_0.

5.4.10   Confident directions

In our heart of hearts, we often believe that all treatment means differ when examined sufficiently precisely. Thus our concern with null hypotheses H_{0ij} is misplaced. As an alternative, we can make statements of *direction*. After having collected data, we consider μ_i and μ_j; assume $\mu_i < \mu_j$. We could decide from the data that $\mu_i < \mu_j$, or that $\mu_i > \mu_j$, or that we don't know—that is, we don't have enough information to decide. These decisions correspond in the testing paradigm to rejecting H_{0ij} in favor of $\mu_i < \mu_j$, rejecting H_{0ij} in favor of $\mu_j < \mu_i$, and failing to reject H_{0ij}. In the confident directions framework, only the decision $\mu_i > \mu_j$ is an error. See Tukey (1991).

   Confident directions procedures are pairwise comparisons testing procedures, but with results interpreted in a directional context. Confident directions procedures bound error rates when making statements about direction. If a testing procedure bounds an error rate at $\mathcal{E}$, then the corresponding confident directions procedure bounds a confident directions error rate at $\mathcal{E}/2$, the factor of 2 arising because we cannot falsely reject in the correct direction.

   Let us reinterpret our usual error rates in terms of directions. Suppose that we use a pairwise comparisons procedure with error rate bounded at $\mathcal{E}$. In a confident directions setting, we have the following:

Strong familywise	The probability of making any incorrect statements of direction is bounded by $\mathcal{E}/2$.
FDR	Incorrect statements of direction will on average be no more than a fraction $\mathcal{E}/2$ of the total number of statements of direction.
Experimentwise	The probability of making any incorrect statements of direction when all the means are very nearly equal is bounded by $\mathcal{E}/2$.
Comparisonwise	The probability of making an incorrect statement of direction for a given comparison is bounded by $\mathcal{E}/2$.

There is no directional analog of simultaneous confidence intervals, so procedures that produce simultaneous intervals should be considered procedures that control the strong familywise error rate (which they do).

5.5   Comparison with Control or the Best

There are some situations where we do not do all pairwise comparisons, but rather make comparisons between a control and the other treatments, or the best responding treatment (highest or lowest average) and the other treatments. For example, you may be producing new standardized mathematics tests for elementary school children, and you need to compare the new tests with the current test to assure comparability of the results. The procedures for comparing to a control or the best are similar.

5.5.1 Comparison with a control

Suppose that there is a special treatment, say treatment g, with which we wish to compare the other $g-1$ treatments. Typically, treatment g is a control treatment. The Dunnett procedure allows us to construct simultaneous $1-\mathcal{E}$ confidence intervals on $\mu_i - \mu_g$, for $i = 1, \ldots, g-1$ when all sample sizes are equal via

$$\bar{y}_i - \bar{y}_g \pm d_\mathcal{E}(g-1, \nu) \sqrt{MS_E} \sqrt{\frac{1}{n_i} + \frac{1}{n_g}} \ ,$$

where ν is the degrees of freedom for MS_E. The value $d_\mathcal{E}(g-1, \nu)$ is tabulated in Appendix Table C.9. These table values are exact when all sample sizes are equal and only approximate when the sizes are not equal.

For testing, we can use

$$u(\mathcal{E}, i, j) = d_\mathcal{E}(g-1, \nu) \ ,$$

which controls the strong familywise error rate and leads to

$$DSD = d_\mathcal{E}(g-1, \nu) \sqrt{MS_E} \sqrt{\frac{1}{n_i} + \frac{1}{n_g}} \ ,$$

the Dunnett significant difference. There is also a step-down modification that still controls the strong familywise error rate and is slightly more powerful. We have $g-1$ t-statistics. Compare the largest (in absolute value) to $d_\mathcal{E}(g-1, \nu)$. If the test fails to reject the null, stop; otherwise compare the second largest to $d_\mathcal{E}(g-2, \nu)$ and so on.

There are also one-sided versions of the confidence and testing procedures. For example, you might reject the null hypothesis of equality only if the noncontrol treatments provide a higher response than the control treatments. For these, test using the critical value

$$u(\mathcal{E}, i, j) = d'_\mathcal{E}(g-1, \nu) \ ,$$

tabulated in Appendix Table C.9, or form simultaneous one-sided confidence intervals on $\mu_i - \mu_g$ with

$$\bar{y}_i - \bar{y}_g \geq d'_\mathcal{E}(g-1, \nu) \sqrt{MS_E} \sqrt{\frac{1}{n_i} + \frac{1}{n_g}}.$$

For t-critical values, a one-sided cutoff is equal to a two-sided cutoff with a doubled $\mathcal{E}$. The same is not true for Dunnett critical values, so that $d'_\mathcal{E}(g-1, \nu) \neq d_{2\mathcal{E}}(g-1, \nu)$.

Example 5.9 Alfalfa meal and turkeys

An experiment is conducted to study the effect of alfalfa meal in the diet of male turkey poults (chicks). There are nine treatments. Treatment 1 is a control treatment; treatments 2 through 9 contain alfalfa meal of two different types in differing proportions. Units consist of 72 pens of eight birds each, so there are eight pens per treatment. One response of interest is average daily weight gains per bird for birds aged 7 to 14 days. We would like to know which alfalfa treatments are significantly different from the control in weight gain, and which are not.

Here are the average weight gains (g/day) for the nine treatments:

$$22.668 \quad 21.542 \quad 20.001 \quad 19.964 \quad 20.893$$
$$21.946 \quad 19.965 \quad 20.062 \quad 21.450$$

The MS_E is 2.487 with 55 degrees of freedom. (The observant student will find this degrees of freedom curious; more on this data set later.) Two-sided, 95% confidence intervals for the differences between control and the other treatments are computed using

$$d_{\mathcal{E}}(g-1,\nu) \sqrt{MS_E} \sqrt{\frac{1}{n_i} + \frac{1}{n_g}} = d_{.05}(8,55) \sqrt{2.487} \sqrt{\frac{1}{8} + \frac{1}{8}}$$

$$= 2.74 \times 1.577/2$$

$$= 2.16 \ .$$

Any treatment with mean less than 2.16 from the control mean of 22.668 is not significantly different from the control. These are treatments 2, 5, 6, and 9.

We apply this method in **R** using the `compare.to.control` function from `cfcdae`.

```
1 > compare.to.control(fit,treat,control=1)
           difference      lower      upper
  * 4 - 1   -2.703750  -4.863788  -0.5437122
  * 7 - 1   -2.702500  -4.862538  -0.5424622
  * 3 - 1   -2.666250  -4.826288  -0.5062122
  * 8 - 1   -2.605000  -4.765038  -0.4449622
    5 - 1   -1.775000  -3.935038   0.3850378
    9 - 1   -1.211451  -3.456599   1.0336974
    2 - 1   -1.125000  -3.285038   1.0350378
    6 - 1   -0.721250  -2.881288   1.4387878
```

It is a good idea to give the control (treatment g) greater replication than the other treatments. The control is involved in every comparison, so it makes sense to estimate its mean more precisely. More specifically, if you had a fixed number of units to spread among the treatments, and you wished to minimize the average variance of the differences $\overline{y}_{g\bullet} - \overline{y}_{i\bullet}$, then you would do best when the ratio n_g/n_i is about equal to $\sqrt{g-1}$.

Give the control more replication

Personally, I rarely use the Dunnett procedure, because I nearly always get the itch to compare the noncontrol treatments with each other as well as with the control.

5.5.2   Comparison with the best

Suppose that the goal of our experiment is to screen a number of treatments and determine those that give the best response—to pick the winner. The multiple comparisons with best (MCB) procedure produces two results:

- It produces a subset of treatments that cannot be distinguished from the best; the treatment having the true largest mean response will be in this subset with probability $1 - \mathcal{E}$.

- It produces simultaneous $1 - \mathcal{E}$ confidence intervals on $\mu_i - \max_{j \neq i} \mu_j$, the difference between a treatment mean and the best of the other treatment means.

The subset selection procedure is the more useful product, so we only discuss the selection procedure.

The best subset consists of all treatments i such that

$$\bar{y}_{i\bullet} > \bar{y}_{j\bullet} - d'_{\mathcal{E}}(g-1, \nu) \sqrt{\text{MS}_{\text{E}}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \quad \text{for all } j \neq i$$

In words, treatment i is in the best subset if its mean response is greater than the largest treatment mean less a one-sided Dunnett allowance. When small responses are good, a treatment i is in the best subset if its mean response is less than the smallest treatment mean plus a one-sided Dunnett allowance.

Example 5.10 Weed control in soybeans

Weeds reduce crop yields, so farmers are always looking for better ways to control weeds. Fourteen weed control treatments were randomized to 56 experimental plots that were planted in soybeans. The plots were later visually assessed for weed control, the fraction of the plot without weeds. The percent responses are given in Table 5.3. We are interested in finding a subset of treatments that contains the treatment giving the best weed control (largest response) with confidence 99%.

For reasons that will be explained in Chapter 6, we will analyze as our response the square root of percent weeds (that is, 100 minus the percent weed control). Because we have subtracted weed control, small values of the transformed response are good. On this scale, the fourteen treatment means are

$$\begin{array}{ccccccc} 1.000 & 2.616 & 2.680 & 2.543 & 2.941 & 1.413 & 1.618 \\ 2.519 & 2.847 & 1.618 & 1.000 & 4.115 & 4.988 & 5.755 \end{array}$$

and the MS_{E} is .547 with 42 degrees of freedom. The smallest treatment

Table 5.3: Percent weed control in soybeans under 14
treatments. Data set `WeedControl`.

1	2	3	4	5	6	7
99	95	92	95	85	98	99
99	92	95	88	92	99	95
99	95	92	95	92	95	99
99	90	92	95	95	99	95

8	9	10	11	12	13	14
95	92	99	99	88	65	75
85	90	95	99	88	65	50
95	95	99	99	85	92	72
97	90	95	99	68	72	68

mean is 1.000, and the Dunnett allowance is

$$
\begin{aligned}
d'_{\mathcal{E}}(g-1,\nu)\;\sqrt{\mathrm{MS_E}}\;\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} \;&=\; d'_{.01}(13,42)\;\sqrt{.547}\;\sqrt{\frac{1}{4}+\frac{1}{4}} \\
&=\; 3.29\times.740\times.707 \\
&=\; 1.72.
\end{aligned}
$$

So, any treatment with a mean of $1+1.72 = 2.72$ or less is included in the
99% grouping. These are treatments 1, 2, 3, 4, 6, 7, 8, 10, and 11.

In **R**, we may use the `compare.to.best` function from `cfcdae`:

```
1 > fit <- lm(sqrt(100-pct.control)~treatment,data=WeedControl)
2 > compare.to.best(fit,treatment,lowisbest = TRUE,confidence = .99)
               difference allowance
  * 14 - 1   4.754856e+00  1.716888
  * 13 - 1   3.988022e+00  1.716888
  * 12 - 1   3.114510e+00  1.716888
  * 5 - 1    1.941476e+00  1.716888
  * 9 - 1    1.847263e+00  1.716888
    3 - 1    1.680337e+00  1.716888
    2 - 1    1.615710e+00  1.716888
    4 - 1    1.543076e+00  1.716888
    8 - 1    1.519293e+00  1.716888
    10 - 1   6.180340e-01  1.716888
    7 - 1    6.180340e-01  1.716888
    6 - 1    4.125704e-01  1.716888
    11 - 1   3.996803e-15  1.716888
  best is 1 0.000000e+00        NA
```

We can see that only treatments 5, 9, 12, 13, and 14 are outside the "best"
group.

5.6 Reality Check on Coverage Rates

We already pointed out that the error rate control for some multiple comparisons procedures is only approximate if the sample sizes are not equal or the tests are dependent. However, even in the "exact" situations, these procedures depend on assumptions about the distribution of the data for the coverage rates to hold: for example normality or constant error variance. These assumptions are often violated—data are frequently nonnormal and error variances are often nonconstant.

Violation of distributional assumptions usually leads to true error rates that are not equal to the nominal $\mathcal{E}$. The amount of discrepancy depends on the nature of the violation. Unequal sample sizes or dependent tests are just another variable to consider.

The point is that we need to get some idea of what the true error is, and not get worked up about the fact that it is not *exactly* equal to $\mathcal{E}$.

> In the real world, coverage and error rates are always approximate.

5.7 A Warning About Conditioning

Except for the protected LSD, the multiple comparisons procedures discussed above do not require the ANOVA F-test to be significant for protection of the experimentwise error rate. They stand apart from the F-test, protecting the experimentwise error rate by other means. In fact, requiring that the ANOVA F-test be significant will alter their error rates.

Bernhardson (1975) reported on how conditioning on the ANOVA F-test being significant affected the per comparison and per experiment error rates of pairwise comparisons, including LSD, HSD, SNK, Duncan's procedure, and Scheffé. Requiring the F to be significant lowered the per comparison error rate of the LSD from 5% to about 1% and lowered the per experiment error rate for HSD from 5% to about 3%, both for 6 to 10 groups. Looking just at those null cases where the F-test rejected, the LSD had a per comparison error rate of 20 to 30% and the HSD per experiment error rate was about 65%—both for 6 to 10 groups. Again looking at just the null cases where the F was significant, even the Scheffé procedure's per experiment error rate increased to 49% for 4 groups, 22% for 6 groups, and down to about 6% for 10 groups.

Requiring the F-test to be significant alters the error rates of pairwise procedures

The problem is that when the ANOVA F-test is significant in the null case, one cause might be an unusually low estimate of the error variance. This unusually low variance estimate gets used in the multiple comparisons procedures leading to smaller than normal HSD's, and so on.

5.8 Some Controversy

Simultaneous inference is deciding which error rate to control and then using an appropriate technique for that error rate. Controversy arises because

- Users cannot always agree on the appropriate error rate.

- Users cannot always agree on what constitutes the appropriate family of tests. Different groupings of the tests lead to different results.

- Standard statistical practice seems to be inconsistent in its application of multiple comparisons ideas. For example, multiple comparisons are fairly common when comparing treatment means, but almost unheard of when examining multiple factors in factorial designs (see Chapter 8).

You as experimenter and data analyst must decide what is the proper approach for inference. See Carmer and Walker (1982) for an amusing allegory on this topic.

More philosophically, multiple comparisons procedures violate the *likelihood principle*, which says that given a statistical model, all of the evidence you have for unknown parameters can be derived from the likelihood function; see Berger and Wolpert (1988) for an extended discussion of the likelihood principle. Confidence intervals and p-values violate the likelihood principle by referencing results obtained to results that might have been obtained rather than simply relying on the likelihood. However, multiple comparisons adjustments go the extra mile by potentially involving other data not directly relevant to the issue at hand (not just other potential outcomes of the experiment at hand, as we do for confidence intervals and p-values) and by accounting for the intentions of the investigator (see Berry 1988).

Likelihood principle

5.9 Further Reading and Extensions

There is much more to the subject of multiple comparisons than what we have discussed here. For example, many procedures for contrasts can be adapted to other linear combinations of parameters, and many of the pairwise comparisons techniques can be adapted to contrasts. A good place to start is Miller (1981), an instant classic when it appeared and still an excellent and readable reference; much of the discussion here follows Miller. Hochberg and Tamhane (1987) contains some of the more recent developments.

The first multiple comparisons technique appears to be the LSD suggested by Fisher (1935). Curiously, the next proposal was the SNK (though not so labeled) by Newman (1939). Multiple comparisons then lay dormant till around 1950, when there was an explosion of ideas: Duncan's multiple range procedure (Duncan 1955), Tukey's HSD (Tukey 1952), Scheffé's all contrasts method (Scheffé 1953), Dunnett's method (Dunnett 1955), and another proposal for SNK (Keuls 1952). The pace of introduction then slowed again. The REGW procedures appeared in 1960 and evolved through the

1970's (Ryan 1960; Einot and Gabriel 1975; Welsch 1977). Improvements in the Bonferroni inequality lead to the modified Bonferroni procedures in the 1970's and later (Holm 1979; Simes 1986; Hochberg 1988; Benjamini and Hochberg 1995).

Procedures sometimes predate a careful understanding of the error rates they control. For example, SNK has often been advocated as a less conservative alternative to the HSD, but the false discovery rate was only defined recently (Benjamini and Hochberg 1995). Furthermore, many textbook introductions to multiple comparisons procedures do not discuss the different error rates, thus leading to considerable confusion over the choice of procedure.

One historical feature of multiple comparisons is the heavy reliance on tables of critical values and the limitations imposed by having tables only for selected percent points or equal sample sizes. Computers and software remove many of these limitations. For example, the software in Lund and Lund (1983) can be used to compute percent points of the Studentized range for $\mathcal{E}$'s not usually tabulated, while the software in Dunnett (1989) can compute critical values for the Dunnett test with unequal sample sizes. When no software for exact computation is available (for example, Studentized range for unequal sample sizes), percent points can be approximated through simulation (see, for example, Ripley 1987).

Hayter (1984) has shown that the Tukey-Kramer adjustment to the HSD procedure is conservative when the sample sizes are not equal.

The issue of optional stopping before testing is important in frequentist statistics. One approach is the Sequential Probability Ratio Test (Wald 1947, Ghosh and Sen 1991). This procedure allows you to test as each data point arrives and still control your Type I error rate. Multiple testing over time is crucial in clinical trials, where it sometimes goes by the name *interim analysis*, and group-sequential designs (data arrive in blocks rather than one at a time) are common. Some classical examples in this regard are found in Pocock 1977, O'Brien and Fleming 1979, and Kim and DeMets 1987.

Berry and Hochberg (1999) provide a fairly gentle introduction to some Bayesian perspectives on multiple comparisons. One outcome discussed there is that a Bayesian might make decisions that look like multiple comparisons are being ignored or accommodated, depending on the kinds of information available and the form of the model. These authors also present an example of a prior distribution for means in the separate means model that does allow for exact equality of some means as well as the computation of a posterior probability that some means are the same.

Neath and Cavanaugh (2006) provide an example approaching Bayesian multiple comparisons as a model selection problem. This is somewhat analogous to the AICc method mentioned in Section 5.4.9.

Duncan (1961) introduced a Bayesian approach to pairwise comparisons, extended and simplified in Waller and Duncan (1969). (This is the same Duncan as the new multiple range test, but a different procedure and different justification.) This approach is a Bayesian decision problem directional differences (greater than or less than). For any pair of means μ_i and μ_j, there are

two decisions to make. In the first decision, we must decide between $\mu_j \leq \mu_i$ and $\mu_j > \mu_i$. In the second decision, we must decide between $\mu_i \leq \mu_j$ and $\mu_i > \mu_j$. If we decide $\mu_j > \mu_i$ when $\mu_j \leq \mu_i$ is true, our loss is $k_1|\mu_i - \mu_j|$. On the other hand, if we decide $\mu_j \leq \mu_i$ when $\mu_j > \mu_i$ is true, our loss is $k_2|\mu_i - \mu_j|$. The overall loss is the sum of the losses for each of the $g(g-1)$ component decisions. In the end, for every pair of treatment means, we have either decided $\mu_j > \mu_i$ or $\mu_i > \mu_j$ or that we cannot rank them relative to each other (analogous to "not significantly different").

The final decision rule takes the familiar form of declare two means different if $|\bar{x}_i - \bar{x}_j| > u\sqrt{\text{MS}_\text{E}}\sqrt{1/n_i + 1/n_j}$, where u depends on the ratio of relative costs of type I and II errors $k = k_1/k_2$, your prior knowledge about the variability of treatment means relative to error variance, and degrees of freedom. The multiplier u increases as k increases and decreases as the prior ratio of variability among versus within treatments increases. Increasing k is like decreasing $\mathcal{E}$ in frequentist comparison; lower values of the ratio of between to within variability lead to more stringent cutoffs, while higher values are more like a per-comparison procedure.

One hesitates to wade into the discussion as to whether data collection stopping rules affect Bayesian analysis. Subject matter experts have been discussing this for decades, with a flurry of recent articles as Bayesian approaches have grown to have wider use. Chapter 4 of Berger and Wolpert (1988) has an extended discussion of the role of stopping rules and their interaction with likelihoods. Similarly, Lindsey (1997) also explores some of the more arcane corners. The issues they discuss revolve around what does it really mean to have an ignorable, or non-informative, stopping rule. Psychologists have been prolific of late in their discussion of stopping rules. For example, Wagenmakers (2007) has a fairly gentle introduction of some of the issues. Rouder (2014) makes some categorical claims along with examples (and with the title "Optional stopping: No problem for Bayesians" it's pretty clear which side he is on), but (Sanborn and Hills 2013) argue the reverse. Blog posts on this subject are lively.

5.10   Problems

We have five groups and three observations per group. The group means are 6.5, 4.5, 5.7, 5.6, and 5.1, and the mean square for error is .75. Compute simultaneous confidence intervals (95% level) for the differences of all treatment pairs.

Exercise 5.1

Consider a completely randomized design with five treatments, four units per treatment, and treatment means

Exercise 5.2

$$3.2892 \quad 10.256 \quad 8.1157 \quad 8.1825 \quad 7.5622 \quad .$$

The MSE is 4.012.

(a) Construct an ANOVA table for this experiment and test the null hypothesis that all treatments have the same mean.

(b) Test the null hypothesis that the average response in treatments 1 and 2 is the same as the average response in treatments 3, 4, and 5.

(c) Use the HSD procedure to compare the means of the five treatments.

Refer to the data in Problem 3.1. Test the null hypothesis that all pairs of workers produce solder joints with the same average strength against the alternative that some workers produce different average strengths. Control the strong familywise error rate at .05.

Exercise 5.3

Refer to the data in Exercise 3.1. Test the null hypothesis that all pairs of diets produce the same average weight liver against the alternative that some diets produce different average weights. Control the FDR at .05.

Exercise 5.4

Use the data from Exercise 3.3. Compute 95% simultaneous confidence intervals for the differences in response between the three treatment groups (acid, pulp, and salt) and the control group.

Exercise 5.5

Use the data from Problem 3.2. Use the Tukey procedure to make all pairwise comparisons between the treatment groups. Summarize your results with an underline diagram.

Problem 5.1

Use the data and context from Problem 3.8. We want to know which of the five treatments differ in mean. Choose one of the error rates, and provide a justification for why you believe that is the correct error rate in this context. Using that error rate, do all pairwise comparisons and report your results.

Problem 5.2

Use the data from Problem 3.9. Test whether the extravert or introvert conditions differ from control in terms of perception of power.

Problem 5.3

In an experiment with four groups, each with five observations, the group means are 12, 16, 21, and 19, and the MSE is 20. A colleague points out that the contrast with coefficients -4, -2, 3, 3 has a rather large sum of squares. No one knows to begin with why this contrast has a large sum of squares, but after some detective work, you discover that the contrast coefficients are roughly the same (except for the overall mean) as the time the samples had to wait in the lab before being analyzed (3, 5, 10, and 10 days). What is the significance of this contrast?

Problem 5.4

Consider an experiment taste-testing six types of chocolate chip cookies: 1 (brand A, chewy, expensive), 2 (brand A, crispy, expensive), 3 (brand B, chewy, inexpensive), 4 (brand B, crispy, inexpensive), 5 (brand C, chewy, expensive), 6 (brand D, crispy, inexpensive). We will use twenty different raters randomly assigned to each type (120 total raters). I have constructed five preplanned contrasts for these treatments, and I obtain p-values of .03, .04, .23, .47, and .68 for these contrasts. Discuss how you would assess the statistical significance of these contrasts, including what issues need to be resolved.

Problem 5.5

A satellite deorbits and pieces come raining down over northern Canada. Unfortunately, this satellite had a nuclear power cell, so now we have radio isotopes spread all over the wilderness. A small plane flies a gamma ray detector over the area and makes gamma ray counts at 10,000 locations in a 40 by 250 grid. At each location, we test the null hypothesis that the expected

Problem 5.6

gamma ray emission rate is equal to the known background rate, versus the alternative that the expected rate is higher than background. We will send teams to investigate and, if required, clean up the locations that are found to be significantly above background.

Discuss the pros and cons of using methods that control the per comparison error rate, false discovery rate, or strong familywise error rate in this situation. Which would you use?

Problem 5.7

An environmental consulting firm has collected 10 soil samples from each of 55 sites and is testing them for dioxin. There is an accepted "background" level of dioxin, and the issue at hand is determining whether any of the sites is above background concentration when tested against the null that they are at background.

What error rate should they control if they want no more than 5% of the rejections to be incorrect rejections? Explain.

What error rate should they control if they want no more than a 5% chance that any incorrect rejections are made? Explain.

Problem 5.8

A laboratory is trying to find compounds that will increase the "grip" of tires in the winter. The real test is to make tires containing the compound, but that is expensive, and the lab can only afford to do that when there is some indication that the compound might actually work. Before making actual tires, the lab screens a large number of compounds using a cheaper, but less precise, process. For each compound screened, the output of the cheaper process is a p-value for the null hypothesis that the compound does not affect winter grip. Because the lab is testing a lot of compounds, there could be a lot of false positives if they take all compounds with a p-value below 5%. The lab doesn't mind have a small fraction of false positives that go on to the expensive test, but they don't want the fraction of false positives to be larger than 10%.

Choose an error rate that gives the lab the control that they want while still selecting as many active compounds as possible. Defend your choice.

Problem 5.9

An issue of *The Journal of Personality and Social Psychology* published an article on ESP (extra-sensory perception). As we all know, many kinds of ESP have been proposed at various times, and the scientific establishment that *any* single person could exhibit *any* kind of ESP would be headline news.

Suppose that the experiment in JPSP went like this. Twenty student volunteers were each tested for two kinds of ESP: telepathy (the ability to determine the card that the experimenter was thinking about) and precognition (the ability to see an event before it happens). The results of the 40 tests were analyzed with a one-sample z-test. In this case, one student had a p-value less than 5% for telepathy, and one had a p-value less than 5% for precognition.

What statistical issues do you see in terms of evidence for or against ESP?

Question 5.1

In an experiment with five groups and 25 degrees of freedom for error, for what numbers of contrasts is the Bonferroni procedure more powerful than the Scheffé procedure?

Consider using BH or SNK in a pairwise comparison problem. For con- **Question 5.2**
creteness, consider the case $g = 5$ and $n = 5$. Compare the cutoffs for
declaring significance. What do you see? For what configurations of group
means is SNK likely to be more powerful than BH?

Chapter 6

Checking Assumptions

We analyze experimental results by using some set of frequentist or Bayesian methods. Despite the many differences between the methods, they are all based on the *assumption* that our data follow the model

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where the parameters α_i's are unknown and the ϵ_{ij}'s are independent normals with constant variance. We have done nothing to ensure that these assumptions are reasonably accurate.

What we did was random assignment of treatments to units, followed by measurement of the response. As discussed briefly in Chapter 2, randomization methods permit us to make inferences based solely on the randomization, but these methods tend to be computationally tedious and difficult to extend. Model-based methods with distributional assumptions usually yield good approximations to the randomization inferences, provided that the model assumptions are themselves reasonably accurate. If we apply the model-based methods in situations where the model assumptions do not hold, the inferences we obtain may be misleading. We thus need to look to the accuracy of the model assumptions.

Historically, the assumptions of independence, constant variance, and normality were our tickets to dependable inference, and statisticians worked hard to ensure that their data were consistent with those assumptions. It was an inferential case of the adage "If all you have is a hammer, the whole world looks like a nail." Fortunately, the past few decades have seen the arrival of new tools in the toolbox, including robust methods, *generalized linear models* (GLM) and Bayesian methods. These, together with better software, permit us to more frequently analyze data as they are, rather than trying to make them look as we think they should be.

For our standard models, the basic assumptions we need to check are that the errors are 1) independent, 2) normally distributed, and 3) have constant variance. (We also require that the errors have mean zero in every treatment group. That is automatically true for the separate-means model, but it might

not be true for reduced models such as polynomial models.) Independence is the most important of these assumptions, and also the most difficult to accommodate when it fails. For the kinds of models we have been using, normality is the least important assumption, particularly for large sample sizes; see Chapter 10 for a different kind of model that is extremely dependent on normality. Constant variance is intermediate, in that nonconstant variance can have a substantial effect on our inferences, but nonconstant variance can also be accommodated in many situations.

Note that the quality of our inference depends on how well the errors ϵ_{ij} conform to our assumptions, but that we do not observe the errors ϵ_{ij}. The closest we can get to the errors are r_{ij}, the residuals from the full model. Thus we must make decisions about how well the errors meet our assumptions based not on the errors themselves, but instead on residual quantities that we can observe. This unobservable nature of the errors can make diagnosis difficult in some situations.

In any real-world data set, we are almost sure to have one or more of the three assumptions be false. For example, real-world data are never exactly normally distributed. Thus there is no profit in formal testing of our assumptions; we already know that they are not true. The good news is that our procedures can still give reasonable inferences when the departures from our assumptions are not too large. This is called *robustness of validity*, which means that our inferences are reasonably valid across a range of departures from our assumptions. Thus the real question is whether the deviations from our assumptions are sufficiently great to cause us to mistrust our inference. At a minimum, we would like to know in what way to mistrust the inference (for example, our confidence intervals are shorter than they should be), and ideally we would like to be able to correct any problems.

In the remaining sections of this chapter, I would like to (1) scare you into believing that violation of our assumptions can lead to poor inference, (2) give you some tools for diagnosing when there may be a problem with our assumptions, and (3) give you some tools for working with data that fail our assumptions. In many cases, these tools could be books unto themselves, so our discussion of these tools, while perhaps longer than the student may prefer, is still far from comprehensive.

Note: although nearly all of our discussion of assumptions will be done in a frequentist context, Bayesian methods are also vulnerable to model misspecification.

Independence, constant variance, normality

Robustness of validity

6.1   Effects of Incorrect Assumptions

Our methods work as advertised when the data meet our assumptions. Some violations of the assumptions have little effect on the quality of our inference, but others can cause almost catastrophic failure. This section gives an overview of how failed assumptions affect inference.

Table 6.1: Skewness and kurtosis for selected distributions.

Distribution	γ_1	γ_2
Normal	0	0
Uniform	0	-1.2
Normal truncated at		
± 1	0	-1.06
± 2	0	-0.63
Student's t (df)		
5	0	6
6	0	3
8	0	1.5
20	0	.38
Chi-square (df)		
1	2.83	12
2	2	6
4	1.41	3
8	1	1.5

6.1.1 Effects of nonnormality

Before describing the effects of nonnormality, we need some way to quantify the degree to which a distribution is nonnormal. For this we will use the *skewness* and *kurtosis*, which measure asymmetry and tail length respectively. The skewness γ_1 and kurtosis γ_2 deal with third and fourth powers of the data:

$$\gamma_1 = \frac{E[(X-\mu)^3]}{\sigma^3} \quad \text{and} \quad \gamma_2 = \frac{E[(X-\mu)^4]}{\sigma^4} - 3.$$

For a normal distribution, both the skewness and kurtosis are 0. Distributions with a longer right tail have positive skewness, while distributions with a longer left tail have negative skewness. Symmetric distributions, like the normal, have zero skewness. Distributions with longer tails than the normal (more outlier prone) have positive kurtosis, and those with shorter tails than the normal (less outlier prone) have negative kurtosis. The "-3" in the definition of kurtosis is there to make the normal distribution have zero kurtosis. Note that neither skewness nor kurtosis depends on location or scale.

Skewness measures asymmetry

Kurtosis measures tail length

Table 6.1 lists the skewness and kurtosis for several distributions, giving you an idea of some plausible values. We could estimate the skewness and kurtosis for the residuals in our analysis, but these values are of limited diagnostic value, as sample estimates of skewness and kurtosis are notoriously variable.

For our discussion of nonnormal data, we will assume that the distribution of responses in each treatment group is the same apart from different means, but we will allow this common distribution to be nonnormal instead

Table 6.2: Actual Type I error rates for ANOVA F-test with nominal 5% error rate for various sample sizes and values of γ_1 and γ_2 using the methods of Gayen (1950).

Four Samples of Size 5

			γ_2				
γ_1	-1	-.5	0	.5	1	1.5	2
0	.0527	.0514	.0500	.0486	.0473	.0459	.0446
.5	.0530	.0516	.0503	.0489	.0476	.0462	.0448
1	.0538	.0524	.0511	.0497	.0484	.0470	.0457
1.5	.0552	.0538	.0525	.0511	.0497	.0484	.0470

$\gamma_1 = 0$ and $\gamma_2 = 1.5$

4 groups of k		k groups of 5		(k_1, k_1, k_2, k_2)	
k	Error	k	Error	k_1, k_2	Error
2	.0427	4	.0459	10,10	.0480
10	.0480	8	.0474	8,12	.0483
20	.0490	16	.0485	5,15	.0500
40	.0495	32	.0492	2,18	.0588

of requiring it to be normal. Our usual point estimates of group means and the common variance ($\overline{y}_{i\bullet}$ and $\mathrm{MS_E}$ respectively) are still unbiased.

The nominal p-values for F-tests are only slightly affected by moderate nonnormality of the errors. For balanced data sets (where all treatment groups have the same sample size), long tails tend to make the F-tests conservative; that is, the nominal p-value is usually a bit larger than it should be; so we reject the null too rarely. Again for balanced data, short tails will tend to make the F-tests liberal; that is, the nominal p-value is usually a bit smaller than it should be, so that we reject the null too frequently. Asymmetry generally has a smaller effect than tail length on p-values. Unbalanced data sets are less predictable and can be less affected by nonnormality than balanced data sets, or even affected in the opposite direction. The effect of nonnormality decreases quickly with sample size. Table 6.2 gives the true Type I error rate of a nominal 5% F-test for various combinations of sample size, skewness, and kurtosis.

Long tails conservative for balanced data

Short tails liberal for balanced data

The situation is not quite so good for confidence intervals, with skewness generally having a larger effect than kurtosis. When the data are normal, two-sided t-confidence intervals have the correct coverage, and the errors are evenly split high and low. When the data are from a distribution with nonzero skewness, two-sided t-confidence intervals still have approximately the correct coverage, but the errors tend to be to one side or the other, rather than split evenly high and low. One-sided confidence intervals for a mean can be seriously in error. The skewness for a contrast is less than that for a single mean, so the errors will be more evenly split. In fact, for a pairwise comparison when the sample sizes are equal, skewness essentially cancels

Skewness affects confidence intervals

out, and confidence intervals behave much as for normal data.

Individual outliers can so influence both treatment means and the mean square for error that the entire inference can change if repeated excluding the outlier. It may be useful here to distinguish between *robustness* (of validity) and *resistance* (to outliers). Robustness of validity means that our procedures give us inferences that are still approximately correct, even when some of our assumptions (such as normality) are incorrect. Thus we say that the ANOVA F-test is robust, because a nominal 5% F-test still rejects the null in about 5% of all samples when the null is true, even when the data are somewhat nonnormal. A procedure is resistant when it is not overwhelmed by one or a few individual data values. Our linear models methods are somewhat robust, but they are not resistant to outliers.

Outliers, robustness, resistance

6.1.2 Effects of nonconstant variance

When there are $g = 2$ groups and the sample sizes are equal, the Type I error rate of the F-test is very insensitive to nonconstant variance. When there are more than two groups or the sample sizes are not equal, the deviation from nominal Type I error rate is noticeable and can in fact be quite large. The basic facts are as follows:

Nonconstant variance affects F-test p-values

1. If all the n_i's are equal, then the effect of unequal variances on the p-value of the F-test is relatively small.

2. If big n_i's go with big variances, then the nominal p-value will be bigger than the true p-value (we overestimate the variance and get a conservative test).

3. If big n_i's go with small variances, then the nominal p-value will be less than the true p-value (we underestimate the variance and get a liberal test).

We can be more quantitative by using an approximation given in Box (1954). Table 6.3 gives the approximate Type I error rates for the usual F-test when error variance is not constant. Clearly, nonconstant variance can dramatically affect our inference. These examples show (approximate) true type I error rates ranging from under .02 to almost .3; these are deviations from the nominal .05 that cannot be ignored.

Our usual form of confidence intervals uses the $\mathrm{MS_E}$ as an estimate of error. When the error variance is not constant, the $\mathrm{MS_E}$ will overestimate the error for contrasts between groups with small errors and underestimate the error for contrasts between groups with large errors. Thus our confidence intervals will be too long when comparing groups with small errors and too short when comparing groups with large errors. The intervals that are too long will have coverage greater than the nominal $1 - \mathcal{E}$, and vice versa for the intervals that are too short. The degree to which these intervals are too long or short can be arbitrarily large depending on sample sizes, the number of groups, and the group error variances.

Nonconstant variance affects confidence intervals

Table 6.3: Approximate Type I error rate $\mathcal{E}$ for nominal 5% ANOVA F-test when the error variance is not constant.

g	σ_i^2	n_i	$\mathcal{E}$
3	1, 1, 1	5, 5, 5	.05
	1, 2, 3	5, 5, 5	.0579
	1, 2, 5	5, 5, 5	.0685
	1, 2, 10	5, 5, 5	.0864
	1, 1, 10	5, 5, 5	.0954
	1, 1, 10	50, 50, 50	.0748
3	1, 2, 5	2, 5, 8	.0202
	1, 2, 5	8, 5, 2	.1833
	1, 2, 10	2, 5, 8	.0178
	1, 2, 10	8, 5, 2	.2831
	1, 2, 10	20, 50, 80	.0116
	1, 2, 10	80, 50, 20	.2384
5	1, 2, 2, 2, 5	5, 5, 5, 5, 5	.0682
	1, 2, 2, 2, 5	2, 2, 5, 8, 8	.0292
	1, 2, 2, 2, 5	8, 8, 5, 2, 2	.1453
	1, 1, 1, 1, 5	5, 5, 5, 5, 5	.0908
	1, 1, 1, 1, 5	2, 2, 5, 8, 8	.0347
	1, 1, 1, 1, 5	8, 8, 5, 2, 2	.2029

6.1.3 Effects of dependence

When the errors are dependent but otherwise meet our assumptions, our estimates of treatment effects are still unbiased, and the MS_{E} is nearly unbiased for σ^2 when the sample size is large. The big change is that the variance of an average is no longer just σ^2 divided by the sample size. This means that our estimates of standard errors for treatment means and contrasts are biased (whether too large or small depends on the pattern of dependence), so that confidence intervals and tests will not have their claimed error rates. The usual ANOVA F-test will be affected for similar reasons.

Variance of average not σ^2/n for dependent data

Let's be a little more careful. The ANOVA F-test is robust to dependence when considered as a randomization test. This means that averaged across all possible randomizations, the F-test will reject the null hypothesis about the correct fraction of times when the null is true. However, when the original data arise with a dependence structure, certain outcomes of the randomization will tend to have too many rejections, while other outcomes of the randomization will have too few.

F robust to dependence averaged across randomizations

More severe problems can arise when there was no randomization across the dependence. For example, treatments may have been assigned to units at random; but when responses were measured, all treatment 1 units were measured, followed by all treatment 2 units, and so on. Random assignment of treatment to units will not help us, even on average, if there is a strong correlation across time in the measurement errors.

Table 6.4: Error rates $\times 100$ of nominal 95% confidence intervals for $\mu_1 - \mu_2$, when neighboring data values have correlation ρ and data patterns are consecutive or alternate.

					ρ				
	$-.3$	$-.2$	$-.1$	0	.1	.2	.3	.4
Con.	.19	1.1	2.8	5	7.4	9.8	12	14
Alt.	12	9.8	7.4	5	2.8	1.1	.19	.001

Example 6.1 Correlated errors

Consider a situation with two treatments and large, equal sample sizes. Suppose that the units have a time order, and that there is a correlation of ρ between the errors ϵ_{ij} for time-adjacent units and a correlation of 0 between the errors of other pairs. For two treatments, the F-test is equivalent to a t-test. The t-test assumes that the difference of the treatment means has variance $2\sigma^2/n$. The actual variance of the difference depends on the correlation ρ and the temporal pattern of the two treatments.

Consider first two temporal patterns for the treatments; call them consecutive and alternate. In the consecutive pattern, all of one treatment occurs, followed by all of the second treatment. In the alternate pattern, the treatments alternate every other unit. For the consecutive pattern, the actual variance of the difference of treatment means is $2(1 + 2\rho)\sigma^2/n$, while for the alternate pattern the variance is $2(1 - 2\rho)\sigma^2/n$. For the usual situation of $\rho > 0$, the alternate pattern gives a more precise comparison than the consecutive pattern, but the estimated variance in the t-test $(2\sigma^2/n)$ is the same for both patterns and correct for neither. So for $\rho > 0$, confidence intervals in the consecutive case are too short by a factor of $1/\sqrt{1 + 2\rho}$, and the intervals will not cover the difference of means as often as they claim, whereas confidence intervals in the alternate case are too long by a factor of $1/\sqrt{1 - 2\rho}$ and will cover the difference of means more often than they claim.

Table 6.4 gives the true error rates for a nominal 95% confidence interval under the type of serial correlation described above and the consecutive and alternate treatment patterns. These will also be the true error rates for the two-group F-test, and the consecutive results will be the true error rates for a confidence interval for a single treatment mean when the data for that treatment are consecutive.

In contrast, consider randomized assignment of treatments for the same kind of units. We could get consecutive or alternate patterns by chance, but that is very unlikely. Under the randomization, each unit has on average one neighbor with the same treatment and one neighbor with the other treatment, tending to make the effects of serial correlation cancel out. Table 6.5 shows median, upper, and lower quartiles of error rates for $\rho = .4$ and sample sizes from 10 to 100 based on 10,000 simulations. The best and worst case error rates are those from Table 6.4; but we can see in Table 6.5 that most randomizations lead to reasonable error rates, and the deviation from the nominal error rate gets smaller as the sample size increases.

Table 6.5: Median, upper and lower quartiles of error rates $\times$ 100 of nominal 95% confidence intervals for $\mu_1 - \mu_2$ when neighboring data values have correlation .4 and treatments are assigned randomly, based on 10,000 simulations.

			n		
	10	20	30	50	100
Lower quartile	3.7	3.9	4.0	4.2	4.5
Median	4.5	4.8	4.8	4.9	5.0
Upper quartile	6.5	5.7	5.8	5.5	5.4

Here is another way of thinking about the effect of serial correlation when treatments are in a consecutive pattern. Positive serial correlation leads to variances for treatment means that are larger than σ^2/n, say $\sigma^2/(En)$, for $E < 1$. The effective sample size En is less than our actual sample size n, because an additional measurement correlated with other measurements doesn't give us a full unit's worth of new information. Thus if we use the nominal sample size, we are being overly optimistic about how much precision we have for estimation and testing.

Positive serial correlation has a smaller effective sample size

The effects of spatial association are similar to those of serial correlation, because the effects are due to correlation itself, not spatial correlation as opposed to temporal correlation.

6.2 Assessing Violations of Assumptions

Our assumptions of independent, normally distributed errors with constant variance are not true for real-world data. However, our procedures may still give us reasonably good inferences, provided that the departures from our assumptions are not too great. Therefore we *assess* the nature and degree to which the assumptions are violated and take corrective measures if they are needed. The p-value of a formal test of some assumption does not by itself tell us the nature and degree of violations, so formal *testing* is of limited utility. Graphical and numerical assessments are the way to go.

Assess — don't test

Our assessments of assumptions about the errors are based on residuals. The raw residuals r_{ij} are simply the differences between the data y_{ij} and the treatment means $\overline{y}_{i\bullet}$. In later chapters there will be more complicated structures for the means, but the raw residuals are always the differences between the data and the fitted value.

Assessments based on residuals

We sometimes modify the raw residuals to make them more interpretable (see Cook and Weisberg 1982). For example, the variance of a raw residual is $\sigma^2(1-H_{ij})$, so we might divide raw residuals by an estimate of their standard error to put all the residuals on an equal footing. (See below for H_{ij}.) This is the *internally Studentized* residual s_{ij}, defined by

Internally Studentized residual

$$s_{ij} = \frac{r_{ij}}{\sqrt{\mathrm{MS_E}(1 - H_{ij})}} \quad .$$

Internally Studentized residuals have a variance of approximately 1.

Alternatively, we might wish to get a sense of how far a data value is from what would be predicted for it from all the other data. This is the *externally Studentized* residual t_{ij}, defined by

$$t_{ij} = s_{ij} \left(\frac{N - g - 1}{N - g - s_{ij}^2} \right)^{1/2} \quad ,$$

where s_{ij} in this formula is the internally Studentized residual. The externally Studentized residual helps us determine whether a data point follows the pattern of the other data. When the data actually come from our assumed model, the externally Studentized residuals t_{ij} follow a t-distribution with $N - g - 1$ degrees of freedom.

<div style="float:right; text-align:right;">Externally
Studentized
residual</div>

The quantity H_{ij} used in computing s_{ij} (and thus t_{ij}) is called the *leverage* and depends on the model being fit to the data and sample sizes; H_{ij} is $1/n_i$ for the separate treatment means model we are using now. Most statistical software will produce leverages and various kinds of residuals.

<div style="float:right; text-align:right;">Leverage</div>

6.2.1 Assessing nonnormality

The normal probability plot (NPP), is a graphical procedure for assessing normality. We plot the ordered data on the vertical axis against the ordered normal scores on the horizontal axis. For assessing the normality of residuals, we plot the ordered residuals on the vertical axis. If you make an NPP of normally distributed data, you get a more or less straight line. It won't be perfectly straight due to sampling variability. If you make an NPP of nonnormal data, the plot will tend to be curved, and the shape of curvature tells you how the data depart from normality.

<div style="float:right; text-align:right;">Normal
probability plot
(NPP)</div>

Normal scores are the expected values for the smallest, second smallest, and so on, up to the largest data point in a sample that really came from a normal distribution with mean 0 and variance 1. One simple approximation to the normal score for the ith point from a sample of size n is the $(i - 3/8)/(n + 1/4)$ percent point of a standard normal.

<div style="float:right; text-align:right;">Normal scores</div>

In our diagnostic setting, we make a normal probability plot of the residuals from fitting the full model; it generally matters little whether we use raw or Studentized residuals. We then examine this plot for systematic deviation from linearity, which would indicate nonnormality. Figure 6.1 shows prototype normal probability plots for long and short tailed data, data skewed to the left and right, and multi-modal and normal data. All sample sizes are 50.

It takes some practice to be able to look at an NPP and tell whether the deviation from linearity is due to nonnormality or sampling variability, and even with practice there is considerable room for error. It is well worth your time to look at a bunch of plots to get a feel for how they may vary.

<div style="float:right; text-align:right;">Practice!</div>
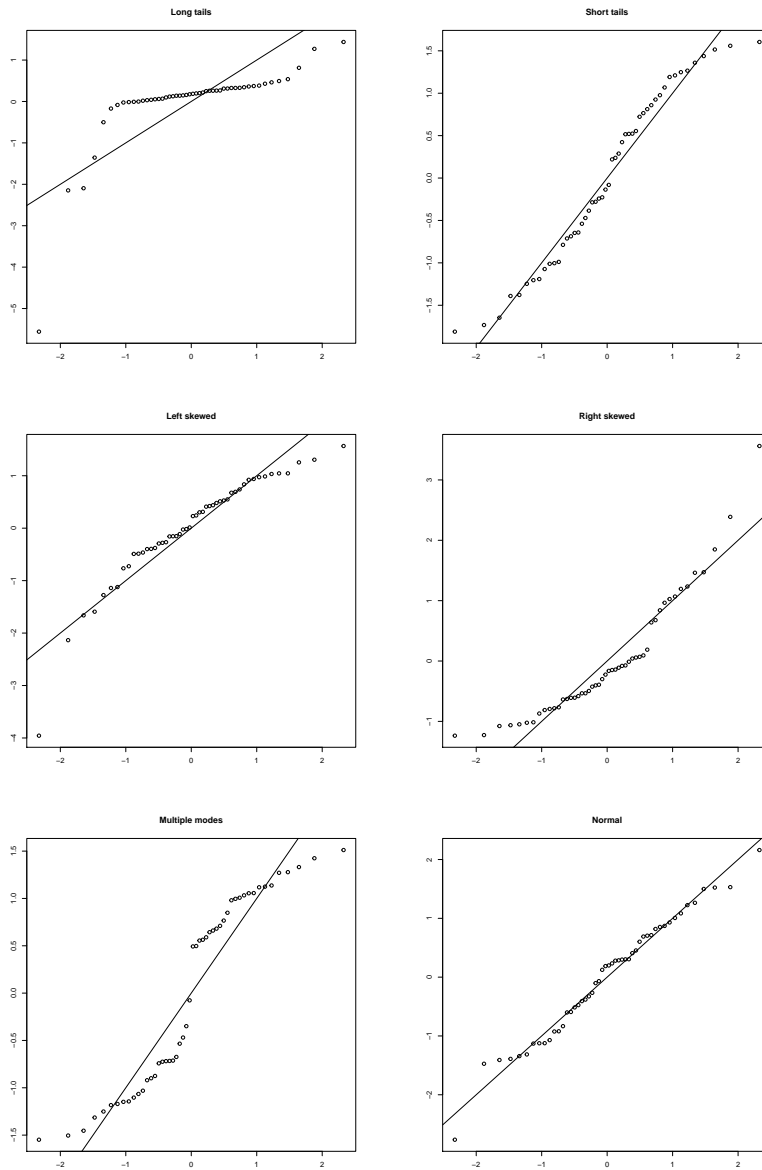
Figure 6.1: Normal probability plots of data with six shapes, with (0,1) line added. All data standardized to zero mean and standard deviation 1.

Outliers are an extreme form of nonnormality. Roughly speaking, an outlier is an observation "different" from the bulk of the data, where different is usually taken to mean far away from or not following the pattern of the bulk of the data. Outliers can show up on an NPP as isolated points in the

Outliers

Table 6.6: Rainfall in acre feet from 52 clouds. Data set
`CloudSeeding`.

Unseeded			Seeded		
1202.6	87.0	26.1	2745.6	274.7	115.3
830.1	81.2	24.4	1697.8	274.7	92.4
372.4	68.5	21.7	1656.0	255.0	40.6
345.5	47.3	17.3	978.0	242.5	32.7
321.2	41.1	11.5	703.4	200.7	31.4
244.3	36.6	4.9	489.1	198.6	17.5
163.0	29.0	4.9	430.0	129.6	7.7
147.8	28.6	1.0	334.1	119.0	4.1
95.0	26.3		302.8	118.3	

corners that lie off the pattern shown by the rest of the data. Some of the points in the corners of the long tails panel of Figure 6.1 would be considered outliers in the context of a model that assumes normally distributed data.

We can use externally Studentized residuals to construct a formal outlier test. Each externally Studentized residual is a test statistic for the null hypothesis that the corresponding data value follows the pattern of the rest of the data, against an alternative that it has a different mean. Large absolute values of the Studentized residual are compatible with the alternative, so we reject the null and declare a given point to be an outlier if that point's Studentized residual exceeds in absolute value the upper $\mathcal{E}/2$ percent point of a t-distribution with $N - g - 1$ degrees of freedom. To test all data values (or equivalently, to test the maximum Studentized residual), make a Bonferroni correction and test the maximum Studentized residual against the upper $\mathcal{E}/(2N)$ percent point of a t-distribution with $N - g - 1$ degrees of freedom. This test can be fooled if there is more than one outlier.

Outlier test

Example 6.2 Cloud seeding

Simpson, Olsen, and Eden (1975) provide data giving the rainfall in acre feet of 52 clouds, 26 of which were chosen at random for seeding with silver oxide. The problem is to determine if seeding has an effect and what size the effect is (if present). Data are given in Table 6.6.

```
1 > fit <- lm(rainfall~seeding,data=CloudSeeding)
2 > anova(fit)
  Analysis of Variance Table

  Response: rainfall
            Df    Sum Sq Mean Sq F value  Pr(>F)
  seeding    1   1000332 1000332   3.993 0.05114 .
  Residuals 50 12526130  250523
  ---
  Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
3 > plot(fit,which=2)
```

Fitting the standard linear model to the data yields a p-value of about .05, giv-
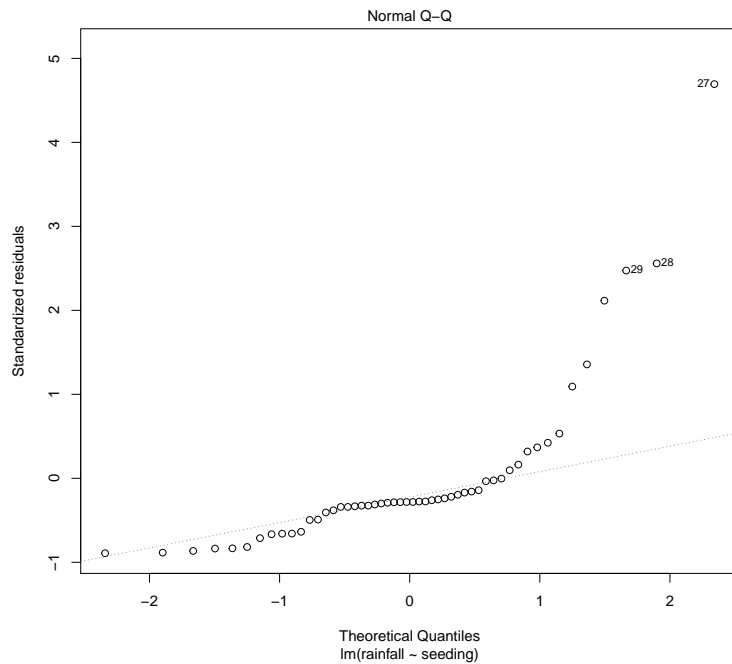
Figure 6.2: Normal probability plot for cloud seeding data.

ing weak evidence of a difference between the treatments. However, before making conclusions, we need to check assumptions.

Figure 6.2 shows an NPP for the cloud seeding data residuals. The plot is angled with the bend in the lower right corner, indicating that the residuals are skewed to the right. This skewness is pretty evident if you make box-plots of the data, or simply look at the data in Table 6.6.

The `plot` method for a fitted linear model can display several types of plots. By default, it shows four different plots, including the normal probability plot. Using `which=2` selects the NPP.

Now compute the externally Studentized residuals.

```
4 > rstudent(fit)
   [1]  2.19411061  1.36772091  0.41990969  0.36539566  0.31620786  0.16082073
   [7] -0.00320393 -0.03386275 -0.14038827 -0.15653518 -0.16824324 -0.19388491
  [13] -0.23670624 -0.24923426 -0.25832867 -0.27369111 -0.27449976 -0.27914975
  [19] -0.27955412 -0.28299133 -0.28845085 -0.29734898 -0.30908053 -0.32243340
  [25] -0.32243340 -0.33032537  6.21229095  2.71710764  2.61380992  1.09427439
  [31]  0.52877750  0.09504057 -0.02417313 -0.21770866 -0.28096164 -0.33780600
  [37] -0.33780600 -0.37769688 -0.40302714 -0.48785195 -0.49211882 -0.63264837
  [43] -0.65430019 -0.65573067 -0.66186225 -0.70871745 -0.81506307 -0.83132908
  [49] -0.83400703 -0.86266333 -0.88289268 -0.89032932
5 > car::outlierTest(fit)
     rstudent unadjusted p-value Bonferonni p
  27 6.212291          1.1009e-07    5.7245e-06
```

The largest (corresponding to 2745.6) is 6.21, and is well beyond any reasonable cutoff for being an outlier. The next largest Studentized residual is 2.71. The outlier test can be automated using `car::outlierTest` as shown in line 5.

```
6 > use <- rep(TRUE,52);use[27] <- FALSE
7 > fit.27 <- lm(rainfall~seeding,data=CloudSeeding,subset=use)
8 > car::outlierTest(fit.27)
     rstudent unadjusted p-value Bonferonni p
  28 4.214633          0.00010977    0.0055984
  29 4.038566          0.00019293    0.0098392
```

If we remove the outlier from the data set and reanalyze, we now find that the largest Studentized residual is 4.21, corresponding to 1697.5. This has a Bonferroni p-value of about .006 for the outlier test. This is an example of *masking*, where one apparently outlying value can hide a second. If we remove this second outlier and repeat the analysis, we now find that 1656 has a Studentized residual of 5.35, again an "outlier". Still more data values will be indicated as outliers as we pick them off one by one.

The problem we have here is not so much that the data are mostly normal with a few outliers, but that the data do not follow a normal distribution at all. The outlier test is based on normality, and doesn't work well for nonnormal data.

6.2.2 Assessing nonconstant variance

There are formal tests for equality of variance—*do not use them!* This is for two reasons. First, p-values from such tests do not tell us what we need to know: the amount of nonconstant variance that is present and how it affects our inferences. Second, classical tests of constant variance (such as Bartlett's test or Hartley's test) are *so incredibly sensitive* to nonnormality that their inferences are worthless in practice.

Don't test equality of variances

We will look for nonconstant variance that occurs when the responses within a treatment group all have the same variance σ_i^2, but the variances differ between groups. We cannot distinguish nonconstant variance within a treatment group from nonnormality of the errors.

Does variance differ by treatment?

We assess nonconstant variance by making a plot of the residuals r_{ij} (or s_{ij} or t_{ij}) on the vertical axis against the fitted values $y_{ij} - r_{ij} = \overline{y}_{i\bullet}$ on the horizontal axis. This plot will look like several vertical stripes of points, one stripe for each treatment group. If the variance is constant, the vertical spread in the stripes will be about the same. Nonconstant variance is revealed as a pattern in the spread of the residuals. Note that groups with larger sample sizes will tend to have some residuals with slightly larger absolute values, simply because the sample size is bigger. It is the overall pattern that we are looking for.

Residual plots reveal nonconstant variance

The most common deviations from constant variance are those where the residual variation depends on the mean. Usually we see variances increasing as the mean increases, but other patterns can occur. When the variance increases with the mean, the residual plot has what is called a right-opening megaphone shape; it's wider on the right than on the left. When the variance decreases with the mean, the megaphone opens to the left. A third possible shape arises when the responses are proportions; proportions around .5 tend to have more variability than proportions near 0 or 1. Other shapes are possible, but these are the most common.

Right-opening megaphone is most common nonconstant variance

A variation on the residuals versus fitted plot that can be helpful is to use the square root of the absolute value of the Studentized residuals instead of the residuals themselves. On this scale, the large residuals get pulled in and the small residuals get pulled up toward 1. These transformed residuals are more clumped and symmetrically distributed, sometimes making nonconstant variance more evident. **R** calls this a Scale-Location plot.

If you absolutely must test equality of variances—for example if change of variance is the treatment effect of interest—Conover, Johnson, and Johnson (1981) suggest a modified Levene test. Let y_{ij} be the data. First compute $\tilde{y}_i$, the median of the data in group i; then compute $d_{ij} = |y_{ij} - \tilde{y}_i|$, the absolute deviations from the group medians. Now treat the d_{ij} as data, and use the ANOVA F-test to test the null hypothesis that the groups have the same average value of d_{ij}. This test for means of the d_{ij} is equivalent to a test for the equality of standard deviations of the original data y_{ij}. The Levene test as described here is a general test and is not tuned to look for specific kinds of nonconstant variance, such as right-opening megaphones. Just as contrasts and polynomial models are more focused than ANOVA, corresponding variants of ANOVA in the Levene test may be more sensitive to specific ways in which constant variance can be violated.

Levene test

Example 6.3 Resin lifetimes, continued

In Example 3.2 we analyzed the $\log_{10}$ lifetimes of an encapsulating resin under different temperature stresses. What happens if we look at the lifetimes on the original scale rather than the log scale?

```
1 > fit <- lm(Time~temp.factor,data=ResinLifetimes)
2 > plot(fit,which=1)
3 > plot(fit,which=3)
4 > car::leveneTest(fit)
  Levene's Test for Homogeneity of Variance (center = median)
        Df F value  Pr(>F)
  group  4  2.3746 0.07286 .
        32
  ---
  Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Line 1 fits the model using the data on the original scale. Line 2 produces the residual plot shown in Figure 6.3. A right-opening megaphone shape is clear, showing that the variability of the residuals increases with the response mean. Line 3 produces the scale-location plot (Figure 6.4), and nonconstant scale is again evident. These scale-assessing residual plots are types 1 and 3 (`which=1` or `which=3`) in **R**.

Although we typically do not test for constant variance, the Levene test is available in the `car` package as shown on line 4. The Levene test for the null hypothesis of constant variance has a p-value of about .07, which is weak evidence against constant variance, even though nonconstant variance is evident in the plots.

6.2.3 Assessing dependence

Serial dependence or *autocorrelation* is one of the more common ways that independence can fail. Serial dependence arises when results close in time tend to be too similar (*positive* dependence) or too dissimilar (*negative* dependence). Positive dependence is far more common. Serial dependence could result from a "drift" in the measuring instruments, a change in skill of the experimenter, changing environmental conditions, and so on. If there is no idea of time order for the units, then there can be no serial dependence.

Serial dependence

A graphical method for detecting serial dependence is to plot the residuals on the vertical axis versus time sequence on the horizontal axis. The plot is sometimes called an *index plot* (that is, residuals-against-time index). Index plots give a visual impression of whether neighbors are too close together (positive dependence), or too far apart (negative dependence). Positive dependence appears as drifting patterns across the plot, while negatively dependent data have residuals that center at zero and rapidly alternate positive and negative.

Index plot to detect serial dependence

The Durbin-Watson statistic is a simple numerical method for checking serial dependence. Let r_k be the residuals sorted into time order. Then the Durbin-Watson statistic is:

Durbin-Watson statistic to detect serial dependence

$$DW = \frac{\sum_{k=1}^{n-1}(r_k - r_{k+1})^2}{\sum_{k=1}^{n} r_k^2}.$$

If there is no serial correlation, the DW should be about 2, give or take sampling variation. Positive serial correlation will make DW less than 2, and
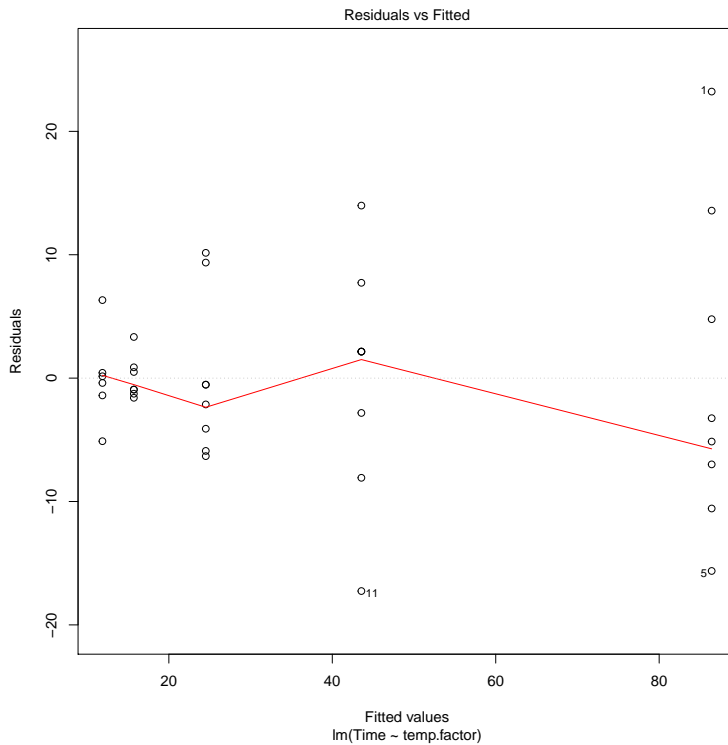
Figure 6.3: Residuals versus fitted values plot for resin lifetime data.

negative serial correlation will make DW more than 2. As a rough rule, serial correlations corresponding to DW outside the range 1.5 to 2.5 are large enough to have a noticeable effect on our inference techniques. Note that DW itself is random and may be outside the range 1.5 to 2.5, even if the errors are uncorrelated. For data sets with long runs of units from the same treatment, the variance of DW is a bit less than $4/N$.

Example 6.4 Temperature differences between thermocouples

Christensen and Blackwood (1993) provide data from five thermocouples that were inserted into a high-temperature furnace to ascertain their relative bias. Sixty-four temperature readings were taken using each thermocouple, with the readings taken simultaneously from the five devices. Table 6.7 gives the differences between thermocouples 3 and 5. We can estimate the relative bias by the average of the observed differences.
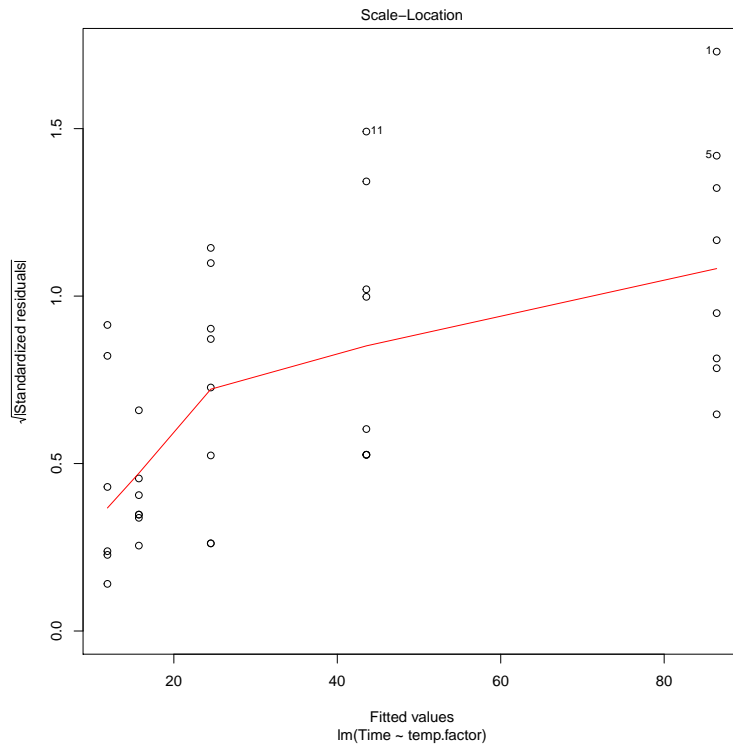
Figure 6.4: Scale-location plot for resin lifetime data.

Table 6.7: Temperature differences in degrees Celsius between two thermocouples for 64 consecutive readings, time order along rows. Data set `Thermocouples`.

3.19	3.15	3.13	3.14	3.14	3.13	3.13	3.11
3.16	3.17	3.17	3.14	3.14	3.14	3.15	3.15
3.14	3.15	3.12	3.05	3.12	3.16	3.15	3.17
3.15	3.16	3.15	3.16	3.15	3.15	3.14	3.14
3.14	3.15	3.13	3.12	3.15	3.17	3.16	3.15
3.13	3.13	3.15	3.15	3.05	3.16	3.15	3.18
3.15	3.15	3.17	3.17	3.14	3.13	3.10	3.14
3.07	3.13	3.13	3.12	3.14	3.15	3.14	3.14

```
1 > fit <- lm(temp.diff~1,data=Thermocouples)
2 > plot(residuals(fit))
3 > abline(h=0)
4 > car::durbinWatsonTest(fit)
   lag Autocorrelation D-W Statistic p-value
    1       0.2131219      1.513854   0.046
  Alternative hypothesis: rho != 0
```
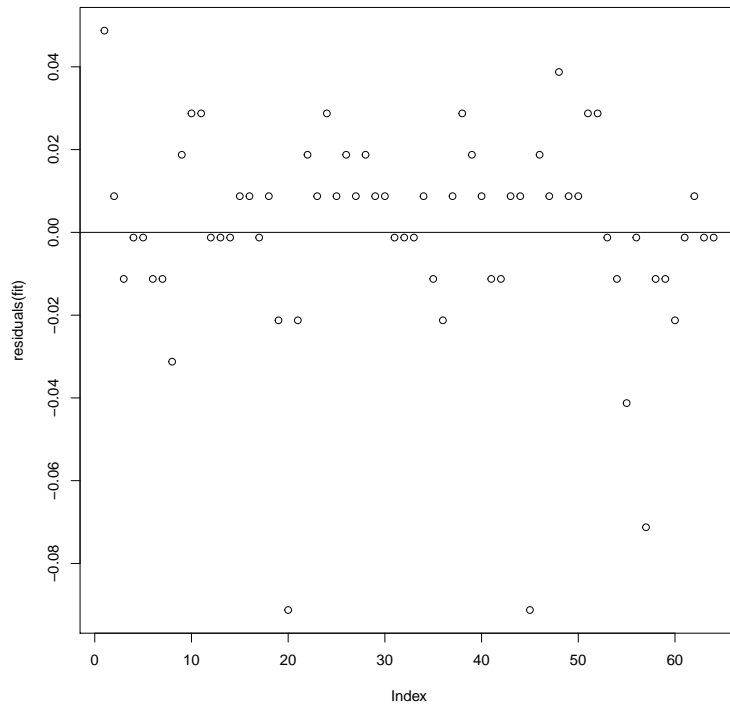
Figure 6.5: Deviations from the mean for paired differences of 64 readings from two thermocouples.

Line 1 fits a single-mean model to the differences, and line 2 plots the residuals. By default, these are plotted in data order. Assuming that the data were entered in time order, this gives a plot of residuals against time, as shown in Figure 6.5. There is a tendency for positive and negative residuals to cluster in time, indicating positive autocorrelation. Line 4 computes the Durbin-Watson statistic. It is marginally significant, but more important is the value of 1.5, indicating that the autocorrelation may be strong enough to affect our inferences.

Spatial association, another common form of dependence, arises when units are distributed in space and neighboring units have responses more similar than distant units. For example, spatial association might occur in an agronomy experiment when neighboring plots tend to have similar fertility, but distant plots could have differing fertilities.

Spatial association

One method for diagnosing spatial association is the *variogram.* We make a plot with a point for every pair of units. The plotting coordinates for a pair are the distance between the pair (horizontal axis) and the squared difference between their residuals (vertical axis). If there is a pattern in this

Variogram to detect spatial association
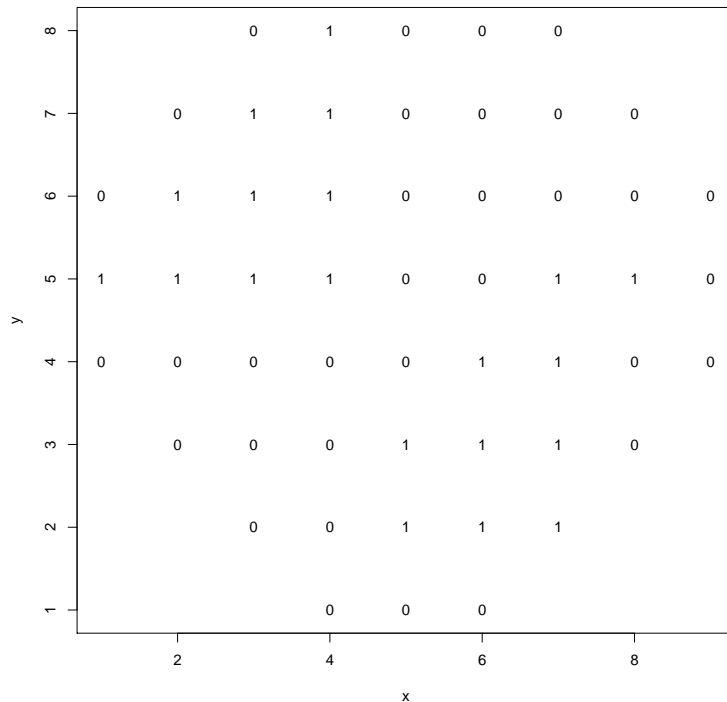
Figure 6.6: Horizontal (x) and vertical (y) locations of good (1) and bad (0) integrated circuits on a wafer. Data set `ICDefects`.

figure—for example, the points in the variogram tend to increase with increasing distance—then we have spatial association.

This plot can look pretty messy, so we usually do some averaging. Let D_{max} be the maximum distance between a pair of units. Choose some number of bins K, say 10 or 15, and then divide the distance values into K groups: those from 0 to D_{max}/K, D_{max}/K up to $2D_{max}/K$, and so on. Now plot the average of the squared difference in residuals for each group of pairs. This plot should be roughly flat for data with no spatial association; it will usually have smaller average squared differences for small distances when there is spatial association.

Plot binned averages in variogram

Example 6.5 Defective integrated circuits on a wafer

Taam and Hamada (1993) provide an example from the manufacture of integrated circuit chips. Many IC chips are made on a single silicon wafer, from which the individual ICs are cut after manufacture. Figure 6.6 (Taam and Hamada's Figure 1) shows the location of good (1) and bad (0) chips on

a single wafer. Clustering of the good chips is readily apparent.

There are 54 chips, of which 20 are good; 37% of chips are good. There are 182 adjacent chip pairs (neighbors either horizontally or vertically). If placement of good and bad chips were independent, we would expect $.37^2 + .63^2 = 53\%$ of the adjacent pairs to be the same. Instead, almost 70% are the same. Similarly, 59% of the diagonal neighbors are the same. This shows nearby chips are more likely to be the same. Curiously, and somewhat unusually, the number of equal pairs also increases at high distances. This occurs because the chips near the edge are more likely to be bad, and the only way to get a pair with a large distance is for them to cross the chip completely.

6.3 Fixing Problems

When our assessments indicate that our data do not meet our assumptions, we must either modify the data so that they do meet the assumptions or change our methods so that the data fit the assumptions of the new methods. Historically, data modification usually meant one of two things: transforming the data to a new scale (for example, analyzing the logarithm of the data) or analyzing data after outliers have been removed. New methods cover a wide array of possibilities, including broad categories such as nonparametric methods, robust methods, and generalized linear models as well as specialty methods that fix a particular problem (e.g., nonconstant variance in analysis of variance). We will see all of these approaches.

There is no step by step recipe for dealing with unmet assumptions. Instead, there is a set of potential approaches, and you might want to try more than one. By great luck, we often find that an approach that fixes one problem, say non-constant variance, will also fix another problem, say non-normality. In the following sections, we will go through different approaches that we can take and discuss how they can be used to address problems in the data.

6.3.1 Transformations

The classic tool for dealing with violations of assumptions is a transformation, or reexpression, of the response. For example, we might analyze the logarithm of the response. The idea is that the responses on the transformed scale match our assumptions more closely, so that we can use standard methods on the transformed data. Transforming the data can work remarkably well, but transforming cannot fix all problems, and it creates some problems of its own.

Transformed data may meet assumptions

There are several schemes for choosing transformations, some of which will be discussed below. For now, we note that transformations often help, and discuss the effect that transformations have on inference.

The null hypothesis tested by an F-test is that all the treatment means are equal. Together with the other assumptions we have about the responses,

the null hypothesis implies that the distributions of the responses in all the treatment groups are exactly the same. Because these distributions are the same before transformation, they will be the same after transformation, provided that we used the same transformation for all the data. Thus we may test the null hypothesis of equal treatment means on any transformation scale that makes our assumptions tenable. By the same argument, we may test pairwise comparisons null hypotheses on any transformation scale.

Transformations don't affect the null

Confidence intervals are more problematic. We construct confidence intervals for means or linear combinations of means, such as contrasts. However, the center described by a mean depends on the scale in which the mean was computed. For example, the average of a data set is not equal to the square of the average of the square roots of the data set. This implies that confidence intervals for means or contrasts of means computed on a transformed scale do not back-transform into confidence intervals for the analogous means or contrasts of means on the original scale.

Transformations affect means

> Transformations and means (or contrasts of means) do not work well together.

A confidence interval for an individual treatment *median* can be obtained by back-transforming a confidence interval for the corresponding mean from the scale where the data satisfy our assumptions. This works because medians are preserved through monotone transformations. If we truly need confidence intervals for differences of means on the original scale, then there is little choice but to do the intervals on the original scale (perhaps using some alternative procedure) and accept whatever inaccuracy results from violated assumptions. Large-sample, approximate confidence intervals on the original scale can sometimes be constructed from data on the transformed scale by using the delta method (Oehlert 1992).

Medians follow transformations

The logarithm is something of a special case. Exponentiating a confidence interval for the *difference* of two means on the log scale leads to a confidence interval for the *ratio* of the means on the original scale. We can also construct an approximate confidence interval for a mean on the original scale using data on the log scale. Land (1972) suggests the following: let $\hat{\mu}$ and $\hat{\sigma}^2$ be estimates of the mean and variance on the log scale, and let $\hat{\eta}^2 = \hat{\sigma}^2/n + \hat{\sigma}^4/[2(n+1)]$ where n is the sample size. Then form a $1 - \mathcal{E}$ confidence interval for the mean on the original scale by computing

Special rules for logs

Land's method

$$\exp(\hat{\mu} + \hat{\sigma}^2/2 \pm z_{\mathcal{E}/2}\,\hat{\eta})\ ,$$

where $z_{\mathcal{E}/2}$ is the upper $\mathcal{E}/2$ percent point of the standard normal.

Nonnormality, particularly asymmetry, can sometimes be lessened by transforming the response to a different scale. Skewness to the right is lessened by a square root, logarithm, or other transformation to a power less than one, while skewness to the left is lessened by a square, cube, or other transformation to a power greater than one. Symmetric long tails do not easily yield to a transformation.

Transformations to improve normality

Power family transformations are general tools for improving normality and equalizing variance; they can only be used for data that are positive. The power family of transformations includes

$$y \to \text{sign}(\lambda)y^\lambda \quad \text{for } \lambda \neq 0$$

and

$$y \to \log(y) \quad \text{for } \lambda = 0 \ ,$$

where $\text{sign}(\lambda)$ is $+1$ for positive λ and -1 for negative λ. The $\log$ function corresponds to λ equal to zero. We multiply by the sign of λ so that the order of the responses is preserved when λ is negative.

Nonnormality, particularly asymmetry, can sometimes be lessened by transforming the response to a different scale. Skewness to the right is lessened by a square root, logarithm, or other transformation to a power less than one, while skewness to the left is lessened by a square, cube, or other transformation to a power greater than one. Symmetric long tails do not easily yield to a transformation.

Power family transformations are not likely to have much effect unless the ratio of the largest to smallest value is bigger than 4 or so. Furthermore, power family transformations only make sense when the data are all positive. When we have data with both signs, we can add a constant to all the data to make them positive before transforming, but different constants added lead to different transformations.

Here is a simple method for finding an approximate variance-stabilizing transformation power λ. Compute the mean and standard deviation for the data in each treatment group. Regress the logarithms of the standard deviations on the logarithms of the group means; let $\hat{\beta}$ be the estimated regression slope. Then the estimated variance stabilizing power transformation is $\lambda = 1 - \hat{\beta}$. If there is no relationship between mean and standard deviation ($\hat{\beta} = 0$), then the estimated transformation is the power 1, which doesn't change the data. If the standard deviation increases proportionally to the mean ($\hat{\beta} = 1$), then the log transformation (power 0) is appropriate for variance stabilization.

The Box-Cox method for determining a transformation power is somewhat more complicated than the simple regression-based estimate, but it tends to find a better power and also yields a confidence interval for λ. Furthermore, Box-Cox can be used on more complicated designs where the simple method is difficult to adapt. Box-Cox transformations rescale the power family transformation to make the different powers easier to compare. Let $\dot{y}$ denote the geometric mean of all the responses, where the geometric mean is the product of all the responses raised to the $1/N$ power:

$$\dot{y} = \left(\prod_{i=1}^{g} \prod_{j=1}^{n_i} y_{ij} \right)^{1/N} .$$

The Box-Cox transformations are then

$$
y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\[3ex] \dot{y} \log(y) & \lambda = 0 \end{cases} \quad .
$$

In the Box-Cox technique, we transform the data using a range of λ values from, say, -2 to 3, and fit the model for each of these transformations. From these we can get $\mathrm{SS_E}(\lambda)$, the sum of squared errors as a function of the transformation power λ. The best transformation power $\lambda^\star$ is the power that minimizes $\mathrm{SS_E}(\lambda)$. We generally use a convenient transformation power λ close to $\lambda^\star$, where by convenient I mean a "pretty" power, like .5 or 0, rather than the actual minimizing power which might be something like .427.

Use best convenient power

The Box-Cox minimizing power $\lambda^\star$ will rarely be exactly 1; when should you actually use a transformation? A graphical answer is obtained by making the suggested transformation and seeing if the residual plot looks better. If there was little change in the variances or the group variances were not that different to start with, then there is little to be gained by making the transformation. A more formal answer can be obtained by computing an approximate $1 - \mathcal{E}$ confidence interval for the transformation power λ. This confidence interval consists of all powers λ such that

Confidence interval for λ

$$
\mathrm{SS_E}(\lambda) \leq \mathrm{SS_E}(\lambda^\star)\left(1 + \frac{F_{\mathcal{E},1,\nu}}{\nu}\right) \;,
$$

where ν is the degrees of freedom for error. Very crudely, if the transformation doesn't decrease the error sum of squares by a factor of at least $\nu/(\nu+4)$, then $\lambda = 1$ is in the confidence interval, and a transformation may not be needed. When I decide whether a transformation is indicated, I tend to rely mostly on a visual judgement of whether the residuals improve after transformation, and secondarily on the confidence interval.

An alternative presentation for Box-Cox transformations plots the likelihood of the model fit to the transformed data and seeks the $\lambda^\star$ that maximizes the likelihood. The likelihood interval for λ is all values of λ that produce a log-likelihood within $\chi^2_{\mathcal{E},\nu}/2$ of the maximum likelihood.

Example 6.6 Cloud seeding, continued

The cloud seeding data introduced in Example 6.2 showed considerable skewness to the right. Thus a square root or logarithm should help make things look more normal.

```
1 > fit <- lm(rainfall~seeding,data=CloudSeeding)
2 > with(CloudSeeding, plot(sort(rainfall[seeding=="nonseeded"]),
      sort(rainfall[seeding=="seeded"]),log='xy'))
3 > abline(.5,1)
4 > boxCox(fit)
5 > fit2 <- lm(log(rainfall)~seeding,data=CloudSeeding)
6 > plot(fit2,which=1)
7 > plot(fit2,which=2)
8 > anova(fit2)
 Analysis of Variance Table

 Response: log(rainfall)
          Df  Sum Sq Mean Sq F value  Pr(>F)
 seeding   1  17.007 17.0071  6.4738 0.01408 *
 Residuals 50 131.353  2.6271
9 > linear.contrast(fit2,seeding,c(-1,1))
   estimates        se  t-value    p-value lower-ci upper-ci
 1  1.143781 0.4495342 2.544369 0.01408266 0.240865 2.046697
```

Line 2 plots the sorted data for seeded and unseeded clouds, both on a logarithmic scale (see panel 1 of Figure 6.7). If the two sets of data have the same distributional shape, this line should be straight. We see it is approximately straight, so even though these two are not normally distributed, it appears that their distributions are similar.

Line 4 produces the Box-Cox plot in panel 2 of Figure 6.7. The best power is around .1, but 0 is not too far off, so we will try a log transformation. Panel 3 of Figure 6.7 shows that the variability is stable, and panel 4 shows that we might have over-transformed a bit, because the residuals are now a little skewed to the left.

Line 8 does the ANOVA for the logged data. The p-value of .014 is reasonably trustworthy, because our assumptions are reasonable on that scale. Note that the p-value on the original scale was .051, which gives us a very different sense of what is happening in the data.

We also wish to estimate the effect of seeding. On the log scale, a 95% confidence interval for the difference between seeded and unseeded is (.24, 2.05). This converts to a confidence interval on the ratio of the means of (1.27, 7.76) by back-exponentiating. A 95% confidence interval for the mean of the seeded cloud rainfalls, based on the original data and using a t-interval, is (179.1, 704.8); this interval is symmetric around the sample mean 442.0. Using Land's method for log-normal data, we get (247.2, 1612.2); this interval is not symmetric around the sample mean and reflects the asymmetry in log-normal data.

Example 6.7 Tearing Facial Tissues

Table 6.8 shows the mass added to facial tissues stretched across the mouth of a jar before the tissues tore. We would like to compare brands and compare wet versus dry.
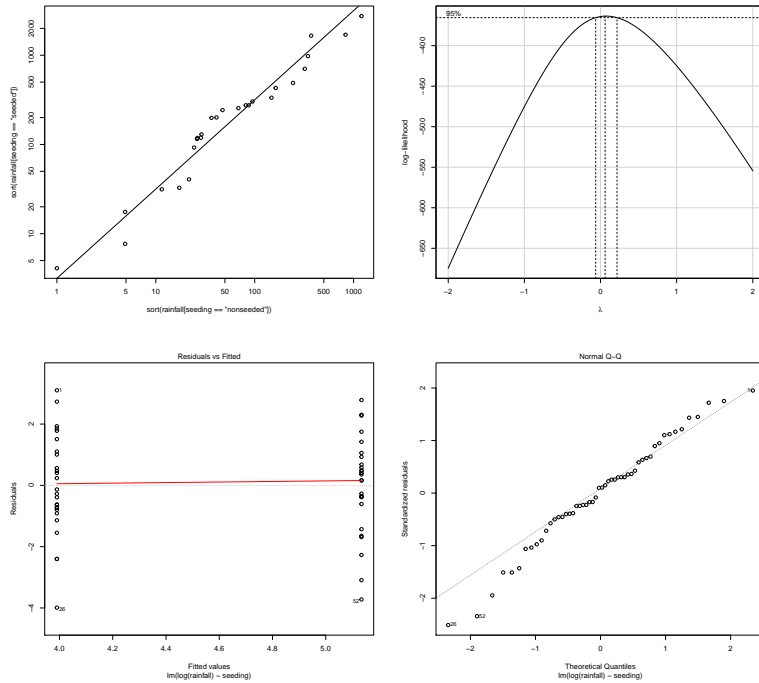
Figure 6.7: Plots for the cloud seeding data. The panels are: sorted responses for seeded clouds plotted against sorted results for unseeded clouds; Box-Cox transformation likelihood plot; residuals versus fitted plot for log-transformed data; normal probability plot for log-transformed data.

Table 6.8: Mass in grams added to facial tissues before tearing. Data from Joel Rumsch. Data set `TissueTearing`.

Brand A, dry	102	101	105	95	97	98	95	101
	100	108	91	99	102	92	98	98
	99	103	102	99	100	98	105	
Brand A, wet	70	71	72	70	71	73	74	70
	73	71	74	71	71	70	73	75
	71	69	70	74	73	69	70	71
Brand B, dry	80	81	79	83	78	77	72	80
	81	75	81	79	78	80	77	76
	76	82	77	79	80	74	75	76
Brand B, wet	45	50	49	50	48	48	49	47
	48	47	46	48	49	46	48	47
	48	48	49	46	46	49	47	49
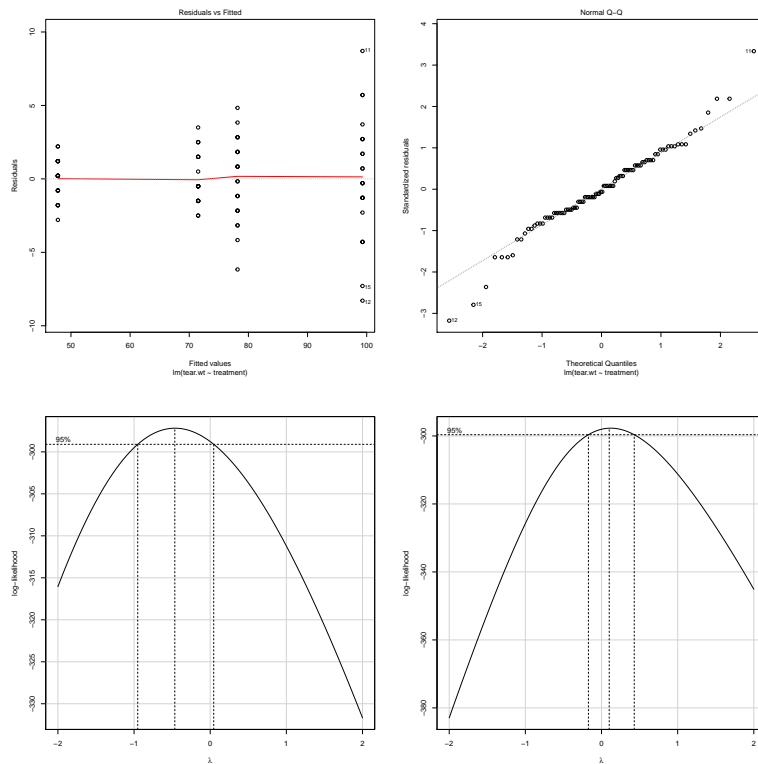
Figure 6.8: Plots for the tearing facial tissues data. The first two panels are residual and normal probability plots. The last two panels are Box-Cox likelihood profiles for the original data and the data reduced by 25.

```
1 > fit <- lm(tear.wt~treatment,data=TissueTearing)
2 > plot(fit,which=1)
3 > plot(fit,which=2)
4 > car::boxCox(fit)
5 > fit2 <- lm(tear.wt-25~treatment)
6 > car::boxCox(fit2)
```

Line 1 fits the model to the tissues data, and lines 2 and 3 make residuals and NPP plots (see Figure 6.8). There is clear evidence of non-constant variance. In addition, the non-constant variance makes the residuals look long tailed.

We use the `boxCox` function in the `car` package in line 4, with the resulting likelihood plot shown in panel 3 of Figure 6.8. This shows the power –.5 (reciprocal square root) is about best, with 0 (logarithm) just barely in the 95% confidence interval for the power, and –1 (reciprocal) just barely outside the interval.

The ratio of the maximum response to the minimum response is about

2.5. As this ratio gets smaller, powers farther and farther from 1 are needed to have much of an effect on unequal variance. Line 6 fits the model with 25 subtracted from the response, and line 7 does a Box-Cox analysis for the second model. The resulting plot is the fourth panel of Figure 6.8. In this case, power 0 (logarithm) is well inside the confidence interval.

Do not simply trust the Box-Cox curve. Refit the data with the selected transformation and see how the residuals look.

```
7 > with(TissueTearing,plot(lm(-1/sqrt(tear.wt)~treatment),which=1))
8 > with(TissueTearing,plot(lm(log(tear.wt)~treatment),which=1))
9 > with(TissueTearing,plot(lm(log(tear.wt-25)~treatment),which=1))
```

Lines 7–9 produce residual plots for the –.5 power, the logarithm, and the logarithm of the data reduced by 25. These are shown in Figure 6.9. The reciprocal square root of the data and the log of the data reduced by 25 have both stabilized the variability. However, the log of the original data, even though it is within the confidence interval, has not fully stabilized the variability.

We wanted to compare brands and wet versus dry, but now we have a problem. It is difficult to take what we learn about differences between treatments on a transformed scale and convert that back to saying something on the original scale. We can do that best if the transformation was a logarithm, so consider continuing with the log of the data less 25.

```
10 > fit3 <- lm(log(tear.wt-25)~treatment,data=TissueTearing)
11 > linear.contrast(fit3,treatment,c(1,0,-1,0))
     estimates         se t-value       p-value  lower-ci   upper-ci
   1 0.3344314 0.01485397 22.51461 3.387903e-39 0.3049301  0.3639327
12 > linear.contrast(fit3,treatment,c(0,1,0,-1))
     estimates         se t-value       p-value  lower-ci  upper-ci
   1 0.7141147 0.01485397 48.07567 5.55236e-67 0.6846135  0.743616
13 > exp(c(.305,.364,.685,.744))
   [1] 1.356625 1.439074 1.983772 2.104336
```

For dry tissues (line 11), the ratio of brand A (less 25) over brand B (less 25) is between 1.36 and 1.44; this is a remarkably narrow interval. The corresponding values for wet tissues are 1.98 and 2.10. Thus brand A is somewhat stronger for dry tissues and much stronger for wet tissues. However, if what we really want to know is the difference in strength on the original scale, then using a transformation has not really helped us answer our question.

6.3.2   Removing Outliers

Former U.S. Supreme Court Justice Potter Stewart has been famously quoted as saying that he could probably never succeed in defining hard core pornography, but "I know it when I see it." Outliers are a bit similar: they are difficult to define, but we all feel confident that we know one when we see one. Let us use the non-technical definition that an outlier is a data point that is far from the pattern established by the rest of the data. Implicit in the definition

<div style="text-align: right;">I know it when I
see it</div>
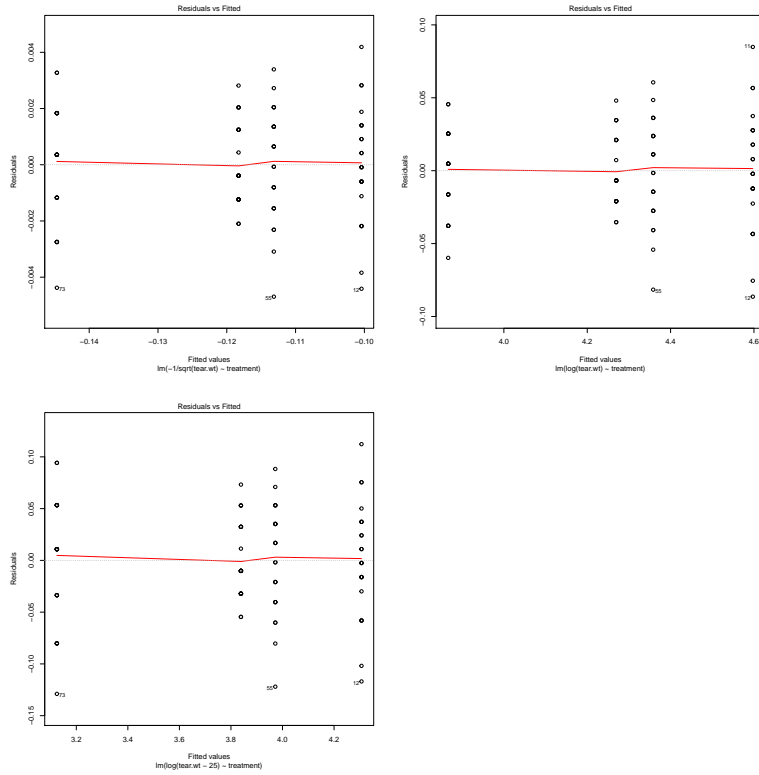
Figure 6.9: Residual plots for transformed facial tissues data. The three panels are for the reciprocal square root of the data, the logarithm of the data, and the logarithm of the data reduced by 25.

is that an outlier must be situated in a such a way that it is unlikely to be that far away by chance. Thus an outlier can only be an outlier in the context of a particular statistical model. A point that might be an outlier if we assume model A might not be an outlier if we assume model B. This applies to both the model for the variability and the model for the means.

Individual outliers can affect our analysis. One long-standing approach for handling outliers is to analyze the data with the outliers included and again with the outliers excluded. If your inferences are the same, you breathe a sigh of relief. If your conclusions change, then you must be careful in interpreting the results, because the results depend rather delicately on a few outlier data values.

Try analysis with and without outliers

Excluding outliers is very tempting, but it must be practiced with great caution and full transparency. Some outliers are truly "bad" data, and their extremity draws our attention to them. For example, we might have miscopied the data so that 17.4 becomes 71.4, an outlier; or perhaps Joe sneezed in a test tube, and the yield on that run was less than satisfactory. Outliers should be checked thoroughly to determine if there is some basis for their

Table 6.9: Width of vaporized tissue in mm for different laser power settings in watts with the laser moving at .01 inches/second. Data set `TissueVaporization`; adapted from D. Deepa.

Power			
5	10	15	19
1.11	1.71	1.92	1.12
1.10	1.41	1.88	2.34
1.21	1.55	1.96	2.41

exclusion.

Outliers can be the most important data; sometimes outliers *are* the information in the data. One of the reasons that the south polar ozone "hole" was so late in discovery was that the algorithms for the satellite ozone data filtered the sudden large drops as being erroneous readings. That is, the data showing the ozone hole were pulled out as being outliers. Routine elimination of outlier data values can be a disaster for learning from your data.

Outliers can be important data

Example 6.8 Tissue Vaporization

An 850 nm laser is being tested for use in cutting tissue. In this study, we wish to study how the power of the laser changes the width of tissue vaporization as the laser in moved at .01 inches per second. There are 12 samples of rat liver. These samples are randomized to 5, 10, 15, or 19 watts (power of the laser), with three samples per power level. After cutting, the average width of vaporized tissue along the incision is taken as the response; see Table 6.9.

```
1 > fit <- lm(width~power.factor-1,data=TissueVaporization)
2 > summary(fit)

  Coefficients:
                Estimate Std. Error t value Pr(>|t|)
  power.factor5    1.1400     0.2149   5.305 0.000724 ***
  power.factor10   1.5567     0.2149   7.244 8.85e-05 ***
  power.factor15   1.9200     0.2149   8.935 1.95e-05 ***
  power.factor19   1.9567     0.2149   9.106 1.70e-05 ***
  ---
  Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

  Residual standard error: 0.3722 on 8 degrees of freedom
  Multiple R-squared:  0.9682,Adjusted R-squared:  0.9523
  F-statistic: 60.85 on 4 and 8 DF,  p-value: 4.998e-06

3 > plot(fit,which=1)
4 > plot(fit,which=2)
5 > watts <- rep(c(5,10,15,19),each=3)
6 > with(TissueVaporization,plot(power,width))
```

Lines 3 and 4 fit and summarize the standard model, and lines 5 and 6 generate residual plots shown in Figure 6.10. There is a pretty clear outlier, and notice that the one low value makes the other values at the same power level

have large positive residuals. The outlier is, in this case, clearly visible when we plot the responses against treatment (third panel of Figure 6.10).

```
7 > use <- (1:12)!=10
8 > fit2 <- lm(width~power.factor-1,data=TissueVaporization,subset=use)
9 > summary(fit2)

                Estimate Std. Error t value Pr(>|t|)
power.factor5    1.14000    0.05261   21.67 1.12e-07 ***
power.factor10   1.55667    0.05261   29.59 1.30e-08 ***
power.factor15   1.92000    0.05261   36.50 3.01e-09 ***
power.factor19   2.37500    0.06443   36.86 2.81e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.09112 on 7 degrees of freedom
Multiple R-squared:  0.9983,Adjusted R-squared:  0.9973
F-statistic:  1009 on 4 and 7 DF,  p-value: 9.705e-10
```

Line 9 generates a variable to indicate data to use (TRUE) and data to omit (FALSE, case 10); line 10 then refits with the outlier left out. Examination of the summary of the fit (line 11) shows that the first three estimated means are unchanged. Showing this is the real reason for using the `vw-1` form of predictor. We also see that the fourth mean is dramatically higher.

In addition to means changing, the standard errors of estimated effects have decreased by a factor of 4. Differences between treatments that were insignificant now look significant. Inference depends strongly on this single value.

One is tempted to speculate that the value of 1.12 might be a transcription error; perhaps it should have been 2.12? Fortunately, in this experiment the widths are determined from digital photographs, and those can be reinspected to determine if this is simply a transcription error.

6.3.3   Robust methods

In many cases, outliers are just data like any other values in your data; the problem is not outliers, the problem is that the model you assumed is not a good fit to the data you have. *Robust* methods are designed for situations where the usual assumptions of independent residuals with constant variability are true, but the data come from a symmetric distribution with longer tails than a normal distribution. That is, robust methods are designed for situations where the data mostly meet our usual assumptions, but are instead more outlier prone.

For example, `bglmm` has an option to assume that the data follow a Student's *t*-distribution with 4 degrees of freedom. This will accommodate longer tailed data than the default normal distribution. Alternatively, the `MASS` package has two functions that are useful in this context. `MASS::rlm` is robust and fairly efficient for a broad range of long-tailed distributions. `MASS::lqs` can accommodate a large number of outliers without being
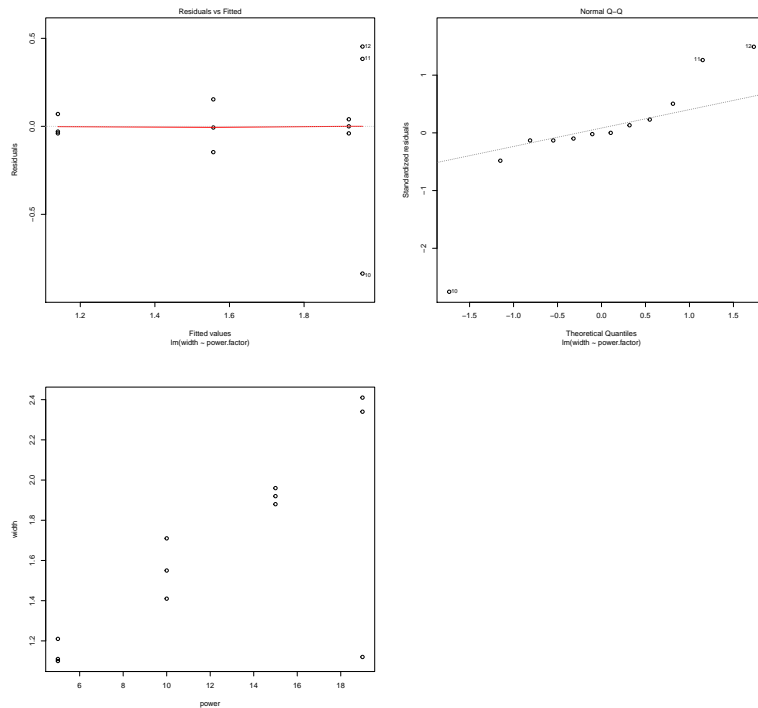
Figure 6.10: Residuals versus predicted plot, residual probability plot, and response versus predictor for liver vaporization width data.

overly perturbed by the outliers.

Example 6.9 Tissue Vaporization, continued

We saw in Example 6.8 that the data contained a large outlier. Let's try refitting this model to the full data set using a robust method.
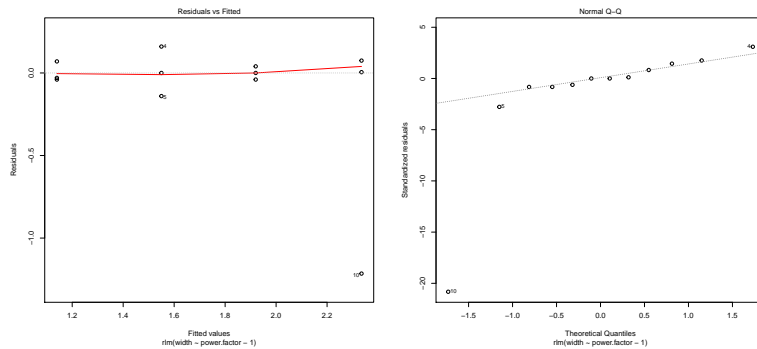
Figure 6.11: Residuals versus predicted plot and residual probability plot for a robust fit of the liver vaporization width data.

```
1 > fit3 <- MASS::rlm(width~power.factor-1,data=TissueVaporization)
2 > summary(fit3)

Call: rlm(formula = width ~ power.factor - 1, data = TissueVaporization)
Residuals:
       Min          1Q      Median          3Q         Max
-1.215e+00 -4.000e-02 -2.639e-05  4.750e-02  1.599e-01

Coefficients:
                Value   Std. Error t value
power.factor5   1.1400  0.0573     19.8788
power.factor10  1.5501  0.0573     27.0291
power.factor15  1.9200  0.0573     33.4800
power.factor19  2.3351  0.0573     40.7187

Residual standard error: 0.0593 on 8 degrees of freedom
3 > plot(fit3,which=1)
4 > plot(fit3,which=2)
5 > anova(MASS::rlm(width~power.factor,data=TissueVaporization))
Analysis of Variance Table

Response: width
             Df Sum Sq Mean Sq F value Pr(>F)
power.factor  3 1.9657 0.65523
Residuals       1.5379
```

Line 1 uses `rlm` from the `MASS` package. The results on line 2 are very close to what we obtained when we omitted the outlier, except that `rlm` figures out how to up or down weight data values on its own. The residual plots from lines 3 and 4 are shown in Figure 6.11. We see that case 10 looks even more outlying than before, but the other two responses at the highest power setting no longer look as unusual as they did before.

Robust methods are good for fitting, but inference for robust methods is not as well developed as what we have seen for other methods. For example, line 5 shows the results for an ANOVA of the model. There are no F-values or p-values, so it leaves one feeling hungry for more.

6.3.4 Nonparametric methods

We have already talked about a randomization test as an alternative to a t-tests; randomization, sometimes called permutation, tests are a class of nonparametric methods. They make no distributional assumptions and derive a reference distribution for a test from the distribution generated by repetitively re-randomizing the data to treatments and observing the test statistic. Given a computer and software, this is relatively straightforward. The `perm` package in **R** contains several randomization test analogues, include `permTS` for two-sample tests and `permKS` for K-sample tests (an alternative to ANOVA).

Rank tests are a second group of nonparametric statistics that can be used instead of tests based on specific distributional assumptions. For example, one can use the Mann-Whitney-Wilcoxon test instead of a t-test (also called the Mann-Whitney U-test or the Wilcoxon rank-sum test—`wilcox.test` in **R**), and the Kruskal-Wallis test instead of ANOVA (`kruskal.test` in **R**). The assumptions for these tests are that the data are independent coming from distributions that all have the same shape and spread, but the medians in different groups might not be the same. Rank tests work with the ranks of the data; the smallest data value gets assigned a rank of 1 up to the largest data value getting a rank of N. Ranks are unaffected by any monotone transformation of the data, so rank tests are unaffected by any monotone transformation of the data.

If your data meet a parameteric assumption (for example, normally distributed), doing inference under that parametric assumption is generally more efficient than robust or nonparametric inference. That is, confidence intervals will be shorter and power will be higher for the parametric inference. But if your data do not meet an obvious parametric assumption, rank-based methods are alternatives.

Rank tests

Example 6.10 Cloud seeding, continued

Since the cloud seeding data arose from a randomized experiment, we could use a randomization test on the difference of the means of the seeded and unseeded cloud rainfalls.

```
1 > perm::permTS(rainfall~seeding,data=CloudSeeding)

 Permutation Test using Asymptotic Approximation

data:  rainfall by treat
Z = -1.9421, p-value = 0.05213
alternative hypothesis:
      true mean seeding=nonseeded - mean seeding=seeded is not equal to 0
sample estimates:
mean seeding=nonseeded - mean seeding=seeded
                                -277.3962

2 > perm::permTS(log(rainfall)~seeding,data=CloudSeeding)

    Permutation Test using Asymptotic Approximation

data:  log(rainfall) by treat
Z = -2.4179, p-value = 0.01561
alternative hypothesis:
      true mean seeding=nonseeded - mean seeding=seeded is not equal to 0
sample estimates:
mean seeding=nonseeded - mean seeding=seeded
                                -1.143781
```

Lines 1 and 2 show the two-sample randomization t-tests for the cloud seeding data on the original and logarithmic scales. Note that the p-value depends on the scale; randomization tests are not invariant under monotone transformations. Note also that the p-values for the randomization tests are very close to those we obtained assuming normality.

```
3 > wilcox.test(rainfall~seeding,data=CloudSeeding)

    Wilcoxon rank sum test with continuity correction

data:  rainfall by treat
W = 203, p-value = 0.01383
alternative hypothesis: true location shift is not equal to 0
```

Line 3 shows the results of the Wilcoxon test. Again, its p-value in this case is close to that of the t-test on the log scale, where the assumptions of the t-test are met. Although not shown, if you repeat the Wilcoxon test with the data after any monotone transformation, you will get the same results.

6.3.5 Methods for nonconstant variance

Dealing with nonconstant variance has provided gainful employment to statisticians for many years, so there are a number of alternative methods to consider. The simplest situation may be when the ratio of the variances in the different groups is known. For example, consider a situation where the response for each experimental unit is the average of the responses for its measurement units. Assume also that error variability is primarily due to variability among the measurement units and that variability is constant. If each experimental unit in treatments 1 and 2 had five measurement units, and for each unit in

Weighted ANOVA when ratio of variances is known

treatments 3 and 4 had seven measurement units, then the variance between experimental units in treatments 3 and 4 would be 5/7 the size of the variance between experimental units in treatments 1 and 2, simply due to different numbers of values in each average. Situations such as this can be handled using *weighted ANOVA*, where each unit receives a weight proportional to the number of measurement units used in its average. In **R**, there is an optional `weights` argument to `lm` that enables weighting.

For pairwise comparisons, the Welch procedure is quite attractive. This procedure is sometimes called the "unpooled" t-test. Let s_i^2 denote the sample variance in treatment i. Then the Welch test statistic for testing $\mu_i = \mu_j$ is

$$t_{ij} = \frac{\overline{y}_{i\bullet} - \overline{y}_{j\bullet}}{\sqrt{s_i^2/n_i + s_j^2/n_j}}.$$

Welch's t for pairwise comparisons with unequal variance

This test statistic is compared to a Student's t distribution with

$$\nu = \left(s_i^2/n_i + s_j^2/n_j\right)^2 \left/ \left(\frac{1}{n_i - 1}\frac{s_i^4}{n_i^2} + \frac{1}{n_j - 1}\frac{s_j^4}{n_j^2}\right)\right.$$

degrees of freedom. For a confidence interval, we compute

$$t_{ij} = \overline{y}_{i\bullet} - \overline{y}_{j\bullet} \pm t_{\mathcal{E}/2,\nu}\sqrt{s_i^2/n_i + s_j^2/n_j}\ ,$$

with ν computed in the same way. More generally, for a contrast we use

$$t = \frac{\sum_i^g w_i \overline{y}_{i\bullet}}{\sqrt{\sum_i^g w_i^2 s_i^2/n_i}}$$

with approximate degrees of freedom

$$\nu = \left(\sum_{i=1}^g w_i^2 s_i^2/n_i\right)^2 \left/ \left(\sum_{i=1}^g \frac{1}{n_i - 1}\frac{w_i^4 s_i^4}{n_i^2}\right)\right..$$

Confidence intervals are computed in an analogous way.

The Welch procedure generally gives observed error rates close to the nominal error rates. Furthermore, the accuracy improves quickly as the sample sizes increase, something that cannot be said for the t and F-tests under nonconstant variance. Better still, there is almost no loss in power for using the Welch procedure, even when the variances are equal. For simple comparisons, the Welch procedure can be used routinely. The problem arises in generalizing it to more complicated situations.

Welch's t works well

In **R**, the `t.test` function (which does both tests and confidence intervals) has an argument `var.equal` that defaults to FALSE. That is, **R** defaults to the unpooled Welch procedure; you must actively select the pooled error version.

The next most complicated procedure is an ANOVA alternative for non-constant variance, and here there are two primary contenders. First, the Welch method for the two-sample t-test has been extended to testing that g treatments all have the same mean. Using the notation

$$w_i = \frac{n_i}{s_i^2}$$

$$u = \sum_{i=1}^{g} w_i$$

$$\tilde{x}_{\bullet\bullet} = \frac{\sum_{i=1}^{g} w_i \overline{x}_{i\bullet}}{u}$$

$$f = \frac{1}{(3/(g^2 - 1)) \sum_{i=1}^{g} [(1 - w_i/u)^2/(n_i - 1)]}$$

the Welch alternative to the F-test is

$$W = \frac{\sum_{i=1}^{g} w_i (\overline{x}_{i\bullet} - \tilde{x}_{\bullet\bullet})^2/(g-1)}{1 + \frac{2(g-2)}{(g^2-1)} \sum_{i=1}^{g} (1 - w_i/u)^2/(n_i - 1)}$$

which should be approximated under the null hypothesis as an F with $g - 1$ and f degrees of freedom.

The Brown-Forsythe method is the second procedure that is less sensitive to nonconstant variance than is the usual ANOVA F-test. Again let s_i^2 denote the sample variance in treatment i, and let $d_i = s_i^2(1 - n_i/N)$. The Brown-Forsythe modified F-test is

Brown-Forsythe
modified F

$$BF = \frac{\sum_{i=1}^{g} n_i (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2}{\sum_{i=1}^{g} s_i^2(1 - n_i/N)} \quad .$$

Under the null hypothesis of equal treatment means, BF is approximately distributed as F with $g - 1$ and ν degrees of freedom, where

$$\nu = \frac{\left(\sum_i d_i\right)^2}{\sum_i d_i^2/(n_i - 1)} \quad .$$

In **R**, the `oneway.test` function implements the Welch modified F-test, and `brown.forsythe.test` in package `cfcdae` implements Brown-Forsythe. Which should you use? Both reduce to the Welch t-test when $g = 2$. Both do a good job of controlling $\mathcal{E}$ when the variances are unequal. The Welch method is preferable when "extreme" means go with small variances, and the Brown-Forsythe method is preferable when the "extreme" means go with large variances, where preferable means more likely to detect the unequal means. See Brown and Forsythe (1974). If you were forced to choose just one, go with the Welch procedure.
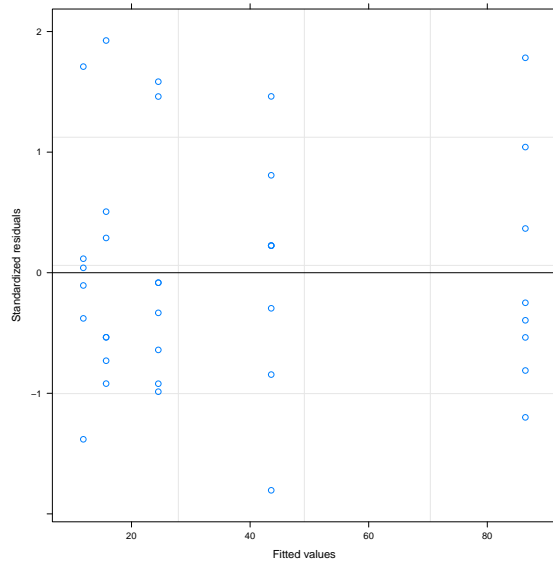
Figure 6.12: Standardized residual plot for resin lifetime data on original scale with separate standard deviations by treatment.

It is also possible to model nonconstant variances. This approach is overkill in the g groups, separate means models we have looked at, because we have other workable procedures such as the Welch or the Brown-Forsythe. However, modeling nonconstant variance will generalize to much more complex situations and models, whereas there is no handy one-size-fits-all generalization of the Welch or Brown-Forsythe.

In **R**, the gls function (*generalized least squares*) has a weights argument that allows us to specify a wide range of models for the error variances. For example, we can specify a different error variance along some grouping factor, or we can specify that the error variance is proportional to a covariate, or a power of a covariate or even a power of the fitted values. More complex forms are also available, and you can create new variance forms as well.

Modeling nonconstant variance sounds like every thing we need, but it is not a complete solution to our problem. It works well for estimation, and it works well for inference so long as sample sizes are large enough for the likelihood ratio test to work well or for normal approximations to be sufficiently accurate. However, we do not always have large sample sizes. This means, for example, that we will be using normal-based confidence intervals instead of t-based confidence intervals, because we do not have an equivalent degrees of freedom.

Example 6.11 Resin lifetimes, continued

Suppose that we need to do inference on the resin lifetime in time units, not in log time units. For example, we might want a confidence interval on the difference in time for the first two temperature levels.

```
1 > fit <- lm(Time~temp.factor,data=ResinLifetimes)
2 > anova(fit)
Analysis of Variance Table

Response: Time
            Df  Sum Sq Mean Sq F value    Pr(>F)
temp.factor  4 28027.9  7007.0  101.81 < 2.2e-16 ***
Residuals   32  2202.4    68.8
3 > brown.forsythe.test(Time~temp.factor,data=ResinLifetimes)

 One-way analysis of means, Brown-Forsythe (unequal variances)

data:  Time and temp.factor
F = 111.74, num df = 4.000, denom df = 18.298, p-value = 1.36e-12
4 > oneway.test(Time~temp.factor,data=ResinLifetimes)

One-way analysis of means (not assuming equal variances)

data:  Time and temp.factor
F = 68.943, num df = 4.000, denom df = 14.373, p-value = 3.273e-09
```

Lines 1 and 2 fit the ordinary linear model for Time in original units and show the ANOVA; we know this ANOVA is not trustworthy. Lines 3 and 4 do the Brown-Forsythe and Welch F-tests. Both of these are highly significant, and in this case all three give the same result.

```
5 > fitw <- gls(Time~temp.factor-1,data=ResinLifetimes,
        weights=varIdent(form=~1|temp.factor))
6 > summary(fitw)
  Generalized least squares fit by REML
    Model: Time ~ temp.factor - 1
    Data: NULL
        AIC      BIC    logLik
    234.0233 248.6806 -107.0116

  Variance function:
   Structure: Different standard deviations per stratum
   Formula: ~1 | temp.factor
   Parameter estimates:
        175        194        213        231        250
   1.0000000 0.7339797 0.4918991 0.1329593 0.2839912

  Coefficients:
                  Value Std.Error   t-value p-value
  temp.factor175 86.42419  4.606382 18.761838       0
  temp.factor194 43.55983  3.380991 12.883747       0
  temp.factor213 24.51965  2.265875 10.821274       0
  temp.factor231 15.71865  0.654748 24.007153       0
  temp.factor250 11.87267  1.510547  7.859846       0
  ...
  Residual standard error: 13.02881
  Degrees of freedom: 37 total; 32 residual
7 > plot(fitw)
8 > 13.029*c(1,.734,.492,.133,.284)
  [1] 13.029000  9.563286  6.410268  1.732857  3.700236
```

Now we use the `gls` function to fit a model with separate means, but we also allow separate error variances for each level of the temperature factor. Line 5 shows the somewhat arcane syntax needed to get the different variances. The variance functions in **R** are extremely powerful and can do quite complex structures, but the side effect is that even simple variance structures are not straightforward.

The summary output on line 6 has several sections. The variance function section says that the error standard deviations for the groups are all relative to group one; the group two standard deviation is 73% of that in group one, group three is 49% that of group one, and so on. The residual root mean square near the bottom (13.029) is for group 1, and the corresponding values for other groups are found proportionally as shown on line 8. The residual plot (line 7, Figure 6.12) looks like constant variance, because the plot shows standardized residuals, and each treatment group has its own individually fit standard deviation.

Looking at the coefficients section of the summary output, one can see that the estimated means are the same as we obtain from a linear model ignoring the nonconstant variance, but the standard errors for the different factor levels are dramatically different. This is due to the different standard deviations fit in each factor level (and somewhat as well due to the sample sizes not being equal).

```
9 > with(ResinLifetimes, t.test(Time[temp.factor==175],Time[temp.factor==194]))

    Welch Two Sample t-test

 data:  Time[temp.factor == 175] and Time[temp.factor == 194]
 t = 7.5016, df = 12.846, p-value = 4.822e-06
 alternative hypothesis: true difference in means is not equal to 0
 95 percent confidence interval:
  30.50490 55.22381
 sample estimates:
 mean of x mean of y
  86.42419  43.55983
0 > vcov(fitw)[1:2,1:2]
                 temp.factor175 temp.factor194
 temp.factor175        21.21875         0.0000
 temp.factor194         0.00000        11.4311
1 > sqrt(21.22+11.43)
 [1] 5.714018
2 > (86.42-43.56)/5.714
 [1] 7.500875
```

Line 9 does the Welch two-sample t procedure for comparing groups one and two on the original time scale. The means and difference are what we would get from the summary on line 6. Line 10 extracts the variances of the coefficients, line 11 computes the standard error of the difference of the means of the first two groups, and line 12 computes a z-test as the difference divided by the standard error. The value from line 12 is the same as the Welch test (give or take round off error). The difference is that for the Welch test we have an estimate of the degrees of freedom to use, but for the results from `gls` we are just assuming the test follows a standard normal.

```
13 > fitw2 <- gls(Time~temp.factor-1,data=ResinLifetimes,
       weights=varPower(form=~fitted(.)))
14 > summary(fitw2)
   Generalized least squares fit by REML
     Model: Time ~ temp.factor - 1
     Data: NULL
          AIC      BIC    logLik
     233.9517 244.2118 -109.9758

   Variance function:
    Structure: Power of variance covariate
    Formula: ~fitted(.)
    Parameter estimates:
       power
   0.7716835

   Coefficients:
                    Value Std.Error   t-value p-value
   temp.factor175 86.42419  5.137080 16.823600       0
   temp.factor194 43.55983  3.027629 14.387442       0
   temp.factor213 24.51965  1.943176 12.618340       0
   temp.factor231 15.71865  1.473997 10.663960       0
   temp.factor250 11.87267  1.282116  9.260211       0
   ...
   Residual standard error: 0.4653629
   Degrees of freedom: 37 total; 32 residual
15 > .465*c(86.42,43.56,24.52,15.72,11.88)^.772
   [1] 14.538485  8.567000  5.497469  3.900446  3.142037
16 > AIC(fit)
   [1] 268.1971
17 > AIC(fitw)
   [1] 234.0233
18 > AIC(fitw2)
   [1] 233.9517
```

We mentioned that there are many different potential variance functions, and line 13 shows an example using a power function. In this case, we use the model that the error standard deviation is proportional to a power of the fitted value. (Note, if the power is 1, a log transformation would exactly equalize the variances.) The summary output in line 14 gives the estimated power (.772) along with the proportionality constant (.465). Line 15 shows the estimated standard deviations. These are similar to, but not the same as, the estimated standard deviations from line 8.

Lines 16–18 show AIC values for the constant variance, group variance, and power variance model fits. Warning: the AIC value for the linear model (`fit`) is *not* comparable to the AIC values for `fitw` and `fitw2`. Look back through the output and you will see that `gls` fits using REML instead of maximum likelihood. We will talk much more about REML later when we estimate variance components. For now, remember that you cannot compare a REML likelihood or AIC to a non-REML likelihood or AIC, and you cannot compare likelihood or AIC from two REML models if they have different mean structures. That significantly restricts what we can compare, but we can compare `fitw` and `fitw2`, because they have the same mean structure and both used REML. Here, `fitw2` is preferred, because even though it has a

slightly lower likelihood, it used fewer parameters than `fitw` to get there. (It is possible to refit the standard model using REML for comparability; its AIC is 248.2, much higher than the two weighted models we considered.)

6.3.6   Generalized Linear Models

We have been focused on data that follow a normal distribution, but there is a whole world of other distributions out there, and data could come from any of them. *Generalized linear models* (GLM) allow us to fit a broader range of response distributions with mean structures similar to those we have used for normal distributions. (Note: it is easy to get confused by nomenclature; generalized linear models are not the same as the general linear model or generalized least squares.) Our standard linear model is just one example of a GLM. You should also be aware that GLMs are a broad subject, and we will barely scratch the surface. McCullagh and Nelder (1989) is the fundamental reference, but there are dozens of others at various positions along the applied to theoretical spectrum.

With our standard linear model, we have data distributed as normal with mean μ, and we describe the structure of the mean with treatment effect terms, regression terms, and so on. This is the *linear predictor*. In the GLM, we have data from a distribution that is not necessarily normal. The mean of each data point is transformed by a *link function*, and then the transformed version of the mean is fit via a linear predictor.

Linear predictor; link function

Note that this is a parameterization of a transformed mean, not the mean of transformed data. This implies that we avoid some of the difficulties we have encountered working with transformed data but trying to make inferences on the original scale. The transformations used for means also take what might be a restricted space (for example, between 0 and 1 for the probability of a binomial or nonnegative for the mean of a Poisson) and map it to an unrestricted space. This avoids problems where a direct modeling of the mean (called an identity link) might might lead to a fitted mean outside the permissible range of values.

Parameterize transformed mean

In **R**, the distribution is indicated by the `family`. The possible choices for the family in **R** are `gaussian` (normal), `inverse.gaussian` (a positively skewed distribution for positive data; not the reciprocal of a normal), `Gamma` (the gamma distribution, but with a capital G to distinguish it from the gamma function; the gamma distribution is also for positive data skewed to the right), `binomial`, and `poisson`. These families cover a much broader range than the normal alone and include distributions with skewness and distributions with variances that depend on the mean. GLMs give us a better chance of working with a distribution that conforms to the properties of the data rather than trying to make our data look like a normal via transformation.

Non-normal distributions

Some distributions in GLM are determined entirely by their mean; the binomial and Poisson are of this form. Other distributions have an additional parameter, called a *dispersion* parameter and usually denoted by ϕ. This parameter adjusts the variability. For example, the variance is the dispersion

Dispersion parameter

Table 6.10: Germination counts (out of 20) for garden pea seeds under four treatments. Data set `PeaGermination`; adapted from K. Becklund.

Hours	Substrate	Counts	
12	Towel	3	2
12	Soil	12	15
24	Towel	17	18
24	Soil	15	13

parameter for a normal distribution, and the reciprocal shape of a gamma distribution is its dispersion parameter. For distributions that do not depend on the dispersion parameter, ϕ is set to 1.

We use maximum likelihood to fit GLMs, and we use the *deviance* (sometimes called *residual deviance*) to assess how closely the model fits. For families where ϕ is known to be 1, the deviance for a model is twice the difference in log likelihood between a model of interest and the saturated model (recall that the saturated model has a rich enough structure to compute a separate mean for each response). For models where ϕ must be fit, the deviance is is twice the difference in log likelihood between a model of interest and the saturated model multiplied by ϕ. We estimate ϕ as the residual deviance divided by the residual degrees of freedom. For a GLM with a normal distribution, the deviance is the sum of squared errors.

Deviance

One view of ANOVA for a sequence of nested models is successive reductions in sum of squared errors together with successive increases in parameters used. There is an *analysis of deviance* that follows the same pattern. Each time we add a term to a model, we use additional parameters and decrease the deviance until we are left with a final residual deviance and residual degrees of freedom.

Analysis of deviance

In models where $\phi = 1$ (binomial and Poisson), terms can be tested by treating the incremental deviance for the term as being distributed under the null as chi-square with degrees of freedom equal to the number of incremental parameters for the term. In fact, this is the likelihood ratio test. When ϕ must be estimated, form a test statistic for a term by dividing its incremental deviance by its incremental degrees of freedom, and then dividing that ratio by our estimate of ϕ. Treat that ratio as having an F distribution under the null, with numerator and denominator degrees of freedom taken as the incremental degrees of freedom for the term and the residual degrees of freedom used to estimate ϕ. For a GLM based on a normal distribution, this equals the usual F-test in ANOVA.

Comparing GLMs

Example 6.12 Germination of garden peas

A middle school science project explores the effect of soaking and substrate treatments on the germination of garden pea seeds. There are eight

batches of 20 seeds. Each batch of seeds is assigned to one of four treatment conditions, two batches per condition. The treatments are presoak for 12 hours and place on wet paper towel; presoak for 12 hours and place on wet potting soil; presoak for 24 hours and place on wet paper towel; and presoak for 24 hours and place on wet potting soil.

The batches will be observed after seven days, with the response being the number of seeds germinating. The batches are the experimental units, the seeds are the measurement units, with a 0/1 response for each measurement unit, and the total as the response for the experimental unit. Table 6.10 gives the results.

These data can be reasonably modeled as a binomial response, with 20 trials and success probability potentially depending on treatment. In **R**, one uses the `glm` function to fit a GLM model, using the `family` argument to indicate the binomial distribution. For a binomial, the response is a matrix with two columns giving the successes and the failures.

```
1 > fit <- glm(cbind(germinated,nongerminated)~treatment,data=PeaGermination,
        family=binomial)
2 > summary(fit)

Call:
glm(formula = cbind(germinated, nongerminated) ~ treatment, family = binomial)

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.3292  -0.7039  -0.3292   0.4966  -0.3487   0.7329   0.3487  -0.4809

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3945     0.2077   1.900   0.0575 .
treatment1    0.3363     0.3164   1.063   0.2878
treatment2    0.4528     0.3204   1.413   0.1576
treatment3   -2.3405     0.3968  -5.899 3.66e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58.5202  on 7  degrees of freedom
Residual deviance:  1.9703  on 4  degrees of freedom
AIC: 33.856

Number of Fisher Scoring iterations: 4
```

Line 1 fits the model. The summary on line two gives the usual parameter estimates and standard errors, plus some additional quantities of interest. Recall that the linear predictor for a binomial in on the logit scale. For example, the linear predictor for treatment 1 is $.3945 - 2.3405 = -1.946$. Transforming back to the probability scale, the estimated success probability is $\exp(-1.946)/(1+\exp(-1.946)) = .125$. This is, of course 1/8 corresponding to 5 of 40 seeds germinating.

The summary also shows *deviance residuals*. If you sum the squares of the deviance residuals you get the deviance for the full model, which is

1.97 for these data. The null deviance (for the single-mean model) is given as 58.52, so the decrease in deviance going to the separate-means model is 56.55; this is accomplished with 3 additional parameters. We can get this more simply as shown in line 3.

```
3 > anova(fit)
  Analysis of Deviance Table

  Model: binomial, link: logit

  Response: cbind(germinated, 20 - germinated)

  Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev
  NULL                        7       58.52
  treatment  3    56.55       4        1.97
4 > pchisq(56.55,3,lower.tail=FALSE)
  [1] 3.206016e-12
```

Line 4 computes the p-value, which is very small indicating strong evidence against the single-mean model.

Example 6.13 Resin lifetimes, continued

The resin lifetime data on the original scale have non-constant variance with the variance roughly proportional to the square of the mean. The GLM with a gamma distribution also has the variance proportional to the square of the mean, so it might be a reasonable alternative to transforming the data.

```
1 > fit <- glm(Time~temp.factor,data=ResinLifetimes,
         family=Gamma(link=log))
2 > summary(fit)

Call:
glm(formula = Time ~ temp.factor, family = Gamma(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.51486  -0.10500  -0.03331   0.05429   0.45970

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.33239    0.03622  92.005  < 2e-16 ***
temp.factor1  1.12687    0.07004  16.090  < 2e-16 ***
temp.factor2  0.44174    0.07004   6.307 4.47e-07 ***
temp.factor3 -0.13292    0.07004  -1.898   0.0668 .
temp.factor4 -0.57755    0.07361  -7.846 5.96e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Gamma family taken to be 0.04791153)

    Null deviance: 20.1272  on 36  degrees of freedom
Residual deviance:  1.5385  on 32  degrees of freedom
AIC: 249.24

Number of Fisher Scoring iterations: 4

3 > (20.127-1.538)/4/.0479
[1] 97.01983
```

Line 1 fits the gamma GLM with a log link for the mean. Line 3 computes the F-statistic for comparing the separate-means model to the single-mean model based on information from the summary in line 2. This F is very similar to the F of 96.36 (with the same degrees of freedom) that we computed for the data on the log scale.

One can apply most of the same post-fit functions we used for lm models to glm models; these include linear.contrast, plot, and so on. Be aware that linear.contrast works on the scale of the linear predictor, which might be log, or logit, or something else.

Sometimes count data look roughly like a Poisson distribution, but instead of having a variance that is equal to the mean as a Poisson would, the variance is greater than the mean. Here are two options for data like that. The function glm.nb in the MASS package will fit a GLM to data following a negative binomial distribution. The negative binomial has a variance $(\mu + \mu^2/r)$ that is larger than its mean (μ); for a negative binomial, the ratio of the variance to the mean increases as the mean increases. The second option is the quasipoisson family for glm. In a quasi-Poisson, the variance is proportional to the mean, but the proportionality constant (the dispersion parameter) can be greater than one.

The quasi-Poisson estimates will be the same as the Poisson estimates, but their standard errors will be multiplied by the square root of the estimated

dispersion parameter. Technically, there is no likelihood for a quasi-Poisson model (only a quasi-likelihood), so **R** does not return a likelihood, AIC, or other likelihood-based quantities. However, deviance based tests described above provide a reasonable approach to model comparison.

Example 6.14 Copepoda

Exercise 2.3 introduced counts of Copepoda in artificial wetland microcosms post snowmelt with either neutral or lowered pH. Given that these are counts of items, it might make sense to model the counts as Poisson distributed.

```
1 > fitp <- glm(count~pH,data=Copepoda,family=poisson())
2 > summary(fitp)

  Deviance Residuals:
      1       2       3       4       5       6
  4.698  -2.173  -2.952  -8.436  -1.605   8.067

  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  5.09496    0.03212 158.638  < 2e-16 ***
  pH1         -0.14148    0.03212  -4.405 1.06e-05 ***

  (Dispersion parameter for poisson family taken to be 1)

      Null deviance: 193.93  on 5  degrees of freedom
  Residual deviance: 174.32  on 4  degrees of freedom
  AIC: 219.29
```

Line 1 fits the Poisson GLM to the data, and line 2 provides a summary. The deviance decreased by 19.6 when we added the single additional parameter, so that looks highly significant. Similarly, the t-test for the pH coefficient is quite large. However, for a Poisson GLM we would expect the residual deviance to be roughly the same size as the residual degrees of freedom, and that is certainly not the case here.

The large relative size of the residual deviance suggests that the data have more variability than Poisson with the same mean should have. To account for this, we fit a quasi-Poisson model in line 3 and show the summary in line 4.

```
3 > fitqp <- glm(count~pH,data=Copepoda,family=quasipoisson())
4 > summary(fitqp)

Deviance Residuals:
     1        2        3        4        5        6
 4.698   -2.173   -2.952   -8.436   -1.605    8.067

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0950     0.2116  24.075 1.77e-05 ***
pH1          -0.1415     0.2116  -0.669     0.54

(Dispersion parameter for quasipoisson family taken to be 43.42048)

    Null deviance: 193.93  on 5  degrees of freedom
Residual deviance: 174.32  on 4  degrees of freedom
AIC: NA
```

The parameter estimates in the linear predictor are exactly the same as before, as are the deviances. However, the dispersion parameter is estimated at 43.42, and this increases the standard errors of the linear parameters by a factor of 6.59 ($= \sqrt{43.42}$). With this correction, pH no longer looks significant.

Even though this model only has a quasi-likelihood, we go ahead and compute the F-test for pH in line 5:

```
5 > (193.93-174.32)/1/43.42
[1] 0.4516352
6 > .669^2
[1] 0.447561
7 > anova(lm(count~pH,data=Copepoda))
Response: count
          Df  Sum Sq Mean Sq F value Pr(>F)
pH         1  3220.2  3220.2  0.4893 0.5228
Residuals  4 26326.7  6581.7
```

Line 6 squares the t-test for pH in line 4, which gives us an F-test very close to what we just computed.

Finally, given that the data have little difference in mean, and thus little anticipated difference in variance, let's try fitting the data with an ordinary linear model as in line 7. Note that in this case the inference from the normal (Gaussian) model is nearly identical to that from the quasi-Poisson model.

6.3.7   Accommodating dependence

Methods for dealing with dependence in data are generally complicated. We will present only the simplest possible accommodation, leaving more more complicated methods to other authors.

An *autoregressive model of order 1* (AR1) is a model for data where there is a time order, and data k steps apart have a correlation ρ^k. Because $|\rho| < 1$, this correlation can be large for nearby data values but decays fairly quickly to 0 as data are separated more in time. In our context, we will assume that

AR1 model

the errors ϵ_{ij} are mean 0, constant variance, but have an AR1 correlation structure across time. In some contexts, one might assume an AR1 structure within certain blocks of data, but independence between blocks of data (think of data collected sequentially but at multiple sites).

While the AR1 model will not be adequate for all data with time depen- dence, it is a good start and can indicate the effect that the dependence is likely to have on our inferences.

Example 6.15 Temperature differences between thermocouples, continued

Example 6.4 gave the difference in temperature readings between two thermocouples at 64 consecutive times. A plot of the differences showed substantial autocorrelation.

```
1 > fit1 <- lm(temp.diff~1,data=Thermocouples)
2 > summary(fit1)
  ...
  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept) 3.141250   0.003138    1001   <2e-16 ***
  ...
3 > fit2 <- gls(temp.diff~1,data=Thermocouples,cor=corAR1(form=~(1:64)))
4 > summary(fit2)
  Generalized least squares fit by REML
    Model: temp.diff ~ 1
    Data: thermocouples
          AIC       BIC   logLik
    -278.9285 -272.4991 142.4643

  Correlation Structure: AR(1)
   Formula: ~(1:64)
   Parameter estimate(s):
       Phi
  0.243579

  Coefficients:
                Value   Std.Error  t-value p-value
  (Intercept) 3.141487 0.004027105 780.0856       0
  ...
```

Line 1 fits the single-mean model to the data via ordinary least squares (OLS), and line 2 shows the summary information for the estimated coef- ficient. Line 3 fits the same mean model but now uses `gls` and sets the correlation structure to be AR1. The argument `form=~x` says that the time steps are in variable `x`. Because our data are already in time order, we can simply write `form=~(1:64)`.

Line 4 shows the summary for the AR1 fit. The lag one autocorrelation is estimated as .24. The estimated mean is essentially the same, but the stan- dard error has increased from .003 to .004. This is the result of the positive autocorrelation in the data.

Note that the output to line 4 indicates that the fit was calculated using REML; this is *restricted maximum likelihood*. Recall than REML likelihoods

Table 6.11: Time to upload and download a 50MB file to three cloud backup services, in seconds. Data are in time sequence down columns. Data set `CloudBackup`.

Svc.	Sec.	Svc.	Sec.	Svc.	Sec.	Svc.	Sec.	Svc.	Sec.
3	564	2	215	2	250	2	248	2	241
1	172	2	203	2	276	1	175	3	461
3	455	3	560	2	265	3	467	3	417
1	147	1	185	3	585	3	490	1	170
2	200	3	560	1	192	1	146	2	208
1	141	2	233	3	510	1	126	1	145

can only be compared to other REML models with the same mean structure and cannot be compared to ML likelihoods.

```
5 > fit2b <- gls(temp.diff~1,data=Thermocouples,cor=corAR1(form=~(1:64)),method='ML')
6 > summary(fit2b)
 Generalized least squares fit by maximum likelihood
   Model: temp.diff ~ 1
   Data: thermocouples
       AIC       BIC    logLik
   -288.1547 -281.678 147.0773

 Correlation Structure: AR(1)
  Formula: ~(1:64)
  Parameter estimate(s):
        Phi
 0.2231942

 Coefficients:
                 Value    Std.Error  t-value p-value
 (Intercept) 3.141461 0.003923022 800.7759       0
 ...
7 > AIC(fit1,fit2b)
      df        AIC
it1    2 -287.0342
it2b   3 -288.1547
```

Line 5 refits the model using maximum likelihood. You can see from lines 4 and 6 that the estimates obtained from REML and ML are similar. Estimates of variances under REML have less bias than those produced by ML; we will see this again in more detail when we estimate variance components.

The important thing to understand now is that likelihoods and AIC values computed under ML and REML are *not* comparable. We can compare likelihoods and AIC values for OLS (`fit1`) and ML (`fit3`) as we do in line 7, but we cannot use likelihood or AIC to compare `fit2` to either `fit1` or `fit3`.

Example 6.16 File backup speed

I need to find a system to backup my computer files to the cloud. All other issues being equal, I want to find the fastest service, so I sign up for

a free trial period with three different providers and test their speeds. Speed will depend on the speed of the backup service itself and also on the speed of my internet service. I expect both of these to vary over time depending on the load on the backup system and the load on my local network branch.

I have a 50MB file as a test case. An experimental run will consist of uploading the file to the backup server, and then downloading it back from the server. The response is the number of seconds to complete the upload/download cycle. I run the upload/download cycle 30 times, using each service 10 times, with the services randomly assigned to the time slots. Table 6.11 gives the results.

We begin by fitting the usual separate-means model in line 1.

```
1 > fit1 <- lm(updowntime~service,data=CloudBackup)
2 > plot(fit1,which=1)
3 > car::boxCox(fit1)
4 > fit2 <- lm(log(updowntime)~service,data=CloudBackup)
5 > summary(fit2)
  ...
  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept)  5.57923    0.02233 249.897  < 2e-16 ***
  service1    -0.51305    0.03157 -16.249 1.83e-15 ***
  service2    -0.13017    0.03157  -4.123  0.00032 ***
  ...
6 > plot(residuals(fit2),type='b');abline(h=0)
7 > car::durbinWatsonTest(fit2)
   lag Autocorrelation D-W Statistic p-value
    1          0.44276         1.06326   0.012
   Alternative hypothesis: rho != 0
```

The residual plot in line 2 (see Figure 6.13 panel 1) suggests non-constant variance, and the Box-Cox plot (panel 2) suggests a log transformation. Line 4 fits the separate-means model to the log-transformed data, with an abbreviated summary in line 5. So far, so good, but I know that the data were collected in time order, and I suspect that there could be autocorrelation. Line 6 plots the residuals in time order (panel 3 of Figure 6.13), and autocorrelation seems to be present. The Durbin-Watson statistic of 1.06 (line 7) confirms that visual diagnosis.

To deal with autocorrelation, we refit using `gls` with an AR1 structure across the 30 time slots in line 8:

```
8 > fit3 <- gls(log(updowntime)~service,cor=corAR1(form= ~(1:30)),data=CloudBackup)
9 > summary(fit3)
...
Correlation Structure: AR(1)
 Formula: ~(1:30)
 Parameter estimate(s):
      Phi
0.562057

Coefficients:
                Value  Std.Error   t-value p-value
(Intercept)  5.579809 0.04224382 132.08581       0
service1    -0.483829 0.02136900 -22.64164       0
service2    -0.138539 0.02482215  -5.58126       0
...
```

Line 9 shows the summary of the fit. The lag-1 autocorrelation is estimated to be .562, which is fairly substantial. The estimate mean and estimated effects do not change much relative to the OLS fit, but the standard errors change dramatically. The standard error of $\widehat{\mu}$ nearly doubles when correlation is taken into account, but the standard errors of the treatment effects are roughly one third *smaller* after accounting for correlation.

Postive autocorrelation will increase the standard error of a mean (relative to uncorrelated data with the same variance), and it will decrease the standard error of differences of means.

The data in this example were entered in time order, so we could indicate time order in line 8 for our 30 data points via `form=~(1:30)`. We could have used data in another order, perhaps ordered by service provider, if we have an additional variable that indicates the order in which the data were run. For example, if `time.slot` is 1 for the first run, 2 for the second run, and so on, then we can use the data in any order and set the correlation via `form=~time.slot`.

6.3.8 Bayesian approaches

Bayesian analysis is also sensitive to unmet assumptions or mis-specification of the model. Thus a Bayesian statistician also needs to detect problems and provide alternatives if needed. Bayesian statisticians have an advantage in this process, as the Bayesian inferential paradigm (posterior distributions, Bayes factor, and so on) is essentially the same regardless of the models used for the data. Thus Bayesian accommodation of unmet assumptions consists of using different model assumptions and turning the Bayesian crank again. No new methods *per se* are needed.

Bayesian methods also depend on assumptions

The `bglmm` function in `cfcdae` has the ability to fit a variety of model forms that will meet many needs when data do not conform to standard assumptions. `bglmm` does this via a `family` argument that works much like that in `glm`, a `varform` argument that allows the user to structure the variance when the family is normal, and a `boxcox` argument that requests a
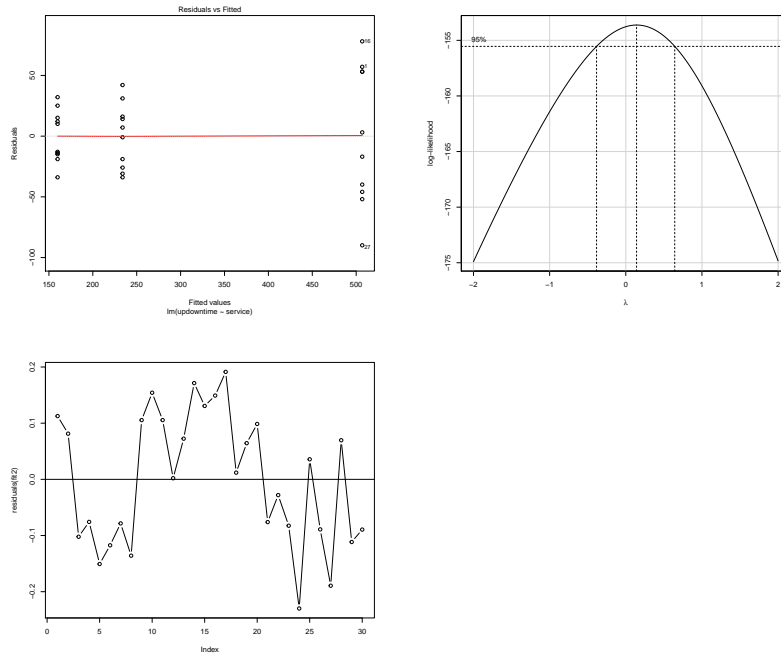
Figure 6.13: Diagnostic plots for the cloud backup example. Panel 1 shows residuals versus fitted values for raw data, and panel 2 shows the corresponding Box-Cox plot. Panel 3 shows residuals in time order for log transformed data.

transformation of the response. Here we briefly revisit several of the previous examples, but using Bayesian methods rather than frequentist methods.

Example 6.17 Cloud seeding, continued

In Example 6.6, we used a Box-Cox analysis to find a transformation of the data. We can do a similar thing in a Bayesian context. The argument `boxcox=TRUE` tells `bglmm` to fit coefficients and a Box-Cox transformation power simultaneously.

```
1 > fit <- bglmm(rainfall~seeding,data=CloudSeeding,boxcox=TRUE)
2 > summary(fit)[,c(1,2,4,6,8:10)]
                      mean   se_mean     2.5%       50%     97.5% n_eff Rhat
  (Intercept)     382.0000   1.19000 290.0000  377.0000  497.000  2040    1
  seeding1        -52.5000   0.41900 -95.4000  -52.7000   -7.300  2880    1
  sigma0          160.0000   0.28500 131.0000  159.0000  195.000  3250    1
  lambda            0.0817   0.00163  -0.0539    0.0815    0.231  2040    1
  sigma.Intercept 1630.0000 26.80000 246.0000 1180.0000 5620.000  3010    1
  sigma.seeding   422.0000   7.06000  40.4000  283.0000 1570.000  3380    1
```

Line 1 shows the fitting command (omitting all the MCMC progress output),

and line 2 shows a partial summary of the results. The MCMC ran well, and the credible interval for `lambda`, the transformation power, is from –.05 to .23. This interval corresponds well to what we got applying `car::boxCox` to the OLS fit. Note that `bglmm` model estimates the coefficients at the same time as the transformation power, making these coefficients extremely difficult to interpret, as they are being applied across a range of transformation powers.

Example 6.18 Tissue Vaporization, continued

In Example 6.9, we used `rlm` to fit the vaporization data robustly. Here we repeat the process, comparing the use of the `gaussian` and `t4` (Student t with 4 degrees of freedom) families.

```
1 > fitb <- bglmm(width~power.factor-1,data=TissueVaporization,adapt_delta = .95)
2 > fitb2 <- bglmm(width~power.factor-1,data=TissueVaporization,family="t4",
           adapt_delta = .95)
3 > summary(fitb)[,c(1,3,4,6,8:10)]
                     mean     sd  2.5%    50% 97.5% n_eff Rhat
 power.factor1      1.110 0.262 0.565 1.110 1.630  3210    1
 power.factor2      1.520 0.266 0.979 1.520 2.020  2960    1
 power.factor3      1.860 0.270 1.300 1.870 2.400  3250    1
 power.factor4      1.900 0.259 1.350 1.910 2.400  2960    1
 sigma0             0.444 0.132 0.266 0.419 0.774  1680    1
 sigma.power.factor 1.710 0.562 0.939 1.600 3.140  2870    1
4 > summary(fitb2)[,c(1,3,4,6,8:10)]
                     mean     sd   2.5%    50% 97.5% n_eff Rhat
 power.factor1      1.130 0.144 0.8190 1.130 1.400  2690    1
 power.factor2      1.540 0.149 1.2300 1.550 1.840  2720    1
 power.factor3      1.900 0.151 1.5800 1.910 2.200  2150    1
 power.factor4      2.210 0.273 1.4500 2.280 2.560  1160    1
 sigma0             0.217 0.116 0.0799 0.189 0.513   787    1
 sigma.power.factor 1.820 0.607 1.0100 1.680 3.350  1750    1
5 > plot(fitb2)
```

Lines 1 and 2 fit the model assuming normal errors, and then again assuming t4 errors. Looking at the summary on line 3, we can see that the third and fourth treatment means are roughly equal, which implies that Bayesian analysis using normal errors is also susceptible to the outlier. Line 4 shows the results assuming t4 errors, and we see that the fourth treatment mean is now substantially higher. We also see that the standard error of our estimate of the fourth treatment mean is higher than the other three means. This is reasonable given that the outlier gets a lower weight in the analysis. Also note that the variability of the mean estimates overall are larger for the Bayesian approach.

Finally, line 5 illustrates that there is a `plot` method for objects of class `bglmm`. It plots Pearson residuals versus fitted values and normal scores, as shown in Figure 6.14. The outlier is not quite as down weighted in the t4 analysis as it was in the `rlm` analysis. We can see this because the non-outlying residuals in the fourth treatment group are a little above zero. They would be centered on zero if the outlier were totally down weighted.
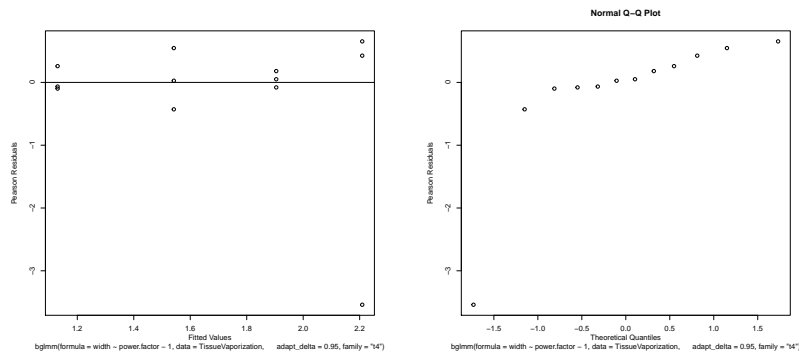
Figure 6.14: Diagnostic plots for the Bayesian analysis of the tissue vaporization data using a `t4` family. Panel 1 shows Pearson residuals versus fitted values, and panel 2 shows a normal plot of the Pearson residuals.

Example 6.19 Resin lifetimes, continued

In Example 6.13, we accommodated non-constant variance by modeling the residual variance as either separate by group or as a power of the mean. We can do both of those in the Bayesian context as well.

```
1 > fit <- bglmm(Time~temp.factor-1,data=ResinLifetimes)
2 > fitb <- bglmm(Time~temp.factor-1,data=ResinLifetimes,varform= ~1|temp.factor)

3 > summary(fitb)[,c(1,3,4,6,8:10)]
                   mean    sd  2.5%   50%  97.5% n_eff Rhat
 temp.factor1     84.80  6.56 70.40 85.10  96.80 2080    1
 temp.factor2     43.10  4.46 34.00 43.30  52.00 2220    1
 temp.factor3     24.50  3.04 18.30 24.40  30.60 2610    1
 temp.factor4     15.70  1.07 13.70 15.70  17.70 1300    1
 temp.factor5     11.90  2.52  6.65 11.90  16.90 2220    1
 sigma0[1]        16.60  5.92  9.26 15.40  31.30 1790    1
 sigma0[2]        12.10  4.44  6.67 11.10  23.50 2130    1
 sigma0[3]         8.20  2.94  4.56  7.55  15.70 2050    1
 sigma0[4]         2.43  1.22  1.23  2.14   5.34 1070    1
 sigma0[5]         5.52  2.80  2.56  4.79  13.30 1880    1
 sigma.temp.factor 54.70 20.10 28.40 50.60 103.00 2800   1
4 > fitb2 <- bglmm(Time~temp.factor-1,data=ResinLifetimes,varform= ~1|temp.factor,
      sigma.scale0=5)
5 > summary(fitb2)[,c(1,3,4,6,8:10)]
                   mean     sd  2.5%   50%  97.5% n_eff  Rhat
 temp.factor1     85.80  4.490 76.40 85.80  94.40  4000 1.000
 temp.factor2     43.30  3.480 36.20 43.40  50.10  4000 1.000
 temp.factor3     24.50  2.450 19.70 24.50  29.40  4000 1.000
 temp.factor4     15.70  0.856 14.00 15.70  17.40  4000 0.999
 temp.factor5     11.80  1.850  8.06 11.90  15.50  4000 1.000
 sigma0[1]        12.30  2.890  7.95 11.90  19.20  4000 1.000
 sigma0[2]         9.55  2.390  6.03  9.20  15.30  4000 0.999
 sigma0[3]         6.77  1.790  4.22  6.48  11.10  4000 0.999
 sigma0[4]         2.14  0.759  1.19  1.98   4.07  2550 1.000
 sigma0[5]         4.31  1.460  2.41  4.00   8.03  4000 1.000
 sigma.temp.factor 54.90 21.400 28.70 50.20 111.00  4000 1.000
```

Line 1 fits the resin lifetimes using a model assuming a constant variance, and line 2 fits the same mean structure but instead allows the variance to be different for each level of temp.factor via the varform argument. The summary from line 3 shows some differences from what we obtained using gls. First, the more extreme means have been shrunk back toward 0, which is the prior expectation for the treatment means in our model. Second, the standard errors of the treatment means are considerably larger than what we obtained with gls. This is reasonable, because the Bayesian approach accounts for the fact that the within group standard deviations are all unknown and could potentially be quite large, whereas the gls approach is using an estimated standard error with the expectation that it will be inflated properly (in the t multiplier) to reflect the variability in its expectation. Also affecting this result is the fact the Bayesian analysis estimates the error standard deviations to be larger. The default prior mean for the error standard deviation(s) is set assuming that the model terms are not very predictive; that is, it is set to be fairly large. In this case, the model is highly predictive, so the prior for the standard deviations is dragging the estimates up. Refitting with a more reasonably prior value, as in line 4, gives summary values from line 5 that are more similar to the gls results. The moral here is that the automatically generated priors are not always appropriate.

The second model we explored using gls was fitting standard deviations

proportional to a power of the fitted value. We can also do that with `bglmm`.

```
 6 > fitb3 <- bglmm(Time~temp.factor-1,data=ResinLifetimes,varform="power",
         sdpower.range = c(0,1))
 7 > summary(fitb3)[,c(1,3,4,6,8:10)]
                  mean     sd  2.5%   50%  97.5% n_eff Rhat
   temp.factor1     4.460 0.0558 4.340 4.460  4.560  3540   1
   temp.factor2     3.770 0.0690 3.640 3.780  3.910  3700   1
   temp.factor3     3.200 0.0913 3.000 3.200  3.370  3720   1
   temp.factor4     2.710 0.1200 2.470 2.710  2.940  3300   1
   temp.factor5     2.450 0.1610 2.110 2.460  2.740  2420   1
   sigma0           1.270 1.0200 0.268 0.965  4.120  1540   1
   power            0.574 0.1900 0.193 0.574  0.938  1470   1
   sigma.temp.factor 5.050 2.5600 2.350 4.380 12.000  1730   1
 8 > exp(c(4.46,3.77,3.20,2.71,2.45))
[1] 86.48751 43.38006 24.53253 15.02928 11.58835
 9 > 1.27*exp(c(4.46,3.77,3.20,2.71,2.45))^.574
[1] 16.429145 11.056261  7.971055  6.016808  5.182652
10 > .965*exp(c(4.46,3.77,3.20,2.71,2.45))^.574
[1] 12.483563  8.401017  6.056747  4.571827  3.937999
```

Line 6 fits the model using the `"power"` option for `varform`, and line 7 gives the summary. *Note:* the power option uses a log link so that the fitted values, which are the exponential of the linear predictor, are guaranteed to positive. Line 8 shows the exponentiation and the resulting fitted means, which are quite close to the `gls` results.

Lines 9 and 10 show the fitted standard deviations using the posterior mean and median values for the error standard deviation. The latter are rather closer to the `gls` results than the former. The difference is due to the large asymmetry in the results for the standard deviation.

We used GLMs to allow us to consider several distributions for our response as alternatives to the normal (Gaussian). The same distributions (and potentially many more), are available in the Bayesian context.

Example 6.20 Germination of garden peas

Example 6.20 introduced data on the germination of garden peas, with eight binomial responses, each based on 20 trials. We can easily replicate that analysis with Bayesian methods using a `family = "binomial"` argument.

```
 1 > fit <- bglmm(cbind(germinated,nongerminated)~treatment,data=PeaGermination,
         family="binomial",adapt_delta=.9)
 1 > summary(fit)[,c(1,3,4,6,8:10)]
                    mean    sd    2.5%    50%  97.5% n_eff  Rhat
   (Intercept)      0.400 0.205  0.0123  0.403  0.808 2990 1.000
   treatment1       0.329 0.312 -0.2650  0.324  0.959 3550 1.000
   treatment2       0.445 0.317 -0.1710  0.441  1.070 3420 0.999
   treatment3      -2.310 0.399 -3.1400 -2.290 -1.580 2500 1.000
   sigma.Intercept  7.760 9.270  0.2790  4.280 32.800 3630 1.000
   sigma.treatment  3.260 2.440  0.9960  2.570 10.300 2200 1.000
```

Line 1 fits the model. As with `glm`, the binomial response is a matrix with

two columns for successes and failures. The summary in line 2 shows results very similar to those obtained using `glm`.

Example 6.21 File backup speed, continued

Finally, for normally distributed data you can specify an AR1 error structure. Example 6.16 introduced backup times for a file to three different remote servers. We observed considerable autocorrelation in the residuals, and thus need to accommodate that correlation in our model.

```
1 > fit <- bglmm(log(updowntime)~service,data=CloudBackup,
    varform=c("ar1","time","timegroups"))
2 > summary(fit)[,c(1,3,4,6,8:10)]
                    mean      sd    2.5%      50%   97.5% n_eff Rhat
 (Intercept)      5.6000  0.0475  5.5200   5.5900   5.720  1690    1
 service1        -0.4890  0.0259 -0.5410  -0.4890  -0.440  3270    1
 service2        -0.1400  0.0286 -0.1950  -0.1400  -0.082  3200    1
 sigma0           0.0882  0.0237  0.0402   0.0875   0.138  1780    1
 rho              0.4620  0.1750  0.1150   0.4650   0.804  1840    1
 sigma.Intercept  7.2000  4.0000  2.5700   6.2300  17.800  3040    1
 sigma.service    0.9540  0.5830  0.3400   0.7990   2.510  2260    1
3 > plot(fit,plottype="pairs",pars=c("sigma0","rho"))
```

Line 1 fits the AR1 model. The `varform` argument is a triple of character strings. The first is `"ar1"` indicating the model form. The second is the name of the variable that indicates time, and the third is the name of the variable that indicates groups that are correlated within group but independent between groups. In this case, the time variable is just 1 through 30, and the groups variable is all 1. The summary from line 2 produces estimates similar to what we found before.

The command in line 3 plots the MCMC values for `sigma0` and `rho` against each other over time. In Figure 6.15, the smoothed scatterplot shows substantial negative correlation between these values: when `rho` is small, `sigma0` is large, and vice versa. This illustrates how parameters in these more complex models can compensate for each other.

6.4 Implications for Design

The major implication for design is that balanced data sets are usually a good idea. Balanced data are less susceptible to the effects of nonnormality and nonconstant variance. Furthermore, when there is nonconstant variance, we can usually determine the direction in which we err for balanced data.

Use balanced designs

When we know that our measurements will be subject to temporal or spatial correlation, we should take care to block and randomize carefully. We can, in principle, use the correlation in our design and analysis to increase precision, but these methods are beyond this text.
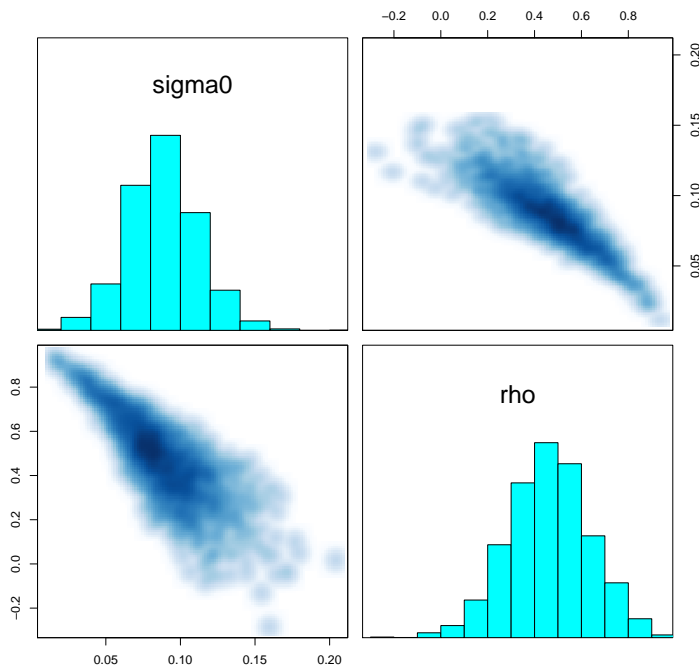
Figure 6.15: Pairwise plot of `rho` and `sigma0` values across
iterations of the Markov Chain, using the cloud backup data.
Diagonal panels show histograms of the values; off diagonal panels
show smoothed scatterplots of the values.

6.5   Further Reading and Extensions

Statisticians started worrying about what would happen to their t-tests and F-tests on real data almost immediately after they started using the tests. See, for example, Pearson (1931). Scheffé (1959) provides a more mathematical introduction to the effects of violated assumptions than we have given here. Ito (1980) also reviews the subject.

Transformations have long been used in Analysis of Variance. Tukey (1957a) puts the power transformations together as a family, and Box and Cox (1964) introduce the scaling required to make the SS_E's comparable. Atkinson (1985) and Hoaglin, Mosteller, and Tukey (1983) give more extensive treatments of transformations for several goals, including symmetry and equalization of spread.

The Type I error rates for nonnormal data were computed using the methods of Gayen (1950). Gayen assumed that the data followed an Edgeworth distribution, which is specified by its first four moments, and then computed

the distribution of the F-ratio (after several pages of awe-inspiring calculus). Our Table 6.2 is computed with his formula (2.30), though note that there are typos in his paper.

Box and Andersen (1955) approached the same problem from a different tack. They computed the mean and expectation of a transformation of the F-ratio under the permutation distribution when the data come from nonnormal distributions. From these moments they compute adjusted degrees of freedom for the F-ratio. They concluded that multiplying the numerator and denominator degrees of freedom by $(1 + \gamma_2/N)$ gave p-values that more closely matched the permutation distribution.

There are two enormous, parallel areas of literature that deal with outliers. One direction is outlier identification, which deals with finding outliers, and to some extent with estimating and testing after outliers are found and removed. Major references include Hawkins (1980), Beckman and Cook (1983), and Barnett and Lewis (1994). The second direction is robustness, which deals with procedures that are valid and efficient for nonnormal data (particularly outlier-prone data). Major references include Andrews *et al.* (1972), Huber (1981), and Hampel *et al.* (1986). Hoaglin, Mosteller, and Tukey (1983) and Rey (1983) provide gentler introductions.

Rank-based, nonparametric methods are a classical alternative to linear methods for nonnormal data. In the simplest situation, the numerical values of the responses are replaced by their ranks, and we then do randomization analysis on the ranks. This is feasible because the randomization distribution of a rank test can often be computed analytically. Rank-based methods have sometimes been advertised as assumption-free; this is not true. Rank methods have their own strengths and weakness. For example, the power of two-sample rank tests for equality of medians can be very low when the two samples have different spreads. Conover (1980) is a standard introduction to nonparametric statistics.

We computed approximate test sizes for F under nonconstant variance using a method given in Box (1954). When our distributional assumptions and the null hypothesis are true, then our observed F-statistic F_{obs} is distributed as F with $g - 1$ and $N - g$ degrees of freedom, and

$$P(F_{\text{obs}} > F_{\mathcal{E}, g-1, N-g}) = \mathcal{E}.$$

If the null is true but we have different variances in the different groups, then F_{obs}/b is distributed approximately as $F(\nu_1, \nu_2)$, where

$$b = \frac{N - g}{N(g - 1)} \frac{\sum_i (N - n_i)\sigma_i^2}{\sum_i (n_i - 1)\sigma_i^2} \ ,$$

$$\nu_1 = \frac{[\sum_i (N - n_i)\sigma_i^2]^2}{[\sum_i n_i\sigma_i^2]^2 + N\sum_i (N - 2n_i)\sigma_i^4} \ ,$$

$$\nu_2 = \frac{[\sum_i (n_i - 1)\sigma_i^2]^2}{\sum_i (n_i - 1)\sigma_i^4} \ .$$

Thus the actual Type I error rate of the usual F-test under nonconstant variance is approximately the probability that an F with ν_1 and ν_2 degrees of freedom is greater than $F_{\mathcal{E}, g-1, N-g}/b$.

The Durbin-Watson statistic was developed in a series of papers (Durbin and Watson 1950, Durbin and Watson 1951, and Durbin and Watson 1971). The distribution of DW is complicated in even simple situations. Ali (1984) gives a (relatively) simple approximation to the distribution of DW.

There are many more methods to test for serial correlation. Several fairly simple related tests are called runs tests. These tests are based on the idea that if the residuals are arranged in time order, then positive serial correlation will lead to "runs" in the residuals. Different procedures measure runs differently. For example, Geary's test is the total number of consecutive pairs of residuals that have the same sign (Geary 1970). Other runs include maximum number of consecutive residuals of the same sign, the number of runs up (residuals increasing) and down (residuals decreasing), and so on.

In some instances we might believe that we know the correlation structure of the errors. For example, in some genetics studies we might believe that correlation can be deduced from pedigree information. If the correlation is known, it can be handled simply and directly by using generalized least squares (Weisberg 1985).

We usually have to use advanced methods from times series or spatial statistics to deal with correlation. Anderson (1954), Durbin (1960), Pierce (1971), and Tsay (1984) all deal with the problem of regression when the residuals are temporally correlated. Kriging is a class of methods for dealing with spatially correlated data that has become widely used, particularly in geology and environmental sciences. Cressie (1991) is a standard reference for spatial statistics. Grondona and Cressie (1991) describe using spatial statistics in the analysis of designed experiments.

6.6   Problems

As part of a larger experiment, 32 male hamsters were assigned to four treatments in a completely randomized fashion, eight hamsters per treatment. The treatments were 0, 1, 10, and 100 nmole of melatonin daily, 1 hour prior to lights out for 12 weeks. The response was paired testes weight (in mg). Below are the means and standard deviations for each treatment group (data from Rollag 1982, data set `Melatonin`). What is the problem with these data and what needs to be done to fix it?

Exercise 6.1

Melatonin	Mean	SD
0 nmole	3296	90
1 nmole	2574	153
10 nmole	1466	207
100 nmole	692	332

Bacteria in solution are often counted by a method known as serial dilution plating. Petri dishes with a nutrient agar are inoculated with a measured

Exercise 6.2

amount of solution. After 3 days of growth, an individual bacterium will have grown into a small colony that can be seen with the naked eye. Counting original bacteria in the inoculum is then done by counting the colonies on the plate. Trouble arises because we don't know how much solution to add. If we get too many bacteria in the inoculum, the petri dish will be covered with a lawn of bacterial growth and we won't be able to identify the colonies. If we get too few bacteria in the inoculum, there may be no colonies to count. The resolution is to make several dilutions of the original solution (1:1, 10:1, 100:1, and so on) and make a plate for each of these dilutions. One of the dilutions should produce a plate with 10 to 100 colonies on it, and that is the one we use. The count in the original sample is obtained by multiplying by the dilution factor.

Suppose that we are trying to compare three different Pasteurization treatments for milk. Fifteen samples of milk are randomly assigned to the three treatments, and we determine the bacterial load in each sample after treatment via serial dilution plating. The following table gives the counts (data set `Pasteurization`).

Treatment	Count				
1	26×10^2	29×10^2	20×10^2	22×10^2	32×10^2
2	35×10^3	23×10^3	20×10^3	30×10^3	27×10^3
3	29×10^5	23×10^5	17×10^5	29×10^5	20×10^5

Test the null hypothesis that the three treatments have the same effect on bacterial concentration.

Exercise 6.3

In order to determine the efficacy and lethal dosage of cardiac relaxants, anesthetized guinea pigs are infused with a drug (the treatment) till death occurs. The total dosage required for death is the response; smaller lethal doses are considered more effective. There are four drugs, and ten guinea pigs are chosen at random for each drug. Lethal dosages follow (data set `LethalDosage`).

1	18.2	16.4	10.0	13.5	13.5	6.7	12.2	18.2	13.5	16.4
2	5.5	12.2	11.0	6.7	16.4	8.2	7.4	12.2	6.7	11.0
3	5.5	5.0	8.2	9.0	10.0	6.0	7.4	5.5	12.2	8.2
4	6.0	7.4	12.2	11.0	5.0	7.4	7.4	5.5	6.7	5.5

Determine which drugs are equivalent, which are more effective, and which less effective.

Exercise 6.4

Four overnight delivery services are tested for "gentleness" by shipping fragile items. The breakage rates observed are given below (data set `Breakage`):

Treatment	Rate				
A	17	20	15	21	28
B	7	11	15	10	10
C	11	9	5	12	6
D	5	4	3	7	6

You immediately realize that the variance is not stable. Find an approximate 95% confidence interval for the transformation power using the Box-Cox method.

Consider the following four plots. Describe what each plot tells you about the assumptions of normality, independence, and constant variance. (Some plots may tell you nothing about assumptions.)

Exercise 6.5

a)

b)

c)

Normal Q–Q Plot

d)
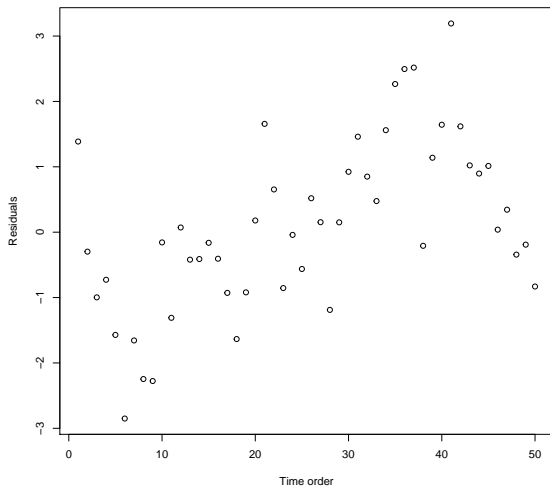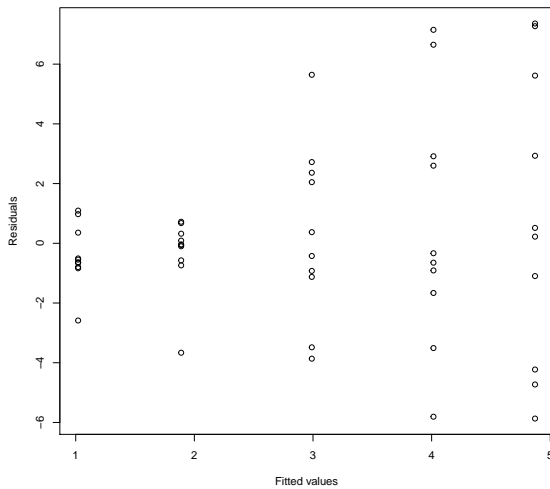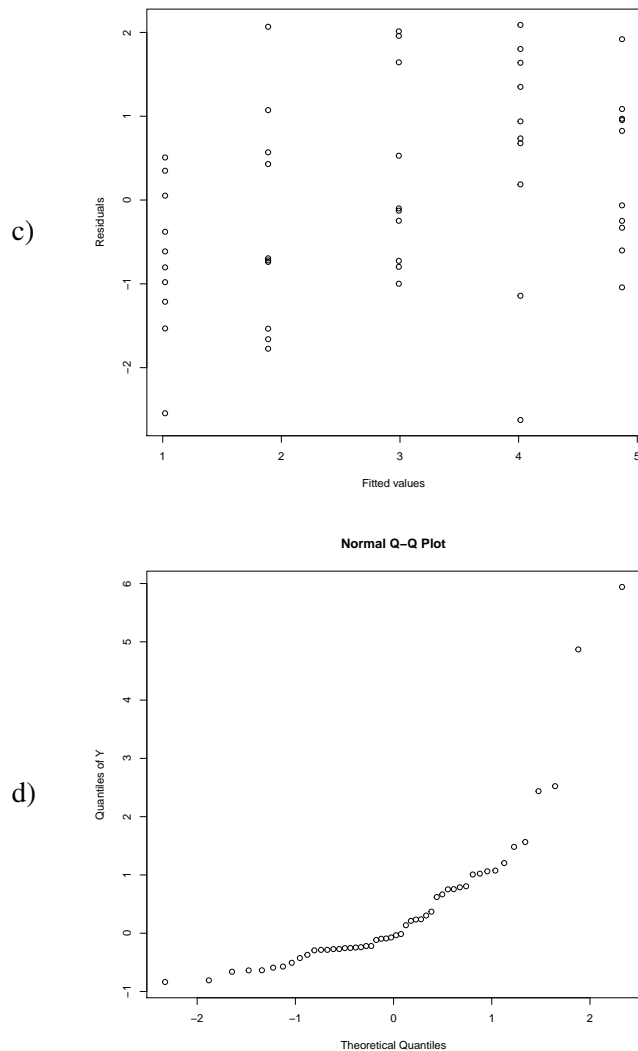
An instrument called a "Visiplume" measures ultraviolet light. By comparing absorption in clear air and absorption in polluted air, the concentration of SO_2 in the polluted air can be estimated. The EPA has a standard method for measuring SO_2, and we wish to compare the two methods across a range of air samples. The recorded response is the ratio of the Visiplume reading to the EPA standard reading. The four experimental conditions are: measurements of SO_2 in an inflated bag (n = 9), measurements of a smoke generator with SO_2 injected (n = 11), measurements at two coal-fired plants (n = 5 and 6). We are interested in whether the Visiplume instrument performs the same relative to the standard method across all experimental conditions, between the coal-fired plants, and between the generated smoke and the real coal-fired smoke. The data follow (McElhoe and Conner 1986, data set `Visiplume`):

Exercise 6.6

Condition	Ratio						
Bag	1.055	1.272	.824	1.019	1.069	.983	1.025
	1.076	1.100					
Smoke	1.131	1.236	1.161	1.219	1.169	1.238	1.197
	1.252	1.435	.827	3.188			
Plant no. 1	.798	.971	.923	1.079	1.065		
Plant no. 2	.950	.978	.762	.733	.823	1.011	

Problem 6.1

We wish to study the competition of grass species: in particular, big bluestem (from the tall grass prairie) versus quack grass (a weed). We set up an experimental garden with 24 plots. These plots were randomly allocated to the six treatments: nitrogen level 1 (200 mg N/kg soil) and no irrigation; nitrogen level 1 and 1cm/week irrigation; nitrogen level 2 (400 mg N/kg soil) and no irrigation; nitrogen level 3 (600 mg N/kg soil) no irrigation; nitrogen level 4 (800 mg N/kg soil) and no irrigation; and nitrogen level 4 and 1 cm/week irrigation. Big bluestem was seeded in these plots and allowed to establish itself. After one year, we added a measured amount of quack grass seed to each plot. After another year, we harvest the grass and measure the fraction of living material in each plot that is big bluestem. We wish to determine the effects (if any) of nitrogen and/or irrigation on the ability of quack grass to invade big bluestem. (Based on Wedin 1990, data set Quackgrass.)

N level	Irrigation	Percent Bluestem			
1	No	97	96	92	95
1	Yes	83	87	78	81
2	No	85	84	78	79
3	No	64	72	63	74
4	No	52	56	44	50
4	Yes	48	58	49	53

(a) Do the data need a transformation? If so, which transformation?

(b) Provide an Analysis of Variance for these data. Are all the treatments equivalent?

(c) Are there significant quadratic effects of nitrogen under nonirrigated conditions?

(d) Is there a significant effect of irrigation?

(e) Under which conditions is big bluestem best able to prevent the invasion by quack grass? Is the response at this set of conditions significantly different from the other conditions?

Problem 6.2

Tajima (1987) describes an experiment examining the effect of a freeze-thaw cycle on the potency of semen used for artificial insemination in chickens. Four semen mixtures are prepared. Each mixture consists of equal volumes of semen from Rhode Island Red and White Leghorn roosters. Mixture

1 has both varieties fresh, mixture 4 has both varieties frozen, and mixtures 2 and 3 each have one variety fresh and the other frozen. Sixteen batches of Rhode Island Red hens are inseminated with the mixtures, using a balanced completely randomized design. The response is the fraction of chicks from each batch that have white feathers (white feathers indicate a White Leghorn father). The observed proportions ranged from .19 to .95.

(a) What problems would you anticipate from these data?
(b) How would you expect to address them?

Problem 6.3

Even good cookies harden over time. This experiment was conducted to determine if adding raffinose (an inhibitor of sucrose crystallization) would improve cookie texture.

Two-hundred twenty cookies are baked. Half of them are made with a control recipe, while the other half include some raffinose. The 110 cookies of each recipe are then randomly divided into 11 groups of ten each. Each group of 10 is put in a plastic bag. On the next day, one bag is chosen at random from each recipe, and the 10 cookies in that bag are measured for texture. The texture measurement is the force required to push a 1.27 cm diameter probe into the cookie. This is repeated on each successive day until the last bag of cookies has been measured on day 11.

For this question, our interest is in modeling how control-recipe cookies harden over time. The control data are shown below, and the full data set is `CookieTexture`.

Day	Force									
1	89.5	79.8	127.6	64.7	90.1	67.1	61.7	52.4	104.1	111.4
2	130.5	159.0	204.1	168.9	123.2	134.6	135.1	119.5	109.8	118.2
3	288.4	375.0	239.2	185.8	188.2	180.7	233.4	242.9	165.9	331.0
4	172.7	369.1	453.6	239.0	269.4	370.2	452.8	563.9	390.5	296.2
5	412.1	315.3	471.1	236.0	806.5	290.7	315.0	314.0	434.6	534.1
6	545.7	443.2	414.6	333.6	277.7	557.5	339.8	801.5	377.0	464.2
7	424.4	554.2	620.6	559.2	671.6	681.2	735.3	324.5	439.7	476.0
8	792.3	495.9	864.7	652.3	584.1	688.2	926.2	671.4	536.6	482.9
9	870.7	1148.2	860.1	704.1	489.2	1440.2	840.3	583.1	463.4	647.3
10	915.4	703.1	899.7	1285.2	770.5	1457.9	1169.1	965.8	614.1	785.6
11	1257.1	909.3	1003.0	497.5	692.7	1012.8	975.4	1080.4	472.8	562.7

Problem 6.4

Water in a cylindrical object will be ejected out if the water in the cylinder is above a hole in the side. We expect that the distance the water shoots out the side will increase as the depth of the water above the hole increases. In this experiment, the cylinder is a 2 liter plastic soda bottle with a hole in the side. We fill the bottle (with the hole covered) 30 times and measure the distance the water is ejected after the hole is uncovered. The runs are randomly assigned to 10 depths (6, 7, 8, ... 15 cm). Data are shown in the following table (from X. Meng, pers. comm., data set `WaterEjection`).

Ejection distances (cm) for 10 depths (cm)									
15	14	13	12	11	10	9	8	7	6
16.4	14.4	13.0	12.7	12.0	10.8	10.8	11.2	6.7	3.4
16.4	15.3	14.0	14.0	11.8	12.0	10.5	10.0	4.2	5.0
16.2	15.0	14.3	13.5	12.2	12.0	11.0	10.0	6.5	5.5

(a) Find a good polynomial fit for these data. Are there outliers or other problems? Does your model make physical sense (for example, what does it predict for the distance at a depth of 1 cm)?
(b) My (perhaps naive) intuition suggests that distance should be directly proportional to ejection velocity, and ejection velocity should be directly proportional to depth. Thus my intuition suggests that a model with a linear term and no intercept should fit the data. How well does this model work? Which model do you prefer?

Question 6.1 What happens to the t-statistic as one of the values becomes extremely large? Look at the data set consisting of the five numbers 0, 0, 0, 0, K, and compute the t-test for testing the null hypothesis that these numbers come from a population with mean 0. What happens to the t-statistic as K goes to infinity?

Question 6.2 Why would we expect the log transformation to be the variance-stabilizing transformation for the data in Exercise 6.2?

Chapter 7

Determining Sample Sizes

Earlier chapters have mostly dealt with analyzing experimental results. In this chapter we turn to design and consider the issues of choosing and assessing sample sizes. You will need to choose sample sizes regardless of how you analyze the data (that is, regardless of whether you use frequentist or Bayesian methods).

As we know, an experimental design is determined by the units, the treatments, and the assignment mechanism. Once we have chosen a pool of experimental units, decided which treatments to use, and settled on a completely randomized design, the major thing left to decide is the sample sizes for the various treatments. Choice of sample size is important because we want our experiment to be as small as possible to save time and money, but big enough to get the job done. What we need is a way to figure out how large an experiment needs to be to meet our goals; a bigger experiment would be wasteful, and a smaller experiment won't meet our needs. Unfortunately, far too many experiments are still sized by dividing available resources by the cost per unit.

Decide how large an experiment is needed

Sample size selection is presented as a problem of selecting the sample size so that a goal is achieved, or, more realistically, so that a goal is achieved with a sufficiently high probability. There are many potential goals, but they fall into two general categories: goals that deal with precision of estimation and goals that deal with (un)certainty of model selection. Precision goals are typically something like "the 95% confidence interval for estimating the difference in means between treatments 1 and 2 will be no longer than 1.5." For model comparison, the goal is typically stated in terms of *power*, that is, the probability of rejecting the null hypothesis when a certain alternative is true. There are also Bayesian analogues.

What is an appropriate level of power? There is no hard and fast rule. A power of 80% is a reasonable lower bound, and I would certainly not want to design for less than 70% power. On the other hand, there are substantial diminishing returns once power is more than 90 or 95%; you can add a lot of units without increasing power very much. Something in the range of 80% to 90% is the usual recommendation for design power.

Appropriate power

The width of a confidence interval depends on the desired coverage, the *error variance*, and the sample size, so we must know the error variance at least roughly before we can compute the required sample size. The power depends on $\mathcal{E}$, the sample sizes, the error variance, and the *actual values of the means under the alternative*. If we have no idea about the size of the error variance, then we cannot say how wide our intervals will be, and we cannot plan an appropriate sample size. If we have no idea what the means might be under the alternative, then we cannot do power analysis.

> The only way to do statistical sample size selection is by specifying and exploiting prior information about parameters in the models. This is true for frequentists as well as Bayesians.

Example 7.1 VOR in ataxia patients

Spinocerebellar ataxias (SCA's) are inherited, degenerative, neurological diseases. Clinical evidence suggests that eye movements and posture are affected by SCA. There are several distinct types of SCA's, and we would like to determine if the types differ in observable ways that could be used to classify patients and measure the progress of the disease. One response believed to be associated with SCA is the "amplitude of the vestibulo-ocular reflex for 20 deg/s^2 velocity ramps;" let's just call it VOR. VOR deals with how your eyes move when trying to focus on a fixed target while you are seated on a chair on a turntable that is rotating increasingly quickly.

We need to choose sample sizes to help us meet three goals regarding VOR in SCA types 1, 5, and 6: first, 95% confidence intervals for pairwise comparisons should be no longer that .5 on the log scale; second, we want power .9 when testing at the $\mathcal{E} = .01$ that these three SCA types have the same mean VOR; third, we want power .95 when testing at the .05 level that the mean response in SCA type 1 is the same as the average of the means of types 5 and 6.

We must specify the means and error variance to compute power, so we use those from the preliminary data. Note that there is only one subject in SCA 6, so our knowledge there is pretty slim and our computed sample sizes involving SCA 6 will not have a very firm foundation.

We have preliminary observations on a total of seventeen patients from SCA groups 1, 5, and 6, with sample sizes 5, 11, and 1. The response appears to have stable variance on the log scale, on which scale the group means of VOR are 2.82, 3.89, and 3.04, and the variance is .075. Thus it looks like the average response (on the original scale) in SCA 5 is about three times that of SCA 1, while the average response of SCA 6 is only about 25% higher than that of SCA 1.

7.1 Sample Size for Confidence Intervals

We can compute confidence intervals for means of treatment groups and contrasts between treatment groups. One sample size criterion is to choose the sample sizes so that confidence intervals of interest are no wider than a maximum allowable width W (margin of error no greater than $W/2$). For the mean of group i, a $1 - \mathcal{E}_I$ confidence interval has width

$$2\, t_{\mathcal{E}_I/2, N-g} \sqrt{\mathrm{MS_E}/n_i} \;\;;$$

for a contrast, the confidence interval has width

$$2\, t_{\mathcal{E}_I/2, N-g} \sqrt{\mathrm{MS_E}} \; \sqrt{\sum_i \frac{w_i^2}{n_i}} \;\;.$$

In principle, the required sample size can be found by equating either of these widths with W and solving for the sample sizes. In practice, we don't know $\mathrm{MS_E}$ until the experiment has been performed, so we must anticipate a reasonable value for $\mathrm{MS_E}$ when planning the experiment.

Assuming that we use equal sample sizes $n_i = n$, we find that

$$n \approx \frac{4\, t_{\mathcal{E}_I/2, g(n-1)}^2 \; \mathrm{MS_E} \sum w_i^2}{W^2} \;\;.$$

This is an approximation because n must be a whole number and the quantity on the right can have a fractional part; what we want is the smallest n such that the left-hand side is at least as big as the right-hand side. The sample size n appears in the degrees of freedom for t on the right-hand side, so we don't have a simple formula for n. We can compute a reasonable lower bound for n by substituting the upper $\mathcal{E}_I/2$ percent point of a normal for $t_{\mathcal{E}_I/2, g(n-1)}^2$. Then increase n from the lower bound until the criterion is met.

Example 7.2 VOR in ataxia patients, continued

Example 7.1 gave a requirement that 95% confidence intervals for pairwise differences should be no wider than .5. The preliminary data had an $\mathrm{MS_E}$ of .075, so that is a plausible value for future data. The starting approximation is then

$$n \approx \frac{4 \times 4 \times .075 \times (1^2 + (-1)^2)}{.5^2} = 9.6 \;\;,$$

so we round up to 10 and start there.

Rather than doing the computations by hand, let's use **R** and a function from `cfcdae`:

```
1 > sample.size.t.interval(.5,.075,2,3,.05)
  [1] 11
2 > sample.size.t.interval(.5,.15,2,3,.05)
  [1] 20
```

Line 1 shows the sample size computation when the width is set to .5, the error variance is .075, the sum of squared coefficients is 2, there are three groups, and the error rate is .05. The required sample size is 11, just one more than our starting approximation.

In this computation we used the error variance from existing data, which was .075 based on 14 degrees of freedom. As we will see in Chapter 10, there is a lot of variability in that estimate, and the true error variance could easily be twice as large. Line 2 shows the computation with a double error variance, yielding a required sample size of 20.

Now the experimenter faces a choice. Sample size 11 is probably big enough, but if she wants to be reasonably sure of meeting the design goal for however the MS_E may turn out in future data, she will need to use almost twice as many units.

Note from the example that doubling the assumed MS_E does not quite double the required sample size. This is because increasing the sample size also increases the degrees of freedom and thus reduces the percent point of t that we use. This effect is strongest for small sample sizes.

Sample size affects df and t-percent point

7.2 Power and Sample Size Analysis for ANOVA

While we prefer to use p-values for analysis, power analysis requires us to work with fixed-level tests for a precise alternative hypothesis. In a fixed level test, we either reject the null hypothesis at the pre-specified level, or we fail to reject the null hypothesis. If we reject a true null hypothesis, we have made a Type I error, and if we fail to reject a false null hypothesis, we have made a Type II error. The probability of making a Type I error is $\mathcal{E}_I$; $\mathcal{E}_I$ is completely under our control. We choose a Type I error rate $\mathcal{E}_I$ (5%, 1%, etc.), and reject H_0 if the p-value is less than $\mathcal{E}_I$. The probability of making a Type II error is $\mathcal{E}_{II}$; the probability of rejecting H_0 when H_0 is false is $1 - \mathcal{E}_{II}$ and is called *power*. The Type II error rate $\mathcal{E}_{II}$ depends on virtually everything: $\mathcal{E}_I$, g, σ^2, the α_i's, and n_i's. Most books use the symbols α and β for the Type I and II error rates. We use $\mathcal{E}$ for error rates, and use subscripts here to distinguish types of errors.

Power is probability of rejecting a false null hypothesis

It is more or less true that we can fix all but one of the interrelated parameters and solve for the missing one. For example, we may choose $\mathcal{E}_I$, g, σ^2, and the α_i's and n_i and then solve for $1 - \mathcal{E}_{II}$. This is called a power analysis, because we are determining the power of the experiment for the alternative specified by the particular α_i's. We may also choose $\mathcal{E}_I$, g, $1 - \mathcal{E}_{II}$, σ^2 and the α_i's and then solve for the sample sizes. This, of course, is called a sample size analysis, because we have specified a required power and now find a sample size that achieves that power. For example, consider a situation with three diets, and $\mathcal{E}_I$ is .05. How large should N be (assuming equal n_i's) to have a 90% chance of rejecting H_0 when σ^2 is 9 and the treatment mean responses are -7, -5, 3 (α_i's are -4, -2, and 6)?

Find minimum sample size that gives desired power

The use of power or sample size analysis begins by deciding on interesting values of the treatment effects and likely ranges for the error variance.

"Interesting" values of treatment effects could be determined in a variety of ways. For example, they could be anticipated effects, or they could be effects seen in a previous study, or they could be consensus scientific opinion, or they could be effects that are of a size to be practically important; in any case, we want to be able to detect interesting effects. For each combination of treatment effects, error variance, sample sizes, and Type I error rate, we may compute the power of the experiment. Sample size computation amounts to repeating this exercise again and again until we find the smallest sample sizes that give us at least as much power as required. Thus what we do is set up a set of circumstances that we would like to detect with a given probability, and then design for those circumstances.

Life is not always so straightforward. We could have several different precise alternatives for which we would like to have adequate power, and we do not know σ exactly. Thus we should perform a sensitivity analysis, trying various combinations of interesting alternatives and reasonable values for σ. This will give us a range for the sample sizes. You could design for the average situation, or you could design for the worst case scenario.

Sensitivity analysis

The ANOVA F-statistic is the ratio of the mean square for treatments to the mean square for error. When the null hypothesis is true, the F-statistic follows an F-distribution with degrees of freedom from the two mean squares. We reject the null when the observed F-statistic is larger than the upper $\mathcal{E}_I$ percent point of the F-distribution. When the null hypothesis is false, the F-statistic follows a *noncentral F-distribution*. Power, the probability of rejecting the null when the null is false, is the probability that the F-statistic (which follows a noncentral F-distribution when the alternative is true) exceeds a cutoff based on the usual (central) F distribution.

F-statistic follows noncentral F-distribution when null is false

This is illustrated in Figure 7.1. The solid line gives a typical null distribution for the F-test. The vertical line is at the 5% cutoff point; 5% of the area under the null curve is to the right, and 95% is to the left. This 5% is the Type I error rate, or $\mathcal{E}_I$. The dashed curve is the distribution of the F-ratio for one alternative. We would reject the null at the 5% level if our F-statistic is greater than the cutoff. The probability of this happening is the area under the alternative distribution curve to the right of the cutoff (the power); the area under the alternative curve to the left of the cutoff is the Type II error rate $\mathcal{E}_{II}$.

Power computed with noncentral F

The noncentral F-distribution has numerator and denominator degrees of freedom the same as the ordinary (central) F, and it also has a *noncentrality parameter* ζ defined by

Noncentrality parameter measures distance from null

$$\zeta = \frac{\sum_i n_i \alpha_i^2}{\sigma^2} \quad .$$

The α_is used in the formula for noncentrality satisfy $\sum n_i \alpha_i = 0$. Another way to think of the noncentrality parameter is to replace the data y_{ij} by their expected values μ_i and fit the null (single mean) model; the numerator of ζ will be the error sum of squares.

The noncentrality parameter measures how far the treatment means are from being equal (α_i^2) relative to the variation of $\overline{y}_{i\bullet}$ (σ^2/n_i). The ordinary
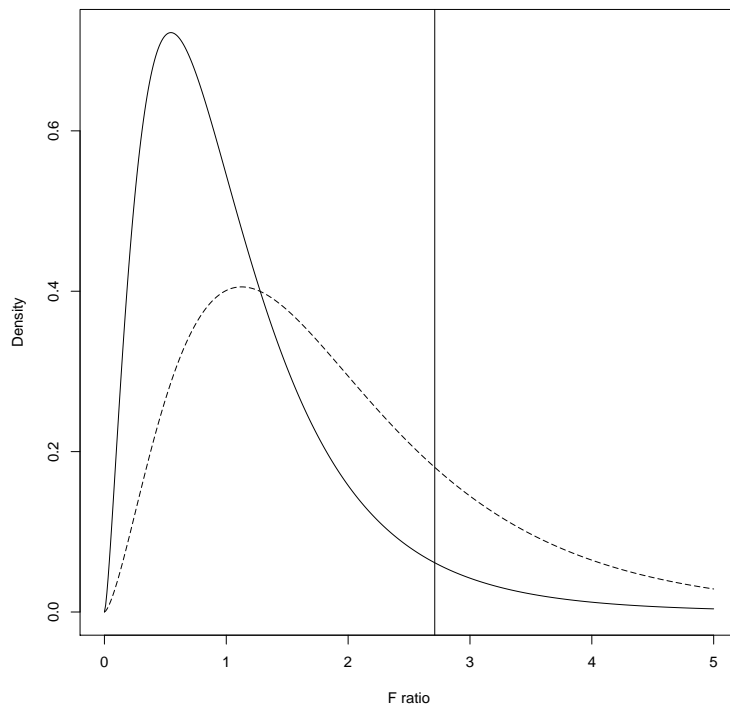
Figure 7.1: Null distribution (sold line) and alternative distribution (dashed line) for an F-test, with the 5% cutoff marked.

central F-distribution has $\zeta = 0$, and the bigger the value of ζ, the more likely we are to reject H_0.

Power is higher when $\mathcal{E}_I$ is larger, when ζ is larger (this can be either treatment means farther apart or smaller error variance), when error degrees of freedom are larger, or when numerator degrees of freedom are smaller (for a fixed value of ζ).

Example 7.3 VOR in ataxia patients, continued

We earlier expressed the goal of having power .9 when testing at the .01 level the null hypothesis that the three SCA groups all had the same mean VOR. This is an incomplete specification of the problem; we can only compute power when we have a specific set of alternative means (or the non-centrality parameter that corresponds to those means). Thus we will need to specify those alternative means before computing power or sample size.

Here are two potential specifications. Suppose that a difference of 50% (.4 on the log scale) is considered clinically relevant. If we design for sit-

uations where the maximum difference between two means is .4, then we will ensure that our sample sizes are big enough when the differences are .4 or larger. The mean triples (1, 1.4, 1.4) and (1, 1.2, 1.4) both represent scenarios where there is a difference of .4 on the log scale. We also need to specify the error variance. As with our confidence interval example, we use the previously estimated MS_E of .075 and also .15 (which is plausible given the variability in MS_E).

The function `sample.size.f.test` from `cfcdae` will do the necessary computations (there are many other sample size calculators available):

```
1 > sample.size.f.test(.9,alpha=.01,means=c(1,1.4,1.4),sigma2=.075)
  $nis
  [1] 14 14 14

  $power
  [1] 0.904201

2 > sample.size.f.test(.9,alpha=.01,means=c(1,1.2,1.4),sigma2=.075)
  $nis
  [1] 18 18 18

  $power
  [1] 0.901666

3 > sample.size.f.test(.9,alpha=.01,means=c(1,1.4,1.4),sigma2=.15)
  $nis
  [1] 27 27 27

  $power
  [1] 0.9127837

4 > sample.size.f.test(.9,alpha=.01,means=c(1,1.2,1.4),sigma2=.15)
  $nis
  [1] 35 35 35

  $power
  [1] 0.9080998

4 > sample.size.f.test(.9,alpha=.01,ncp1=((-.2)^2+0+(.2)^2)/.075,ngrps=3)
  $nis
  [1] 18 18 18

  $power
  [1] 0.901666
```

Lines 1 and 2 compute the sample size for the two configurations of means when the error variance is .075, and lines 3 and 4 do parallel computations when the error variance is .15. The results returned are the minimum sample sizes and the actual power with those sample sizes. When the required sample size is small, the actual power can be considerably greater than the requested power. As with the confidence interval example, doubling the proposed error variance nearly doubles the required sample sizes.

Note that line 2 needed a larger sample size to obtain power .9 than did line 1. In fact, the noncentrality parameter for line 1 is as large as it can be and still have the maximum difference between means equal to .4. There are

other configurations that also achieve this, for example, (1, 1, 1.4). On the other hand, the noncentrality parameter for line 2 is as small as it can be and still have the maximum difference between two means equal to 4. Thus these two configurations bookend the possibilities with a maximum difference of .4.

Line 5 shows an alternate set of arguments for `sample.size.f.test`. Instead of passing the means and the error variance, you may pass the number of groups (g) and the noncentrality parameter computed when $n = 1$.

Here is a useful trick for choosing sample size. Sometimes it is difficult to specify an interesting alternative completely; that is, we can't specify all the means or effects α_i, but we can say that any configuration of means that has two means that differ by an amount D or more would be interesting. The smallest possible value for the noncentrality parameter when this condition is met is $nD^2/(2\sigma^2)$, corresponding to two means D units apart and all the other means in the middle (with zero α_i's). If we design for this alternative, then we will have at least as much power for any other alternative with two treatments D units apart. This is essentially what we did in line 5 of the example.

Specify minimum difference

Some people talk about *observed power*; they shouldn't. Observed power is the power that you compute if you take the results of an experiment (treatment means $\overline{y}_{i\bullet}$, MS_E, sample sizes, and so on) and use those results to compute power. Observed power is a monotone function of the p-value for the ANOVA. Data sets with a small p-value will have high observed power, and those with a larger p-value will have low observed power. The observed power tells you nothing that the p-value did not already tell you.

Observed power

7.3 Power for a Contrast

The Analysis of Variance F-test is sensitive to all departures from the null hypothesis of equal treatment means. A contrast is sensitive to particular departures from the null. In some situations, we may be particularly interested in one or two contrasts, and less interested in other contrasts. In that case, we might wish to design our experiment so that the contrasts of particular interest had adequate power.

Suppose that we have a contrast with coefficients $\{w_i\}$. Test the null hypothesis that the contrast has expected value zero by using an F-test (the sum of squares for the contrast divided by the MS_E). The F-test has 1 and $N - g$ degrees of freedom and noncentrality parameter

Noncentrality parameter for a contrast

$$\frac{(\sum_{i=1}^{g} w_i \alpha_i)^2}{\sigma^2 \sum_{i=1}^{g} w_i^2/n_i} .$$

We now use software for 1 numerator degree of freedom to compute power.

Example 7.4 VOR in ataxia patients, continued

Suppose that we are particularly interested in comparing the VOR for SCA 1 to the average VOR for SCA 5 and 6 using a contrast with coefficients (1, -.5, -.5). We want to know the power for this contrast when testing at the .05 level when the means are actually (1, 1.4, 1.4) or (1, 1.2, 1.4) and the error variance is either .075 or .15.

In **R**, we cannot use the `sample.size.f.test` function, because it links numerator degrees of freedom, denominator degrees of freedom, and noncentrality parameter in a way that is incorrect for a contrast. Instead, we use the `power.f.test` function and set those quantities directly.

```
1 > ncp1a <- (1*1 - .5*1.4 - .5*1.4)^2/.075/(1+.5^2+.5^2)
2 > ncp1b <- (1*1 - .5*1.2 - .5*1.4)^2/.075/(1+.5^2+.5^2)
3 > n <- 5:20;power.f.test(ncp=n*ncp1a,df1=1,df2=3*(n-1),alpha=.05)
  [1] 0.6877270 0.7794260 0.8468854 0.8953215 0.9293914 0.9529386 0.9689671
  [8] 0.9797332 0.9868797 0.9915739 0.9946282 0.9965984 0.9978594 0.9986608
 [15] 0.9991667 0.9994841
4 > n <- 5:20;power.f.test(ncp=n*ncp1b,df1=1,df2=3*(n-1),alpha=.05)
  [1] 0.4521937 0.5357938 0.6099223 0.6747138 0.7306307 0.7783538 0.8186855
  [8] 0.8524744 0.8805618 0.9037462 0.9227620 0.9382684 0.9508460 0.9609982
 [15] 0.9691556 0.9756827
```

Lines 1 and 2 compute the noncentrality parameters for $n = 1$ for the two configurations of means. The actual noncentrality parameters will be n times these values. Lines 3 and 4 then compute the power for a range of sample sizes from 5 to 20.

We wanted power .95. We get that power for the sixth and thirteenth sample sizes, corresponding to sample sizes of 10 and 17. The sample size is smaller for the first configuration of means, because the average of the SCA 5 and 6 means is 1.4 (.4 from SCA 1) in that configuration, compared to 1.3 (.3 from SCA 1) in the second configuration. This leads to more power and a lower required sample size.

7.4   Sample Size for Bayesian Analysis

Bayesian analysis uses a prior distribution on parameters. Bayesian design, including sample size selection, also uses a prior distribution on parameters. The first thing to note is that the design prior and analysis prior need not be the same. In fact, you probably want them to be different. We often use weakly informative priors when doing analysis. The advantage of a weakly informative prior is that it is more palatable to a broader range of researchers, so the results stemming from that prior are also likely to be broadly acceptable. On the other hand, we want the best information available when designing the experiment. Such a prior will be much tighter than the weakly informative analysis prior, but it is almost surely much broader than the single point priors (that is, precise, completely specified alternatives) used by frequentists when doing sample size analysis.

Design prior and analysis prior

Bayesian sample size analysis involves choosing sample sizes so that we either achieve a goal on average, or the probability of achieving a goal is

acceptably high. This probability is computed on the basis of three distributions: the prior used in the design of the experiment, the prior used in analyzing the experiment, and the distribution of yet-to-be observed data given the parameter values (the likelihood). In contrast, frequentists only deal with the likelihood. This multi-layering of distributions means that doing Bayesian sample size analysis is more computationally intensive than the corresponding frequentist exercise.

As with the frequentist situation, the goals generally revolve around precision or model selection. Unlike the frequentist setting, however, there are several ways to specify precision or model selection goals.

7.4.1 Precision

A full Bayesian, decision-theoretic selection of sample size for estimation requires the statistician to make a decision $(n, \omega(x_1, \ldots, x_n))$, consisting of a sample size n and a method ω for estimating the quantity of interest. There is also a loss associated with any decision. For example, we might assume that the loss is $c_1(\omega(x_1, \ldots, x_n) - \mu)^2 + c_2 n$. This says that we lose c_1 times our squared error of estimation (the cost of not estimating μ exactly) plus c_2 times the sample size (the cost of collecting the data). Based on the likelihood and the prior distributions, the full decision-theoretic approach says to choose the sample size n and estimator ω that minimize the expected loss. (If we use squared error loss as shown above, the estimator will be the posterior expected value.) In many cases, the expected loss is *roughly* $c_1 V/n + c_2 n$, leading to a sample size of approximately $n = \sqrt{c_1 V/c_2}$. The cost (loss) per unit c_2 is often fairly obvious, but the cost of estimation error is often very difficult to specify, making the decision theoretic approach difficult to implement.

Full Bayesian solution minimizes expected costs

In the absence of a fully specified loss function, consider choosing sample size to meet a precision goal analogous to those we used in the frequentist formulation. In the very simplest of Bayesian settings, one can compute the precision of an estimate analytically. For example, suppose we have a prior that $\mu \sim N(0, \sigma_\mu)$, and we observe data $y_i, i = 1, ..., n$ that are $N(\mu, \sigma_y)$. In that case, the posterior is normal, and the posterior standard deviation of μ is

$$\sqrt{\frac{1}{\frac{n}{\sigma_y^2} + \frac{1}{\sigma_\mu^2}}}$$

With σ_μ and σ_y specified, we simply compute twice the normal percent point times this posterior standard deviation and solve for the n that gives us a credible interval that meets our goal.

In the usual situation where the standard deviations are not known, we need a procedure that will work on average across all the potential values of both standard deviations and data. In general, this is an intractable problem, but Joseph and Bélisle (1997) discuss some solutions based on work of Adcock (1988). For a single normal mean, Adcock proposes the following likelihood and priors. Assume that conditional on μ and σ_ϵ, the data are

Adcock's priors

independent and follow a normal distribution with that mean and standard deviation. Assume that $1/\sigma_\epsilon^2$ is distributed as $1/(\nu\sigma_0^2)$ times a chi-square distribution with ν degrees of freedom. Finally, assume that the prior for μ is normal with mean μ_0 and variance σ_ϵ^2/n_0. Note first that the prior variability for μ is proportional to that of the data in the likelihood. This seems unlikely in practice, but nevertheless, Adcock's method works fairly well. Note second that the assumptions on the prior can be interpreted as the prior on μ provides n_0 units worth of information.

Suppose that we choose to use an interval of length ℓ. Because the standard deviations vary from data set to data set, the best interval of length ℓ will have coverage that varies from data set to data set. The Average Coverage Criterion says to choose n large enough that the average coverage (averaging over potential values of σ and data) meets our coverage goal.

Using Adcock's assumptions, if we want an interval of length W with average coverage $1 - \mathcal{E}$, the sample size should be:

$$n \geq \frac{4\,\sigma_0^2\,t_{\mathcal{E}/2,\nu}^2}{W^2} - n_0$$

If we have a separate-means problem and we assume that sample sizes are the same and all means have the same prior information (same n_0), then for a contrast the average coverage criterion is

<div align="right">Adcock's method
for Average
Coverage</div>

$$n \geq \frac{4\,\sigma_0^2\,\sum w_i^2\,t_{\mathcal{E}/2,\nu}^2}{W^2} - n_0$$

Note that the t cutoff does not depend on n. This will lead to greater required n than we saw for the frequentist approach if prior information is low (that is, if n_0 and/or ν are small), but it can lead to smaller required n if prior information is high (n_0 and/or ν large).

There are other criteria besides average coverage. For example, the Average Length Criterion works in the opposite direction. It says to always choose an interval with the desired coverage. These intervals will have varying lengths depending on the data, so we choose the sample size so that the average length meets our goal. The average length criterion leads to a considerably more complex sample size calculation.

Example 7.5 VOR in ataxia patients, continued

Let's use the average coverage criterion to select a sample size for the pairwise differences in the VOR situation, where we want a width of .5 and coverage of 95%. Our prior information is that the $\mathrm{MS_E}$ of pilot data was .075 with 14 degrees of freedom.

```
1 > sample.size.bayes.interval(.5,.075,14,1,2)
  [1] 11
2 > sample.size.bayes.interval(.5,.075,14,3,2)
  [1] 9
3 > sample.size.bayes.interval(.5,.075,24,1,2)
  [1] 10
4 > sample.size.bayes.interval(.5,.075,4,1,2)
  [1] 18
```

Line 1 gets the Bayesian sample size determination based on a needed width of .5, an expected error variance of .075, 14 degrees of freedom in our variance prior, 1 units worth of prior information on the contrast mean, a sum of squared contrast coefficients of 2, and a default error rate of .05. The calculated sample size is 11, which is the same as we obtained for a confidence interval. This 11, however, is based on averaging across some smaller variances and some larger variances, which lead to smaller and larger required sample sizes.

Line 2 shows the trivial effect that having more prior information about the contrast mean: every additional unit's worth of information we have from the prior is one less we need in the posterior. Lines 3 and 4 illustrate the effect of having more or less information (degrees of freedom) in our estimate of the error variance.

One can still determine the sample size needed to reach a goal of specified average coverage at a given length, even when the specific assumptions of the Adcock method are not true, but it is more complicated and requires simulation. That is, the process is slow and not precise like a formula (making it more precise makes it slower). Basically, we estimate the average coverage for an interval width and a set of sample sizes. Then we keep trying different sample sizes until we find the smallest one that meets our criterion. Each determination of coverage for a given sample size involves some number M of simulation runs. For each of the M runs:

Simulation-based coverage estimates

- Randomly select parameter values from the (design) prior;

- Randomly select data distributed according to these parameter values;

- Fit the model to the data using the (analysis) prior, obtaining a sample from the posterior distribution of the quantity of interest;

- Determine the coverage of an interval of the specified length.

Average the M sample coverages to estimate the average coverage for the sample size.

This simulation approach works in models that meet our typical assumptions (albeit slowly), but it also works for non-normally distributed data, data with non-constant error structures, correlated data, and so on. We simply need to be able to generate data from the model of interest, fit the model of interest, and tolerate the time it takes to do the repeated simulations.

Simulation works broadly

The simulation approach can be slow, but you may be spending a lot of time and money running your experiment, so it is worth some time investment up front to get the sample sizes correct.

7.4.2 Model Selection

A full Bayesian, decision-theoretic selection of sample size for model selection requires the statistician to make a decision $(n, \omega(x_1, \ldots, x_n))$, consisting of a sample size n and a method ω for selecting the model based on data and the priors. There is a loss associated with any decision. For example, we might assume a loss c_1 when we select model 1 when model 2 is correct, a loss c_2 when we select model 2 when model 1 is correct, and a loss $c_3 n$ for the cost of collecting data. Based on the likelihood and the prior distributions, the full decision-theoretic approach says to choose the sample size n and method ω that minimize the expected loss. (The method ω will be selection via Bayes factor with the cutoff determined by the relative costs and prior probabilities of the models.) This can also be adapted to optional collection of additional data: at any sample size n, collect another data point if the expected loss with an additional data point is less than the expected loss with the current sample size.

To implement the decision-theoretic approach, you need to be able to specify the three costs, the (design) priors for the parameters, and the prior probabilities for the models. This full specification is particularly valuable in high-stakes situations such as clinical trials, where the loss function could be quite complex and account for many factors, for example, disease prevalence, where you need to be more certain about decisions that will affect more people; see Berry (2006). However, we will not always have a complete specification of the needed quantities (especially the losses), so less form approaches are also needed.

The Bayes factor is the standard Bayesian tool for selecting between two models for the data, so selecting a sample size based on its expected behavior under different models has been the primary approach to sample size selection. For example, when sampling from the prior distribution under the alternative (model 2), we might select a sample size so that the Bayes factor BF_{21} will be 3 or greater with probability 80%. In general, we select the sample size so that when sampling under the alternative we have probability $1 - \mathcal{E}_{II}$ of having BF_{21} be at least K.

As we have seen before, an analytical solution to the Bayesian problem is only available in the simplest models, so we will need to use simulation to determine the sample size. Simulation is slow, but it is adaptable to all kinds of models. In general, the approach is to determine the probability that the Bayes factor meets our goal for a given sample size, and then vary the sample size until we find the one that is just large enough to meet the goal with the required probability. Each probability determination is done via simulation. For each of M iterations we:

- Randomly select parameter values from the (design) prior for the alternative;

- Randomly select data distributed according to these parameter values;

- Fit the model to the data using the (analysis) prior, obtaining the Bayes factor;

Estimate the probability as the fraction of simulated Bayes factors that exceed K.

Example 7.6 VOR in ataxia patients, continued

Let's simulate the distribution of the Bayes factor to determine a sample size that gives us probability .95 that the Bayes factor will be 3 or greater using criteria analogous to Example 7.4. We have three groups, we have an existing $MS_E = .075$ based on 14 degrees of freedom, and we would like to determine the sample size for a configuration of means (0,.4,.4). Assume that y_{ij} is $N(\mu_i, \sigma_0)$, μ_i is $N(\mu_{0i}, \sigma_\mu)$, and we can produce samples from the distributions for σ_0 and σ_μ.

```
1 > BFfrac <- function(K,nis,mu0,sigma.mu,sigma0,M=1000) {
      count <- 0
      g <- length(nis)
      x <- factor(rep(1:g,nis))
      N <- sum(nis)
      for(i in 1:M) {
        mus <- rnorm(g,mu0,sigma.mu[i])
        y <- rnorm(N,rep(mus,nis),sigma0[i])
        fit <- lmBF(y~x,data.frame(x,y),progress=FALSE)
        if(extractBF(fit)[1] > K) count <- count+1
      }
      count/M
    }
2 > BFfrac(3,rep(10,3),c(0,.4,.4),rep(.0001,1000),rep(.075^.5,1000),1000)
  [1] 0.788
3 > BFfrac(3,rep(15,3),c(0,.4,.4),rep(.0001,1000),rep(.075^.5,1000),1000)
  [1] 0.961
4 > BFfrac(3,rep(14,3),c(0,.4,.4),rep(.0001,1000),rep(.075^.5,1000),1000)
  [1] 0.95
5 > BFfrac(3,rep(14,3),c(0,.4,.4),rep(.0001,1000),sqrt(.075*14/rchisq(1000,14)),1000)
  [1] 0.88
6 > BFfrac(3,rep(18,3),c(0,.4,.4),rep(.0001,1000),sqrt(.075*14/rchisq(1000,14)),1000)
  [1] 0.945
7 > BFfrac(3,rep(18,3),c(0,.4,.4),rep(.1,1000),sqrt(.075*14/rchisq(1000,14)),1000)
  [1] 0.894
8 > BFfrac(3,rep(20,3),c(0,.4,.4),rep(.1,1000),sqrt(.075*14/rchisq(1000,14)),1000)
  [1] 0.907
9 > BFfrac(3,rep(22,3),c(0,.4,.4),rep(.1,1000),sqrt(.075*14/rchisq(1000,14)),1000)
  [1] 0.922
10 > BFfrac(3,rep(25,3),c(0,.4,.4),rep(.1,1000),sqrt(.075*14/rchisq(1000,14)),1000)
  [1] 0.935
11 > BFfrac(3,rep(28,3),c(0,.4,.4),rep(.1,1000),sqrt(.075*14/rchisq(1000,14)),1000)
  [1] 0.953
```

Line 1 creates a function to simulate the rate at which the Bayes factor achieves a goal. The arguments of the function are the Bayes factor goal K, the sample sizes in the treatment groups, the prior means μ_0, samples from the distribution of σ_μ, samples from the distribution of σ_0, and the number of iterations in the simulation.

Line 2 represents a situation very similar to the frequentist case. We use a tiny σ_μ, meaning that μ is essentially known and equal to its prior mean, and

we assume that σ_0 is known. We set the prior means equal to $(0,.4,.4)$ and use the sample size that gave us power .95 power. We see that the probability that the Bayes factor is greater than 3 is less than .8; $BF_{21} > 3$ is a more stringent condition than $p < .05$. Lines 3 and 4 find a sample size that yields a probability of at least .95.

Line 5 makes things more realistic by assuming that we do not know σ_0 exactly. Now we take samples from a prior distribution for σ_0 that matches our knowledge from the previous data anlysis. This gives us σ_0 values ranging from about .2 to about .45 ($\sqrt{.075} = .274$). Line 6 shows that we need to increase the sample size to (roughly) 18 to get the probability up to .95.

Finally, lines 7–11 consider the situation where we do not know the mean exactly and instead let them vary around μ_0 with a standard deviation of .1. Now we need a sample size of 28 to get a probability of .95. What is happening is that when the configuration of means selects means that are farther apart than suggested by μ_0, then the probability is increased, but not by a lot. Conversely, when the means are closer together than suggested by μ_0, then the probability is decreased, and possibly by a lot. Thus we need a larger sample size to counter the instances where the means are more similar that suggested by μ_0.

If we had used the frequentist sample size calculation when our prior information is really more like what is represented in this example, we would have produced an experiment that did not meet our power requirements.

7.5 More about Units and Measurement Units

Thinking about sample size, cost, and power brings us back to some issues involved in choosing experimental units and measurement units. The basic problems are those of dividing fixed resources (there is never enough money, time, material, etc.) and trying to get the most bang for the buck.

Consider first the situation where there is a fixed amount of experimental material that can be divided into experimental units. In agronomy, the limited resource might be an agricultural field of a fixed size. In textiles, the limited resource might be a bolt of cloth of fixed size. The problem is choosing into how many units the field or bolt should be divided. Larger units have the advantage that their responses tend to have smaller variance, since these responses are computed from more material. Their disadvantage is that you end up with fewer units to average across. Smaller units have the opposite properties; there are more of them, but they have higher variance.

Subdividing spatial units

There is usually some positive spatial association between neighboring areas of experimental material. Because of that, the variance of the average of k adjacent spatial units is greater than the variance of the average of k randomly chosen units. (How much greater is very experiment specific.) This greater variance for contiguous blocks implies that randomizing treatments across more little units will lead to smaller variances for treatment averages and comparisons than using fewer big units.

More little units generally better

There are limits to this splitting, of course. For example, there may be an expensive or time-consuming analytical measurement that must be made on each unit. An upper bound on time or cost thus limits the number of units that can be considered. A second limit comes from edge guard wastage. When units are treated and analyzed *in situ* rather then being physically separated, it is common to exclude from analysis the edge of each unit. This is done because treatments may spill over and have effects on neighboring units; excluding the edge reduces this spillover. The limit arises because as the units become smaller and smaller, more and more of the unit becomes edge, and we eventually we have little analyzable center left.

A second situation occurs when we have experimental units and measurement units. Are we better off taking more measurements on fewer units or fewer measurement on more units? In general, we have more power and shorter confidence intervals if we take fewer measurements on more units. However, this approach may have a higher cost per unit of information.

<div style="float:right">More units or measurement units?</div>

For example, consider an experiment where we wish to study the possible effects of heated animal pens on winter weight gain. Each animal will be a measurement unit, and each pen is an experimental unit. We have g treatments with n pens per treatment ($N = gn$ total pens) and r animals per pen. The cost of the experiment might well be represented as $C_1 + gnC_2 + gnrC_3$. That is, there is a fixed cost, a cost per pen, and a cost per animal. The cost per pen is no doubt very high. Let σ_1^2 be the variation from pen to pen, and let σ_2^2 be the variation from animal to animal. Then the variance of a treatment average is

<div style="float:right">Costs may vary by unit type</div>

$$\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{nr}.$$

The question is now, "What values of n and r give us minimal variance of a treatment average for fixed total cost?" We need to know a great deal about the costs and sources of variation before we can complete the exercise.

7.6 Allocation of Units for Two Special Cases

We have considered computing power and sample size for balanced allocations of units to treatments. Indeed, Chapter 6 gave some compelling reasons for favoring balanced designs. However, there are some situations where unequal sample sizes could increase the power for alternatives of interest. We examine two of these.

Suppose that one of the g treatments is a control treatment, say treatment 1, and we are only interested in determining whether the other treatments differ from treatment 1. That is, we wish to compare treatment 2 to control, treatment 3 to control, ..., treatment g to control, but we don't compare noncontrol treatments. This is the standard setup where Dunnett's test is applied. For such an experiment, the control plays a special role (it appears in all contrasts), so it makes sense that we should estimate the control response more precisely by putting more units on the control. In fact, we can show that

<div style="float:right">Comparison with control</div>

we should choose group sizes so that the noncontrol treatments sizes (n_t) are equal and the control treatment size (n_c) is about $n_c = n_t\sqrt{g-1}$.

A second special case occurs when the g treatments correspond to numerical levels or doses. For example, the treatments could correspond to four different temperatures of a reaction vessel, and we can view the differences in responses at the four treatments as linear, quadratic, and cubic temperature effects. If one of these effects is of particular interest, we can allocate units to treatments in such a way to make the standard error for that selected effect small.

Suppose that we believe that the temperature effect, if it is nonzero, is essentially linear with only small nonlinearities. Thus we would be most interested in estimating the linear effect and less interested in estimating the quadratic and cubic effects. In such a situation, we could put more units at the lowest and highest temperatures, thereby decreasing the variance for the linear effect contrast. We would still need to keep some observations in the intermediate groups to estimate quadratic and cubic effects, though we wouldn't need as many as in the high and low groups since determining curvature is assumed to be of less importance than determining the presence of a linear effect.

Note that we need to exercise some caution. If our assumptions about shape of the response and importance of different contrasts are incorrect, we could wind up with an experiment that is much less informative than the equal sample size design. For example, suppose we are near the peak of a quadratic response instead of on an essentially linear response. Then the linear contrast (on which we spent all our units to lower its variance) is estimating zero, and the quadratic contrast, which in this case is the one with all the interesting information, has a high variance.

7.7   Further Reading and Extensions

When the null hypothesis is true, the treatment and error sums of squares are distributed as σ^2 times chi-square distributions. Mathematically, the ratio of two independent chi-squares, each divided by their degrees of freedom, has an F-distribution; thus the F-ratio has an F-distribution when the null is true. When the null hypothesis is false, the error sum of squares still has its chi-square distribution, but the treatment sum of squares has a *noncentral chi-square* distribution. Here we briefly describe the noncentral chi-square.

If $Z_1, Z_2, \cdots, Z_n$ are independent normal random variables with mean 0 and variance 1, then $Z_1^2 + Z_2^2 + \cdots + Z_n^2$ (a sum of squares) has a chi-square distribution with n degrees of freedom, denoted by χ_n^2. If the Z_i's have variance σ^2, then their sum of squares is distributed as σ^2 times a χ_n^2. Now suppose that the Z_i's are independent with means δ_i and variance σ^2. Then the sum of squares $Z_1^2 + Z_2^2 + \cdots + Z_n^2$ has a distribution which is σ^2 times a *noncentral* chi-square distribution with n degrees of freedom and noncentrality parameter $\sum_{i=1}^{n} \delta_i^2 / \sigma^2$. Let $\chi_n^2(\zeta)$ denote a noncentral chi-square with n

degrees of freedom and noncentrality parameter ζ. If the noncentrality parameter is zero, we just have an ordinary chi-square.

In Analysis of Variance, the treatment sum of squares has a distribution that is σ^2 times a noncentral chi-square distribution with $g - 1$ degrees of freedom and noncentrality parameter $\sum_{i=1}^{g} n_i \alpha_i^2 / \sigma^2$. See Appendix A. The mean square for treatments thus has a distribution

$$\text{MS}_{\text{trt}} \sim \frac{\sigma^2}{g-1} \chi^2_{g-1} \left(\frac{\sum_{i=1}^{g} n_i \alpha_i^2}{\sigma^2} \right) \quad .$$

The expected value of a noncentral chi-square is the sum of its degrees of freedom and noncentrality parameter, so the expected value of the mean square for treatments is $\sigma^2 + \sum_{i=1}^{g} n_i \alpha_i^2 / (g - 1)$. When the null is false, the F-ratio is a noncentral chi-square divided by a central chi-square (each divided by its degrees of freedom); this is a noncentral F-distribution, with the noncentrality of the F coming from the noncentrality of the numerator chi-square.

7.8   Problems

Find the smallest sample size giving power of at least .7 when testing equality of six groups at the .05 level when $\zeta = 4n$. **Exercise 7.1**

We are planning an experiment comparing three fertilizers. We will have six experimental units per fertilizer and will do our test at the 5% level. One of the fertilizers is the standard and the other two are new; the standard fertilizer has an average yield of 10, and we would like to be able to detect the situation when the new fertilizers have average yield 11 each. We expect the error variance to be about 4. What sample size would we need if we want power .9? **Exercise 7.2**

What is the probability of rejecting the null hypothesis when there are four groups, the sum of the squared treatment effects is 6, the error variance is 3, the group sample sizes are 4, and $\mathcal{E}$ is .01? **Exercise 7.3**

I conduct an experiment doing fixed-level testing with $\mathcal{E} = .05$; I know that for a given set of alternatives my power will be .85. True or False? **Exercise 7.4**

1. The probability of rejecting the null hypothesis when the null hypothesis is false is .15.

2. The probability of failing to reject the null hypothesis when the null hypothesis is true is .05.

Consider two experiments. The first has two treatments with four units in each and a noncentrality parameter of 20. The second has six treatments with two units in each and also has a noncentrality parameter of 20. Which experiment has greater power? Explain your answer. **Exercise 7.5**

(a) We are considering an experiment with $n = 20$ units, with 5 allocated to each of $g = 4$ treatments. When considering power, we are planning for treatment means of $7, 9, 11, 13$ and $\sigma^2 = 10$. Give the noncentrality parameter and degrees of freedom for this test.
(b) Suppose that we consider a different experiment, one with 4 units for each of $g = 5$ treatments ($n = 20$ again) and we are considering power for the treatment means $7, 9, 10, 11, 13$ and $\sigma^2 = 10$. Give the noncentrality parameter and degrees of freedom for this test
(c) Which test is more powerful? **Exercise 7.6**

We can run an experiment with three treatments, four units per treatment, and error variance 2; or we can run an experiment with three treatments, two units per treatment, and error variance 1. (The extra expense of the less variable method made us reduce the sample size.) We should have the same treatment effects in both cases. Which experiment has more power, and why? **Exercise 7.7**

We are interested in the effects of soy additives to diets on the blood concentration of estradiol in premenopausal women. We have historical data on six subjects, each of whose estradiol concentration was measured at the same stage of the menstrual cycle over two consecutive cycles. On the log scale, **Problem 7.1**

the error variance is about .109. In our experiment, we will have a pretreatment measurement, followed by a treatment, followed by a posttreatment measurement. Our response is the difference (post − pre), so the variance of our response should be about .218. Half the women will receive the soy treatment, and the other half will receive a control treatment.

How large should the sample size be if we want power .9 when testing at the .05 level for the alternative that the soy treatment raises the estradiol concentration 25% (about .22 log units)?

Problem 7.2 Nondigestible carbohydrates can be used in diet foods, but they may have effects on colonic hydrogen production in humans. We want to test to see if inulin, fructooligosaccharide, and lactulose are equivalent in their hydrogen production. Preliminary data suggest that the treatment means could be about 45, 32, and 60 respectively, with the error variance conservatively estimated at 35. How many subjects do we need to have power .95 for this situation when testing at the $\mathcal{E}_I = .01$ level?

Problem 7.3 Consider the situation of Exercise 3.5. The data we have appear to depend linearly on delay with no quadratic component. Suppose that the true expected value for the contrast with coefficients (1,-2,1) is 1 (representing a slight amount of curvature) and that the error variance is 60. What sample size would be needed to have power .9 when testing at the .01 level?

Problem 7.4 Suppose that I have planned an experiment with three treatments, 20 units per treatment, and anticipated error standard deviation $\sigma = 10$. Will my power increase more if I spend money to double my sample size (to 40 units per treatment), or spend money to halve my σ to 5? Explain your answer.

Problem 7.5 There are three treatments with treatment effects −1, 0, and 1. For a particular error variance of σ^2, we need 20 units per treatment to achieve power .8 when testing at the .05 significance level.

(a) Approximately how many units will we need per treatment to get power .8 when testing at the .05 level if we instead assume that the treatment effects are −0.5, 0 and 0.5?

(b) Approximately how many units will we need per treatment to get power .8 when testing at the .05 level if the treatment effects are –1, 0, and 1 but the error variance is $2\sigma^2$?

You are designing a great experiment for your boss. It has four treatments and an anticipated error variance of 3. Your boss gives you an alternative mean scenario, and you work out that you need 10 units per treatment to achieve 90% power when testing at error rate 0.01. You're all set to send the instructions off to the technicians when you boss rushes in and confesses, "Oops, 3 was the anticipated error standard deviation, not the variance. The anticipated error variance is 9." Approximately how many units are you going to need per treatment for the same alternative to achieve 90% power when testing at 0.01. Explain your answer.

Problem 7.6

Chapter 8

Factorial Treatment Structure

We have been working with completely randomized designs, where g treatments are assigned at random to N units. Up till now, the treatments have had no structure; they were just g treatments. *Factorial treatment structure* exists when the g treatments are the combinations of the levels of two or more factors. We call these combination treatments *factor-level combinations* or *factorial combinations* to emphasize that each treatment is a combination of one level of each of the factors. We have not changed the randomization; we still have a completely randomized design. It is just that now we are considering treatments that have a factorial structure. We will learn that there are compelling reasons for preferring a factorial experiment to a sequence of experiments investigating the factors separately.

Factorials combine the levels of two or more factors to create treatments

8.1 Factorial Structure

It is best to start with some examples of factorial treatment structure.

- Nelson, Kriby, and Johnson (1990) studied the effects of six dietary supplements on the occurrence of leg abnormalities in young chickens. The six treatments were the combinations of two levels of phosphorus supplement and three levels of calcium supplement. The phosphorus supplement alone is not a treatment, and neither is the calcium supplement. The combination of phosphorus and calcium supplements is a treatment in this design.

- Ellering (pers. comm.) studied corrosion-prevention coatings for metal guitar strings, examining how extraction rate (which is another way of looking at application time), curing time, and curing temperature affect the retention of the coating after a salt water bath. There are two levels of rate (6 or 12 inches/minute), four levels of curing temperature (20, 60, 90, or 120 degrees C), and two levels of curing time (30 or 60 minutes), for a total of 16 treatments applied to 48 steel samples.

Table 8.1: Number of sprouting barley seeds out of 100, by water used for springing and age of seeds. Data from Hareland and Madson (1989); data set `SproutingBarley`.

ml H_2O	Age of Seeds (weeks)				
	1	3	6	9	12
	11	7	9	13	20
4	9	16	19	35	37
	6	17	35	28	45
	8	1	5	1	11
8	3	7	9	10	15
	3	3	9	9	25

- Van de Ven (pers. comm.) studied the effect of thickness and surface treatment on the transmission of laser light through a piece of clear PVC (polyvinyl chloride). The thickness factor had six levels, and the surface treatment had three levels (no sanding, sand the front, sand the front and back). Combined, there are 18 treatments.

- Finally, Hunt and Larson (1990) studied the effects of sixteen treatments on zinc retention in the bodies of rats. The treatments were the combinations of two levels of zinc in the usual diet, two levels of zinc in the final meal, and four levels of protein in the final meal. Again, it is the combination of factor levels that makes a factorial treatment.

We begin our study of factorial treatment structure by looking at two-factor designs. We may present the responses of a two-way factorial as a table with rows corresponding to the levels of one factor (which we call factor A) and columns corresponding to the levels of the second factor (factor B). For example, Table 8.1 shows the results of an experiment on sprouting barley. Barley seeds are divided into 30 lots of 100 seeds each. The 30 lots are divided at random into ten groups of three lots each, with each group receiving a different treatment. The ten treatments are the factorial combinations of amount of water used for sprouting (factor A) with two levels, and age of the seeds (factor B) with five levels. The response measured is the number of seeds sprouting.

Two-factor designs

We use the notation y_{ijk} to indicate responses in the two-way factorial. In this notation, y_{ijk} is the kth response in the treatment formed from the ith level of factor A and the jth level of factor B. Thus in Table 8.1, $y_{2,5,3} = 25$. For a four by three factorial design (factor A has four levels, factor B has three levels), we could tabulate the responses as in Table 8.2. This table is just a convenient representation that emphasizes the factorial structure; treatments were still assigned to units at random.

Multiple subscripts denote factor levels and replication

Notice in both Tables 8.1 and 8.2 that we have the same number of responses in every factor-level combination. This is called *balance*. Balance turns out to be important for the standard analysis of factorial responses.

Balanced data have equal replication

Table 8.2: A two-way factorial treatment structure.

	B1	B2	B3
A1	y_{111} $\vdots$ y_{11n}	y_{121} $\vdots$ y_{12n}	y_{131} $\vdots$ y_{13n}
A2	y_{211} $\vdots$ y_{21n}	y_{221} $\vdots$ y_{22n}	y_{231} $\vdots$ y_{23n}
A3	y_{311} $\vdots$ y_{31n}	y_{321} $\vdots$ y_{32n}	y_{331} $\vdots$ y_{33n}
A4	y_{411} $\vdots$ y_{41n}	y_{421} $\vdots$ y_{42n}	y_{431} $\vdots$ y_{43n}

> We will assume for now that our data are balanced with n responses in every factor-level combination.

Chapter 9 will consider analysis of unbalanced factorials.

8.2 Factorial Analysis: Main Effect and Interaction

When our treatments have a factorial structure, we may also use a factorial analysis of the data. The major concepts of this factorial analysis are main effect and interaction.

Consider a two-way factorial where factor A has four levels and factor B has three levels, as in Table 8.2. There are $g = 12$ treatments, with 11 degrees of freedom between the treatments. We use i and j to index the levels of factors A and B. The expected values in the twelve treatments may be denoted μ_{ij}, coefficients for a contrast in the twelve means may be denoted w_{ij} (where as usual $\sum_{ij} w_{ij} = 0$), and the contrast sum is $\sum_{ij} w_{ij}\mu_{ij}$. Similarly, $\overline{y}_{ij\bullet}$ is the observed mean in the ij treatment group, and $\overline{y}_{i\bullet\bullet}$ and $\overline{y}_{\bullet j\bullet}$ are the observed means for all responses having level i of factor A or level j of B, respectively. It is often convenient to visualize the expected values, means, and contrast coefficients in matrix form, as in Table 8.3. *(Treatment, row, and column means)*

For the moment, forget about factor B and consider the experiment to be a completely randomized design just in factor A (it *is* completely randomized in factor A). Analyzing this design with four "treatments," we may compute a sum of squares with 3 degrees of freedom. The variation summarized by this sum of squares is denoted SS_A and depends on just the level of factor A. The expected value for the mean of the responses in row i is $\mu + \alpha_i$, where we assume that $\sum_i \alpha_i = 0$. *(Factor A ignoring factor B)*

Table 8.3: Matrix arrangement of (a) expected values, (b) means, and (c) contrast coefficients in a four by three factorial.

(a)		
μ_{11}	μ_{12}	μ_{13}
μ_{21}	μ_{22}	μ_{23}
μ_{31}	μ_{32}	μ_{33}
μ_{41}	μ_{42}	μ_{43}

(b)		
$\bar{y}_{11\bullet}$	$\bar{y}_{12\bullet}$	$\bar{y}_{13\bullet}$
$\bar{y}_{21\bullet}$	$\bar{y}_{22\bullet}$	$\bar{y}_{23\bullet}$
$\bar{y}_{31\bullet}$	$\bar{y}_{32\bullet}$	$\bar{y}_{33\bullet}$
$\bar{y}_{41\bullet}$	$\bar{y}_{42\bullet}$	$\bar{y}_{43\bullet}$

(c)		
w_{11}	w_{12}	w_{13}
w_{21}	w_{22}	w_{23}
w_{31}	w_{32}	w_{33}
w_{41}	w_{42}	w_{43}

Now, reverse the roles of A and B. Ignore factor A and consider the experiment to be a completely randomized design in factor B. We have an experiment with three "treatments" and treatment sum of squares SS_B with 2 degrees of freedom. The expected value for the mean of the responses in column j is $\mu + \beta_j$, where we assume that $\sum_j \beta_j = 0$.

Factor B ignoring factor A

The effects α_i and β_j are called the *main effects* of factors A and B, respectively. The main effect of factor A describes variation due solely to the level of factor A (row of the response matrix), and the main effect of factor B describes variation due solely to the level of factor B (column of the response matrix). We have analogously that SS_A and SS_B are main-effects sums of squares.

A main effect describes variation due to a single factor

The variation described by the main effects is variation that occurs from row to row or column to column of the data matrix. The example has twelve treatments and 11 degrees of freedom between treatments. We have described 5 degrees of freedom using the A and B main effects, so there must be 6 more degrees of freedom left to model. These 6 remaining degrees of freedom describe variation that arises from changing rows and columns simultaneously. We call such variation *interaction* between factors A and B, or between the rows and columns, and denote it by SS_{AB}.

Interaction is variation not described by main effects

Here is another way to think about main effect and interaction. The main effect of rows tells us how the response changes when we move from one row to another, averaged across all columns. The main effect of columns tells us how the response changes when we move from one column to another, averaged across all rows. The interaction tells us how the change in response depends on columns when moving between rows, or how the change in response depends on rows when moving between columns. Interaction between factors A and B means that the change in mean response going from level i_1 of factor A to level i_2 of factor A depends on the level of factor B under consideration. We can't simply say that changing the level of factor A changes the response by a given amount; we may need a different amount of change for each level of factor B.

We can make our description of main-effect and interaction variation more precise by using contrasts. Any contrast in factor A (ignoring B) has four coefficients $w_i^\star$ and observed value $w^\star(\{\bar{y}_{i\bullet\bullet}\})$. This is a contrast in the four row means. We can make an equivalent contrast in the twelve treatment means by using the coefficients $w_{ij} = w_i^\star/3$. This contrast just repeats $w_i^\star$ across each row and then divides by the number of columns to match up

Table 8.4: Example main-effects and interaction contrast coefficients for a four by three factorial design.

A

$$
\begin{array}{|ccc|}
\hline
-3 & -3 & -3 \\
-1 & -1 & -1 \\
1 & 1 & 1 \\
3 & 3 & 3 \\
\hline
\end{array}
\qquad
\begin{array}{|ccc|}
\hline
1 & 1 & 1 \\
-1 & -1 & -1 \\
-1 & -1 & -1 \\
1 & 1 & 1 \\
\hline
\end{array}
\qquad
\begin{array}{|ccc|}
\hline
-1 & -1 & -1 \\
3 & 3 & 3 \\
-3 & -3 & -3 \\
1 & 1 & 1 \\
\hline
\end{array}
$$

B

$$
\begin{array}{|ccc|}
\hline
-1 & 0 & 1 \\
-1 & 0 & 1 \\
-1 & 0 & 1 \\
-1 & 0 & 1 \\
\hline
\end{array}
\qquad
\begin{array}{|ccc|}
\hline
1 & -2 & 1 \\
1 & -2 & 1 \\
1 & -2 & 1 \\
1 & -2 & 1 \\
\hline
\end{array}
$$

AB

$$
\begin{array}{|ccc|}
\hline
3 & 0 & -3 \\
1 & 0 & -1 \\
-1 & 0 & 1 \\
-3 & 0 & 3 \\
\hline
\end{array}
\qquad
\begin{array}{|ccc|}
\hline
-1 & 0 & 1 \\
1 & 0 & -1 \\
1 & 0 & -1 \\
-1 & 0 & 1 \\
\hline
\end{array}
\qquad
\begin{array}{|ccc|}
\hline
1 & 0 & -1 \\
-3 & 0 & 3 \\
3 & 0 & -3 \\
-1 & 0 & 1 \\
\hline
\end{array}
$$

$$
\begin{array}{|ccc|}
\hline
-3 & 6 & -3 \\
-1 & 2 & -1 \\
1 & -2 & 1 \\
3 & -6 & 3 \\
\hline
\end{array}
\qquad
\begin{array}{|ccc|}
\hline
1 & -2 & 1 \\
-1 & 2 & -1 \\
-1 & 2 & -1 \\
1 & -2 & 1 \\
\hline
\end{array}
\qquad
\begin{array}{|ccc|}
\hline
-1 & 2 & -1 \\
3 & -6 & 3 \\
-3 & 6 & -3 \\
1 & -2 & 1 \\
\hline
\end{array}
$$

with the division used when computing row means. Factor A has four levels, so three orthogonal contrasts partition SS_A. There are three analogous orthogonal w_{ij} contrasts that partition the same variation. (See Question 8.1.) Table 8.4 shows one set of three orthogonal contrasts describing the factor A variation; many other sets would do as well. *(Main-effects contrasts)*

The variation in SS_B can be described by two orthogonal contrasts between the three levels of factor B. Equivalently, we can describe SS_B with orthogonal contrasts in the twelve treatment means, using a matrix of contrast coefficients that is constant on columns (that is, $w_{1j} = w_{2j} = w_{3j} = w_{4j}$ for all columns j). Table 8.4 also shows one set of orthogonal contrasts for factor B.

Inspection of Table 8.4 shows that not only are the factor A contrasts orthogonal to each other, and the factor B contrasts orthogonal to each other, but the factor A contrasts are also orthogonal to the factor B contrasts. This orthogonality depends on balanced data and is the key reason why balanced data are easier to analyze. *(A contrasts orthogonal to B contrasts for balanced data)*

There are 11 degrees of freedom between the twelve treatments, and the A and B contrasts describe 5 of those 11 degrees of freedom. The 6 additional degrees of freedom are interaction degrees of freedom; sample interaction contrasts are also shown in Table 8.4. Again, inspection shows that the interaction contrasts are orthogonal to both sets of main-effects contrasts. Thus the 11 degrees of freedom between-treatment sum of squares can be *(Interaction contrasts)*

partitioned using contrasts into SS_A, SS_B, and SS_{AB}.

Look once again at the form of the contrast coefficients in Table 8.4. Row-main-effects contrast coefficients are constant along each row, and add to zero down each column. Column-main-effects contrasts are constant down each column and add to zero along each row. Interaction contrasts add to zero down columns and along rows. This pattern of zero sums will occur again when we look at parameters in factorial models.

Contrast coefficients satisfy zero-sum restrictions

8.3 Advantages of Factorials

Before discussing advantages, let us first recall the difference between factorial treatment structure and factorial analysis. Factorial analysis is an option we have when the treatments have factorial structure; we can always ignore main effects and interaction and just analyze the g treatment groups.

Factorial structure versus analysis

It is easiest to see the advantages of factorial treatment structure by comparing it to a design wherein we only vary the levels of a single factor. This second design is sometimes referred to as "one-at-a-time." The sprouting data in Table 8.1 were from a factorial experiment where the levels of sprouting water and seed age were varied. We might instead use two one-at-a-time designs. In the first, we fix the sprouting water at the lower level and vary the seed age across the five levels. In the second experiment, we fix the seed age at the middle level, and vary the sprouting water across two levels.

One-at-a-time designs

Factorial treatment structure has two advantages:

1. When the factors interact, factorial experiments can estimate the interaction. One-at-at-time experiments cannot estimate interaction. Use of one-at-a-time experiments in the presence of interaction can lead to serious misunderstanding of how the response varies as a function of the factors.

2. When the factors do not interact, factorial experiments are more efficient than one-at-a-time experiments, in that the units can be used to assess the (main) effects for both factors. Units in a one-at-a-time experiment can only be used to assess the effects of one factor.

> There are two times when you should use factorial treatment structure: (1) when your factors interact, and (2) when your factors do not interact.

Factorial structure is a win, whether or not we have interaction.

The argument for factorial analysis is somewhat less compelling. We usually wish to have a model for the data that is as simple as possible. When there is no interaction, then main effects alone are sufficient to describe the means of the responses. Such a model (or data) is said to be *additive*. An additive model is simpler (in particular, uses fewer degrees of freedom) than a model with a mean for every treatment. When interaction is moderate compared to main effects, the factorial analysis is still useful. However, in some

Additive model has only main effects

experiments the interactions are so large that the idea of main effects as the primary actors and interaction as fine tuning becomes untenable. For such experiments it may be better to revert to an analysis of g treatment groups, ignoring factorial structure.

Example 8.1 Pure interactive response

Consider a chemistry experiment involving two catalysts where, unknown to us, both catalysts must be present for the reaction to proceed. The response is one or zero depending on whether or not the reaction occurs. The four treatments are the factorial combinations of Catalyst A present or absent, and Catalyst B present or absent. We will have a response of one for the combination of both catalysts, but the other three responses will be zero. While it is possible to break this down as main effect and interaction, it is clearly more comprehensible to say that the response is one when both catalysts are present and zero otherwise. Note here that the factorial treatment structure was still a good idea, just not the main-effects/interactions analysis.

8.4   Visualizing Interaction

An *interaction plot*, also called a *profile plot*, is a graphic for assessing the relative size of main effects and interaction; an example is shown in Figure **??**. Consider first a two-factor factorial design. We construct an interaction plot in a "connect-the-dots" fashion. Choose a factor, say A, to put on the horizontal axis. For each factor level combination, plot the pair $(i, \overline{y}_{ij\bullet})$. Then "connect-the-dots" corresponding to the points with the same level of factor B; that is, connect $(1, \overline{y}_{1j\bullet})$, $(2, \overline{y}_{2j\bullet})$, up to $(a, \overline{y}_{aj\bullet})$. In our four by three prototype factorial, the level of factor A will be a number between one and four; there will be three points plotted above one, three points plotted above two, and so on; and there will be three "connect-the-dots" lines, one for each level of factor B.

For additive data, the change in response moving between levels of factor A does not depend on the level of factor B. In an interaction plot, that similarity in change of level shows up as parallel line segments. Thus interaction is small compared to the main effects when the connect-the-dots lines are parallel, or nearly so. Even with visible interaction, the degree of interaction may be sufficiently small that the main-effects-plus-interaction description is still useful. It is worth noting that we sometimes get visually different impressions of the interaction by reversing the roles of factors A and B.

Interaction plots connect-the-dots between treatment means

Interaction plot shows relative size of main effects and interaction

Example 8.2 Chick body weights

Table 8.5 shows the treatment means for the six treatments in the Nelson, Kriby, and Johnson (1990) experiment. We can use `cfcdae::interactplot()` to visualize the interaction as shown on line 1.

Table 8.5: Average chick body weights (in grams) under mineral supplements. Data from Nelson, Kriby, and Johnson (1990); data set `ChickBodyWeight`.

	Calcium %		
Phosphorus %	.6	.9	1.2
.25	584	489	453
.50	616	606	621
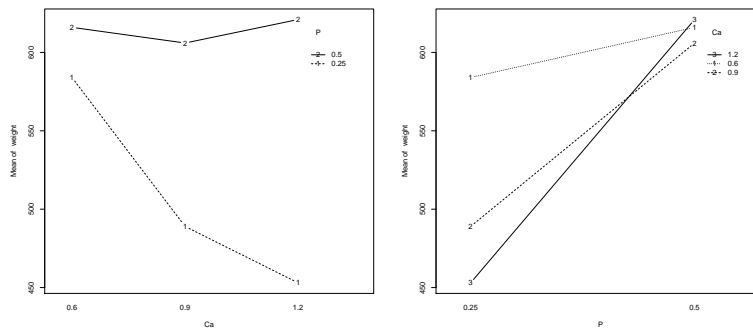
Figure 8.1: Interaction plot of chick body weights data with calcium on the horizontal axis.

```
1 > with(ChickBodyWeight,interactplot(Ca,P,weight))
2 > with(ChickBodyWeight,interactplot(P,Ca,weight))
```

We have put the calcium factor on the horizontal axis and have phosphorus levels indicate the separate traces; see the first panel of Figure 8.1. Here, interaction is seemingly clear. At the upper level of phosphorus, chick weight does not depend on calcium. At the lower level of phosphorus, weight decreases with increasing calcium. Thus the effect of changing calcium levels depends on the level of phosphorus.

It is important to know that reversing the roles of horizontal factor and trace factor does not change the information presented in the interaction plot, but you might find that one order is more understandable than another. For example, line 2 produces the second panel of Figure 8.1. Somehow, to me, the first version of the plot is more understandable.

Interaction is only "seemingly" clear in Figure 8.1, because the basic interaction plot does not tell us anything about variability. If the standard error of those treatment means is 100, then what looks like interaction could simply be random variation. The observed means that we plot are subject to error, so the line segments will not be exactly parallel—even if the true means are additive. The degree to which the lines are not parallel must be interpreted in light of the likely size of the variation in the observed means. As the data

Interpret "parallel" in light of variability

Table 8.6: Coating retention after salt water bath, in percent.
Temperature in degrees C, time in minutes, rate in inches/minute.
Data from N. Ellering, data set `StringCoating`.

		Temperature			
Time	Rate	20	60	90	120
30	6	3.6	7.3	10.8	39.3
		3.4	4.9	5.7	47.9
		2.3	1.7	6.8	34.9
30	12	2.7	3.0	3.0	7.0
		7.2	4.1	2.7	4.0
		5.4	5.4	6.2	7.1
60	6	34.2	37.1	93.2	92.2
		46.2	48.9	87.2	95.8
		37.2	48.5	96.0	91.2
60	12	23.5	29.5	52.3	82.1
		17.0	34.5	47.8	84.7
		15.1	36.6	48.8	86.6

become more variable, greater departures from parallel line segments become
more likely, even for truly additive data.

Example 8.3 String coating retention

Table 8.6 shows the percent retention of coating for the 48 units tested
in the guitar string data. This is a three-factor model (four by two by two).
For more than two factors, we can either look at two factors at a time (one
horizontal factor and one trace factor), or we can use one horizontal factor
and plot a trace for each combination of two or more factors.

We can include pointwise (uncorrected for multiple comparisons) con-
fidence intervals for each treatment mean in an interaction plot by using a
`confidence` argument.

```
1 > with(StringCoating,interactplot(temp.factor,time.factor:rate.factor,pct.retained,
    confidence=.95))
2 > with(StringCoating,interactplot(temp.factor,time.factor:rate.factor,pct.retained,
    confidence=.95,pooled=FALSE))
3 > with(StringCoating,interactplot(temp.factor,rate.factor,pct.retained,
    confidence=.95))
4 > with(StringCoating,interactplot(temp.factor,rate.factor,pct.retained,
    confidence=.95,sigma2=14.1,df=32))
```

Line 1 creates the interaction plot shown in panel 1 of Figure 8.2. Note that
we get traces for combinations of two (or more) variables by entering their in-
teraction as the trace variable (here, `time.factor:rate.factor`). We
can clearly see that the deviations from parallel line segments are far beyond
sampling variation.

Line 2 produces a similar plot (panel 2 of Figure 8.2), except instead

Table 8.7: Zinc retention after a diet with two levels of zinc, and a last meal with two levels of zinc and four levels of protein. Data from Hunt and Larson (1990), data set `ZincRetention`.

Meal Zinc	Diet Zinc	Meal Protein			
		1	2	3	4
1	1	52	73	76	80
1	2	74	89	89	88
2	1	48	44	59	64
2	2	62	56	69	79

of using a pooled estimate of variance across all treatments (which is the default), it uses error variances estimated separately for each treatment; this is chosen via `pooled=FALSE`. Unpooled variances will be a more reasonable representation when the error variance is not constant, and it does not appear to be constant in these data. Note, however, that confidence intervals with individually-estimated variances will typically be wider and more variable, because each of these estimates of variance is done on relatively few degrees of freedom.

interactplot only knows about the factors included in its arguments; if there are other factors that you do not include in its arguments, any variation due to those factors will appear to be error variation, probably making the confidence intervals too wide. That is, part of the mean structure will taken as part of the error variance, making the error variance appear to be too large. Line 3 uses temperature as the horizontal factor but only uses rate as a trace factor (instead of both rate and time). Panel 3 of Figure 8.2 shows the resulting plot, and it does have much wider confidence intervals. Line 4 shows that you can tell interactplot what error mean square and error degrees of freedom to use for confidence intervals via the sigma2 and df arguments. In practice, the values for sigma2 and df will not be known until you fit the full model to the data.

Example 8.4 Zinc retention

Finally, let's look at the zinc retention data of Hunt and Larson (1990); treatment means are shown in Table 8.7. Because we only have treatment means and no measure of variability, any judgement we make regarding interaction is somewhat tenuous.

```
1 > with(ZincRetention,interactplot(m.protein,m.zinc:d.zinc,retention))
2 > with(ZincRetention,interactplot(m.protein,m.zinc,retention))
3 > with(ZincRetention,interactplot(m.protein,d.zinc,retention))
4 > with(ZincRetention,interactplot(m.zinc,d.zinc,retention))
```

These commands produce all two-way interaction plots (line 1–3), and a three-way interaction plot with mean protein on the horizontal axis (line 4). These plots are in the four panels of Figure 8.3. The first two panels look
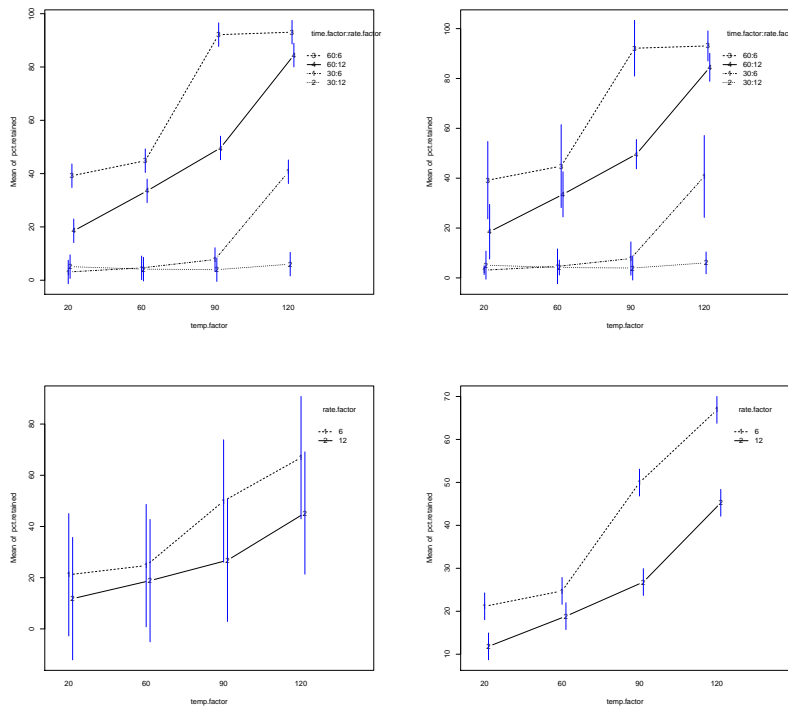
Figure 8.2: Interaction plots of StringCoating data with the temperature factor on the horizontal axis. Panels 1 and 2 show traces for the combinations of the time and rate factors, with panel 1 using a pooled estimate of variance, and panel 2 using separate estimates of variance for each treatment. Panels 3 and 4 show traces for the rate factor only. Panel 3 is misleading, because it includes variation caused by time and its interactions in the confidence intervals. Panel 4 corrects this by using a separately specified residual variance.

fairly parallel. These two panels include diet zinc as one of the factors, suggesting that diet zinc does not interact with either factor. In contrast, panel 3 suggests a possible interaction between meal protein and meal zinc: there seems to be a much larger difference in response between levels of meal zinc at level 2 of meal protein. Of course, we do not have standard errors, so this could just be random variation.

The fourth panel is the three factor interaction plot. Traces 1 and 2 have level 1 of meal zinc but differ on diet zinc. These lines look roughly parallel, because while diet zinc has an effect on the mean, it seems to have the same effect at every level of meal protein when meal zinc is 1. Similarly, traces 3 and 4 have level 2 of meal zinc and look roughly parallel. Joining these observations with the fact that the gap between traces 1 and 2 is roughly the same as the gap between traces 3 and 4, we would conclude that there is little
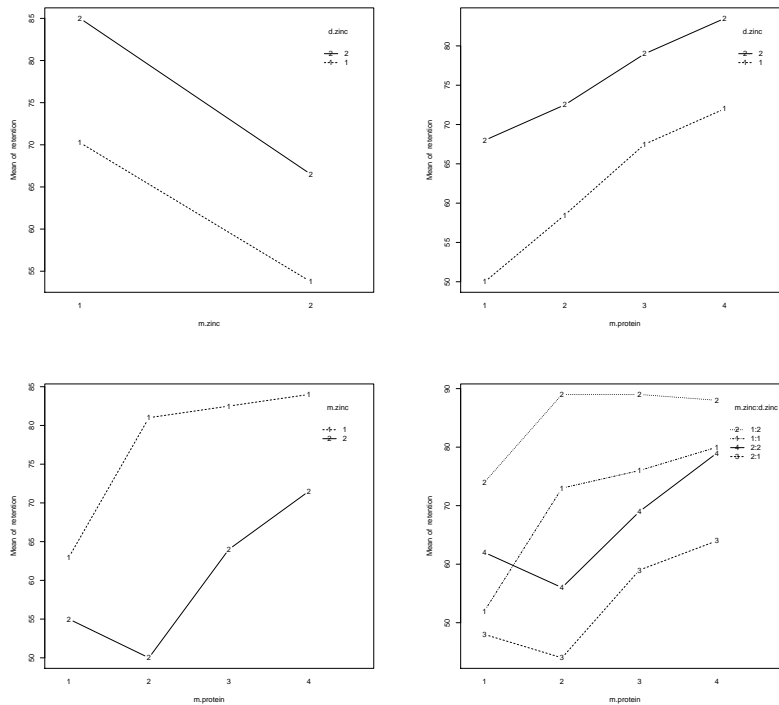
Figure 8.3: Interaction plots of zinc retention data. Panel 1: meal zinc horizontal, diet zinc as trace; panel 2: meal protein horizontal, diet zinc as trace; panel 3: meal protein horizontal, meal zinc as trace; panel 4: meal protein horizontal, combinations of meal zinc and diet zinc as trace.

evidence for a three factor interaction.

8.5   Models with Parameters

Let us now look at the factorial analysis model for a two-way factorial treatment structure. Factor A has a levels, factor B has b levels, and there are n experimental units assigned to each factor-level combination. The kth response at the ith level of A and jth level of B is y_{ijk}. The model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \, ,$$

where i runs from 1 to a, j runs from 1 to b, k runs from 1 to n, and the ϵ_{ijk}'s are independent and normally distributed with mean zero and variance σ^2. The α_i, β_j, and $\alpha\beta_{ij}$ parameters in this model are fixed, unknown constants. There is a total of $N = nab$ experimental units.

A has a levels, B has b levels, n replications

Factorial model

$$
\begin{array}{ccccc}
\text{responses} & & \text{overall mean} & & \text{row effects} \\[4pt]
\begin{bmatrix} y_{111} & y_{121} \\ y_{211} & y_{221} \\ y_{311} & y_{321} \end{bmatrix} & = & \begin{bmatrix} \mu & \mu \\ \mu & \mu \\ \mu & \mu \end{bmatrix} & + & \begin{bmatrix} \alpha_1 & \alpha_1 \\ \alpha_2 & \alpha_2 \\ \alpha_3 & \alpha_3 \end{bmatrix} \quad +
\end{array}
$$

$$
\begin{array}{ccccc}
 & \text{column effects} & & \text{interaction effects} & \\[4pt]
 & \begin{bmatrix} \beta_1 & \beta_2 \\ \beta_1 & \beta_2 \\ \beta_1 & \beta_2 \end{bmatrix} & + & \begin{bmatrix} \alpha\beta_{11} & \alpha\beta_{12} \\ \alpha\beta_{21} & \alpha\beta_{22} \\ \alpha\beta_{31} & \alpha\beta_{32} \end{bmatrix} & +
\end{array}
$$

$$
\begin{array}{c}
\text{random errors} \\[4pt]
\begin{bmatrix} \epsilon_{111} & \epsilon_{121} \\ \epsilon_{211} & \epsilon_{221} \\ \epsilon_{311} & \epsilon_{321} \end{bmatrix}
\end{array}
$$

Display 8.1: Breakdown of a three by two table into factorial effects.

Another way of viewing the model is that the table of responses is broken down into a set of tables which, when summed element by element, give the response. Display 8.1 is an example of this breakdown for a three by two factorial with $n = 1$.

The term μ is called the overall mean; it is the expected value for the responses averaged across all treatments. The term α_i is called the main effect of A at level i. It is the average effect (averaged over levels of B) for level i of factor A. Since the average of all the row averages must be the overall average, these row effects α_i must sum to zero. The same is true for β_j, which is the main effect of factor B at level j. The term $\alpha\beta_{ij}$ is called the interaction effect of A and B in the ij treatment. Do not confuse $\alpha\beta_{ij}$ with the product of α_i and β_j; they are different ideas. The interaction effect is a measure of how far the treatment means differ from additivity. Because the average effect in the ith row must be α_i, the sum of the interaction effects in the ith row must be zero. Similarly, the sum of the interaction effects in the jth column must be zero.

Main effects

Interaction effects

The expected value of the response for treatment ij is

$$E\, y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} \; .$$

There are ab different treatment means, but we have $1 + a + b + ab$ parameters, so we have vastly overparameterized. Recall that in Chapter 3 we had to choose a set of restrictions to make treatment effects well defined; we must again choose some restrictions for factorial models. We will use the following set of restrictions on the parameters:

Expected value

Zero-sum restrictions on parameters

$$
\begin{aligned}
\widehat{\mu} &= \overline{y}_{\bullet\bullet\bullet} \\
\widehat{\alpha}_i &= \overline{y}_{i\bullet\bullet} - \widehat{\mu} = \overline{y}_{i\bullet\bullet} - \overline{y}_{\bullet\bullet\bullet} \\
\widehat{\beta}_j &= \overline{y}_{\bullet j\bullet} - \widehat{\mu} = \overline{y}_{\bullet j\bullet} - \overline{y}_{\bullet\bullet\bullet} \\
\widehat{\alpha\beta}_{ij} &= \overline{y}_{ij\bullet} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j \\
&= \overline{y}_{ij\bullet} - \overline{y}_{i\bullet\bullet} - \overline{y}_{\bullet j\bullet} + \overline{y}_{\bullet\bullet\bullet}
\end{aligned}
$$

Display 8.2: Estimators for main effects and interactions in a two-way factorial.

$$
0 = \sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} \alpha\beta_{ij} = \sum_{j=1}^{b} \alpha\beta_{ij} .
$$

This set of restrictions is standard and matches the description of the parameters in the preceding paragraph. The α_i values must sum to 0, so at most $a - 1$ of them can vary freely; there are $a - 1$ degrees of freedom for factor A. Similarly, the β_j values must sum to 0, so at most $b - 1$ of them can vary freely, giving $b - 1$ degrees of freedom for factor B. For the interaction, we have ab effects, but they must add to 0 when summed over i or j. We can show that this leads to $(a - 1)(b - 1)$ degrees of freedom for the interaction. Note that the parameters obey the same restrictions as the corresponding contrasts: main-effects contrasts and effects add to zero across the subscript, and interaction contrasts and effects add to zero across rows or columns.

Main-effect and interaction degrees of freedom

When we add the degrees of freedom for A, B, and AB, we get $a - 1 + b - 1 + (a - 1)(b - 1) = ab - 1 = g - 1$. That is, the $ab - 1$ degrees of freedom between the means of the ab factor level combinations have been partitioned into three sets: A, B, and the AB interaction. Within each factor-level combination there are $n - 1$ degrees of freedom about the treatment mean. The error degrees of freedom are $N - g = N - ab = (n - 1)ab$, exactly as we would get ignoring factorial structure.

Main effects and interactions partition between treatments variability

The laser transmission data had a six by three factorial structure with $n = 5$. Thus there are 5 degrees of freedom for factor A, 2 degree of freedom for factor B, 10 degrees of freedom for the AB interaction, and 72 degrees of freedom for error.

Display 8.2 gives the formulae for estimating the effects in a balanced two-way factorial. Estimate μ by the mean of all the data $\overline{y}_{\bullet\bullet\bullet}$. Estimate $\mu + \alpha_i$ by the mean of all responses that had treatment A at level i, $\overline{y}_{i\bullet\bullet}$. To get an estimate of α_i itself, subtract our estimate of μ from our estimate of $\mu + \alpha_i$. Do similarly for factor B, using $\overline{y}_{\bullet j\bullet}$ as an estimate of $\mu + \beta_j$. We can extend this basic idea to estimate the interaction terms $\alpha\beta_{ij}$. The expected value in treatment ij is $\mu + \alpha_i + \beta_j + \alpha\beta_{ij}$, which we can estimate by $\overline{y}_{ij\bullet}$, the observed treatment mean. To get an estimate of $\alpha\beta_{ij}$, simply

Estimating factorial effects

Table 8.8: Total free amino acids in cheddar cheese after 56 days of ripening under four nonstarter bacteria treatments. Data from P. Swearingen; data set `CheeseAminoAcid56`.

Control	R50#10	R21#2	Both
1.697	2.032	2.211	2.091
1.601	2.017	1.673	2.255
1.830	2.409	1.973	2.987

subtract the estimates of the lower order parameters (parameters that contain no additional subscripts beyond those found in this term) from the estimate of the treatment mean.

We examine the estimated effects to determine which treatment levels lead to large or small responses, and where factors interact (that is, which combinations of levels have large interaction effects).

Example 8.5 Nonstarter bacteria in cheddar cheese

Cheese is made by bacterial fermentation of Pasteurized milk. Most of the bacteria are purposefully added; these are the starter cultures. Some "wild" bacteria are also present in cheese; these are nonstarter bacteria. This experiment explores how intentionally-added nonstarter bacteria affect cheese quality. We use two strains of nonstarter bacteria: R50#10 and R21#2. Our four treatments will be control, addition of R50, addition of R21, and addition of a blend of R50 and R21. Twelve cheeses are made, three for each of the four treatments, with the treatments being randomized to the cheeses. After 56 days of ripening, each cheese is measured for total free amino acids (a measure of bacterial activity related to cheese quality). Responses are given in Table 8.8.

Let's estimate the effects in these data. The four treatment means are

$$
\begin{aligned}
\overline{y}_{11\bullet} &= (1.697 + 1.601 + 1.830)/3 = 1.709 \quad \text{Control} \\
\overline{y}_{21\bullet} &= (2.032 + 2.017 + 2.409)/3 = 2.153 \quad \text{R50} \\
\overline{y}_{12\bullet} &= (2.211 + 1.673 + 1.973)/3 = 1.952 \quad \text{R21} \\
\overline{y}_{22\bullet} &= (2.091 + 2.255 + 2.987)/3 = 2.444 \quad \text{Blend.}
\end{aligned}
$$

The grand mean is the total of all the data divided by 12,

$$
\overline{y}_{\bullet\bullet\bullet} = 24.776/12 = 2.065 \; ;
$$

the R50 (row or first factor) means are

$$
\begin{aligned}
\overline{y}_{1\bullet\bullet} &= (1.709 + 1.952)/2 = 1.831 \\
\overline{y}_{2\bullet\bullet} &= (2.153 + 2.444)/2 = 2.299 \; ;
\end{aligned}
$$

and the R21 (column or second factor) means are

$$
\begin{aligned}
\overline{y}_{\bullet1\bullet} &= (1.709 + 2.153)/2 = 1.931 \\
\overline{y}_{\bullet2\bullet} &= (1.952 + 2.444)/2 = 2.198 \; .
\end{aligned}
$$

Using the formulae in Display 8.2 we have the estimates

$$\widehat{\mu} \quad = \overline{y}_{\bullet\bullet\bullet} \qquad\qquad = \ 2.065$$

$$\begin{aligned}
\widehat{\alpha}_1 &= 1.831 - 2.065 &= -.234 \\
\widehat{\alpha}_2 &= 2.299 - 2.065 &= .234
\end{aligned} \quad .$$

$$\begin{aligned}
\widehat{\beta}_1 &= 1.931 - 2.065 &= -.134 \\
\widehat{\beta}_2 &= 2.198 - 2.065 &= .134
\end{aligned}$$

Finally, use the treatment means and the previously estimated effects to get the estimated interaction effects:

$$\begin{aligned}
\widehat{\alpha\beta}_{11} &= 1.709 - (2.065 + -.234 + -.134) &= .012 \\
\widehat{\alpha\beta}_{21} &= 2.153 - (2.065 + .234 + -.134) &= -.012 \\
\widehat{\alpha\beta}_{12} &= 1.952 - (2.065 + -.234 + .134) &= -.012 \\
\widehat{\alpha\beta}_{22} &= 2.444 - (2.065 + .234 + .134) &= .012 \ .
\end{aligned}$$

Of course, these computations are trivial using **R**:

```
1 > fit <- lm(freeAminoAcid~r50*r21,data=CheeseAminoAcid56)
2 > fit <- lm(freeAminoAcid~r50+r21+r50:r21,data=CheeseAminoAcid56)
3 > summary(fit)
 ...
 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)  2.06467    0.08694  23.748 1.05e-08 ***
 r501        -0.23383    0.08694  -2.690   0.0275 *
 r211        -0.13367    0.08694  -1.537   0.1627
 r501:r211    0.01217    0.08694   0.140   0.8922
 ...
4 > model.effects(fit,"r50")
          no          yes
 -0.2338333   0.2338333
5 > model.effects(fit,"r21")
          no          yes
 -0.1336667   0.1336667
6 > model.effects(fit,"r50:r21")
               no           yes
 no    0.01216667 -0.01216667
 yes  -0.01216667  0.01216667
```

Lines 1 and 2 fit the same factorial model. Line 1 uses the shortcut notation factor1*factor2, which expands into main effects and interaction, as was done explicitly in line 2. Note that interaction is indicated by two or more factors joined by a colon.

The summary in line 3 only prints information for non-redundant coefficients, verifying what we computed by hand. Lines 4–6 use model.effects to print all the coefficients, even the redundant ones. This is not terribly helpful when there are only two levels, but it can save time when there are more levels.

It is also possible to get predicted values and standard errors of prediction using functions in the `effects` package.

```
 7 > int.effect <- effects::effect('r50:r21',fit)
 8 > int.effect

   r50*r21 effect
       r21
  r50         no      yes
    no  1.709333 1.952333
    yes 2.152667 2.444333
 9 > plot(int.effect)
10 > int.effect$se
   [1] 0.1738849 0.1738849 0.1738849 0.1738849
11 > r21.effect <- effects::effect('r21',fit)
   NOTE: r21 is not a high-order term in the model
12 > r21.effect

   r21 effect
   r21
        no      yes
   1.931000 2.198333
```

Line 7 extracts the interaction "effect." This is actually a linear combination of what we call model effects or coefficients yielding the fitted value for the margin of factor level combinations indicated by the factor; some call this the "least squares mean". For this two-factor interaction, the effect is $\widehat{\mu} + \widehat{\alpha}_i + \widehat{\alpha}_j + \widehat{\alpha\beta}_{ij}$. Line 8 prints the "effect", which in this case is the table of treatment means. Line 9 produces a nice plot, Figure 8.4, related to an interaction plot, that can help you visualize the fitted values. The `effect` function also computes standard errors for the fitted values, which we extract in line 10. Looking at the effect for a term that is included in a higher order interaction can be very misleading. Line 11 shows that you get a warning when you try to do that. This effect is shown in line 12.

8.6 The Analysis of Variance for Balanced Factorials

We have described the Analysis of Variance as an algorithm for partitioning variability in data, a method for testing null hypotheses, and a method for comparing models for data. The same roles hold in factorial analysis, but we now have more null hypotheses to test and/or models to compare.

We partition the variability in the data by using ANOVA. There is a source of variability for every term in our model; for a two-factor analysis, these are factor A, factor B, the AB interaction, and error. In a one-factor ANOVA, we obtained the sum of squares for treatments by first squaring an estimated effect (for example, $\widehat{\alpha}_i^2$), then multiplying by the number of units receiving that effect (n_i), and finally adding over the index of the effect (for example, add over i for α_i). The total sum of squares was found by summing the squared deviations of the data from the overall mean, and the error sum of squares was found by summing the squared deviations of the data

ANOVA source for every term in model

Sum of squares

r50*r21 effect plot
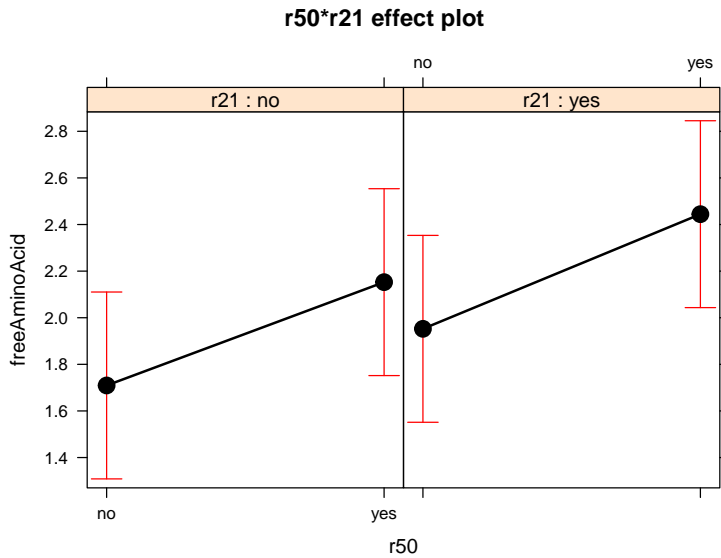
Figure 8.4: Effect plot for the two-factor interaction of the 56-day cheese amino acid data.

Term	Sum of Squares	Degrees of Freedom
A	$\sum\limits_{i=1}^{a} bn(\widehat{\alpha}_i)^2$	$a-1$
B	$\sum\limits_{j=1}^{b} an(\widehat{\beta}_j)^2$	$b-1$
AB	$\sum\limits_{i=1,j=1}^{a,b} n(\widehat{\alpha\beta}_{ij})^2$	$(a-1)(b-1)$
Error	$\sum\limits_{i=1,j=1,k=1}^{a,b,n} (y_{ijk} - \bar{y}_{ij\bullet})^2$	$ab(n-1)$
Total	$\sum\limits_{i=1,j=1,k=1}^{a,b,n} (y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2$	$abn-1$

Display 8.3: Sums of squares in a balanced two-way factorial.

from the treatment means. We follow exactly the same program for balanced factorials, obtaining the formulae in Display 8.3.

The sums of squares must add up in various ways. For example

$$\mathrm{SS_T = SS_A + SS_B + SS_{AB}} + SSE \ .$$

Also recall that $\mathrm{SS_A}$, $\mathrm{SS_B}$, and $\mathrm{SS_{AB}}$ must add up to the sum of squares between treatments, when considering the experiment to have $g = ab$ treatments, so that

<div align="right">SS partitions</div>

$$\sum_{i=1,j=1}^{a,b} n(\overline{y}_{ij\bullet} - \overline{y}_{\bullet\bullet\bullet})^2 = \mathrm{SS_A + SS_B + SS_{AB}} \ .$$

These identities can provide useful checks on ANOVA computations.

We display the results of an ANOVA decomposition in an Analysis of Variance table. As before, the ANOVA table has columns for source, degrees of freedom, sum of squares, mean square, and F. For the two-way factorial, the sources of variation are factor A, factor B, the AB interaction, and error, so the table looks like this:

<div align="right">Two-factor ANOVA table</div>

Source	DF	SS	MS	F
A	a-1	$\mathrm{SS_A}$	$\mathrm{SS_A}/(a-1)$	$\mathrm{MS_A/MS_E}$
B	b-1	$\mathrm{SS_B}$	$\mathrm{SS_B}/(b-1)$	$\mathrm{MS_B/MS_E}$
AB	(a-1)(b-1)	$\mathrm{SS_{AB}}$	$\mathrm{SS_{AB}}/[(a-1)(b-1)]$	$\mathrm{MS_{AB}/MS_E}$
Error	(n-1)ab	$\mathrm{SS_E}$	$\mathrm{SS_E}/[(n-1)ab]$	

Tests or model comparisons require assumptions on the errors. We have assumed that the errors ϵ_{ijk} are independent and normally distributed with constant variance. When the assumptions are true, the sums of squares as random variables are independent of each other and the tests discussed below are valid.

<div align="right">Normality needed for testing</div>

To test the null hypothesis $H_0\colon \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ against the alternative that some α_i's are not zero, we use the F-statistic $\mathrm{MS_A/MS_E}$ with $a - 1$ and $ab(n-1)$ degrees of freedom. This is a test of the main effect of A. The p-value is calculated as before. To test $H_0\colon \beta_1 = \beta_2 = \cdots = \beta_b = 0$ against the alternative hypothesis that at least one β is nonzero, use the F-statistic $\mathrm{MS_B/MS_E}$, with $b - 1$ and $ab(n-1)$ degrees of freedom. Similarly, the test statistic for the null hypothesis that the $\alpha\beta$ interaction terms are all zero is $\mathrm{MS_{AB}/MS_E}$, with $(a-1)(b-1)$ and $ab(n-1)$ degrees of freedom. Alternatively, these tests may be viewed as comparisons between models that include and exclude the terms under consideration.

<div align="right">F-tests for factorial null hypotheses</div>

> It usually does not make sense to test the null hypothesis that the coefficients of a main effect are zero if we include in the model an interaction that includes the main effect.

Thus we would usually only test the main effect of A if the AB interaction is not needed in, or excluded from, the model. We will return to this is more detail when we discuss hierarchy and unbalanced designs.

Example 8.6 Nonstarter bacteria, continued

We compute sums of squares using the effects of Example 8.5 and the formulae of Display 8.3.

$$SS_{R50} = 6 \times ((-.234)^2 + .234^2) = .656$$
$$SS_{R21} = 6 \times ((-.134)^2 + .134^2) = .214$$
$$SS_{R50.R21} = 3 \times (.012^2 + (-.012)^2 + (-.012)^2 + .012^2) = .002$$

Computing SS_E is more work:

$$SS_E = (1.697 - 1.709)^2 + (2.032 - 2.153)^2 + (2.211 - 1.952)^2$$
$$+ (2.091 - 2.444)^2 + \cdots + (2.987 - 2.444)^2 = .726 .$$

We have $a = 2$ and $b = 2$, so the main effects and the two-factor interaction have 1 degree of freedom each; there are $12 - 4 = 8$ error degrees of freedom. Combining, we get the ANOVA table:

Source	DF	SS	MS	F	p-value
R50	1	.656	.656	7.23	.028
R21	1	.214	.214	2.36	.16
R50.R21	1	.002	.002	.02	.89
Error	8	.726	.091		

The large p-values indicate that we have no evidence that R21 interacts with R50 or causes a change in total free amino acids. The p-value of .028 indicates modest evidence that R50 may affect total free amino acids.

It is common to test the null hypotheses (or compare models) in just the way we have described here. *But we just did three tests! Don't we need to consider multiple comparisons? How about a Bonferroni adjustment or similar?* I think that we should consider a Bonferroni correct, but common practice ignores the issue.

Bonferroni adjustment?

Getting the ANOVA table for a factorial model is straightforward in **R**.

```
8 > anova(fit)
nalysis of Variance Table

esponse: freeAminoAcid
          Df  Sum Sq Mean Sq F value  Pr(>F)
50         1 0.65614 0.65614  7.2335 0.02752 *
21         1 0.21440 0.21440  2.3636 0.16275
50:r21     1 0.00178 0.00178  0.0196 0.89217
esiduals   8 0.72566 0.09071
```

8.7   General Factorial Models

The model and analysis of a multi-way factorial are similar to those of a two-way factorial. Consider a four-way factorial with factors A, B, C and D,

which match with the letters α, β, γ, and δ. The model is

$$
\begin{aligned}
y_{ijklm} ={}& \mu + \alpha_i + \beta_j + \gamma_k + \delta_l \\
& + \alpha\beta_{ij} + \alpha\gamma_{ik} + \alpha\delta_{il} + \beta\gamma_{jk} + \beta\delta_{jl} + \gamma\delta_{kl} \\
& + \alpha\beta\gamma_{ijk} + \alpha\beta\delta_{ijl} + \alpha\gamma\delta_{ikl} + \beta\gamma\delta_{jkl} \\
& + \alpha\beta\gamma\delta_{ijkl} \\
& + \epsilon_{ijklm} \;\;.
\end{aligned}
$$

The first line contains the overall mean and main effects for the four factors; the second line has all six two-factor interactions; the third line has three-factor interactions; the fourth line has the four-factor interaction; and the last line has the error. Just as a two-factor interaction describes how a main effect changes depending on the level of a second factor, a three-factor interaction like $\alpha\beta\gamma_{ijk}$ describes how a two-factor interaction changes depending on the level of a third factor. Similarly, four-factor interactions describe how three-factor interactions depend on a fourth factor, and so on for higher order interactions.

Multi-factor interactions

We still have the assumption that the ϵ's are independent normals with mean 0 and variance σ^2. Analogous with the two-factor case, we restrict our effects so that they will add to zero when summed over any subscript. For example,

Zero-sum restrictions on parameters

$$
0 = \sum_l \delta_l = \sum_k \beta\gamma_{jk} = \sum_j \alpha\beta\delta_{ijl} = \sum_i \alpha\beta\gamma\delta_{ijkl} \;\;.
$$

These zero-sum restrictions make the model parameters unique. The $abcd - 1$ degrees of freedom between the $abcd$ treatments are assorted among the terms as follows. Each term contains some number of factors—one, two, three, or four—and each factor has some number of levels—a, b, c, or d. To get the degrees of freedom for a term, subtract one from the number of levels for each factor in the term and take the product. Thus, for the ABD term, we have $(a-1)(b-1)(d-1)$ degrees of freedom.

Degrees of freedom for general factorials

Effects in the model are estimated analogously with how we estimated effects for a two-way factorial, building up from overall mean, to main effects, to two-factor interactions, to three-factor interactions, and so on. The estimate of the overall mean is $\widehat{\mu} = \sum_{ijklm} y_{ijklm}/N = \overline{y}_{\bullet\bullet\bullet\bullet\bullet}$. Main-effect and two-factor interaction estimates are just like for two-factor factorials, ignoring all factors but the two of interest. For example, to estimate a main effect, say the kth level of factor C, we take the mean of all responses that received the kth level of factor C, and subtract out the lower order estimated effects, here just $\widehat{\mu}$:

Main effects and two-factor estimates as before

$$
\widehat{\gamma}_k = \overline{y}_{\bullet\bullet k\bullet\bullet} - \widehat{\mu} \;\;.
$$

For a three-way interaction, say the ijkth level of factors A, B, and C, we take the mean response at the ijk combination of factors A, B, and C, and then subtract out the lower order terms—the overall mean; main effects of A, B, and C; and two-factor interactions in A, B, and C:

Multi-way effects for general factorials

$$
\widehat{\alpha\beta\gamma}_{ijk} = \overline{y}_{ijk\bullet\bullet} - (\widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j + \widehat{\gamma}_k + \widehat{\alpha\beta}_{ij} + \widehat{\alpha\gamma}_{ik} + \widehat{\beta\gamma}_{jk}) \;\;.
$$

Simply continue this general rule for higher order interactions.

The rules for computing sums of squares follow the usual pattern: square each effect, multiply by the number of units that receive that effect, and add over the levels. Thus,

$$\mathrm{SS_{ABD}} = \sum_{ijl} nc(\widehat{\alpha\beta\delta}_{ijl})^2 \; ,$$

and so on.

As with the two-factor factorial, the results of the Analysis of Variance are summarized in a table with the usual columns and a row for each term in the model. We test the null hypothesis that the effects in a given term are all zeroes by taking the ratio of the mean square for that term to the mean square for error and comparing this observed F to the F-distribution with the corresponding numerator and denominator degrees of freedom. Alternatively, we can consider these F-tests to be tests of whether a given term is needed in a model for the data.

It is clear by now that the computations for a multi-way factorial are tedious at best and should be performed on a computer using statistical software. However, you might be stranded on a desert island (or in an exam room) and need to do a factorial analysis by hand. Here is a technique for multi-way factorials that reorganizes the computations required for computing factorial effects; some find this easier for hand work. The general approach is to compute an effect, subtract it from the data, and then compute the next effect on the differences from the preceding step. This way we only need to subtract out lower order terms once, and it is easier to keep track of things.

First compute the overall mean $\widehat{\mu}$ and subtract it from all the data values. Now, compute the mean of the differences at each level of factor A. Because we have already subtracted out the overall mean, these means are the estimated effects for factor A. Now subtract these factor A effects from their corresponding entries in the differences. Proceed similarly with the other main effects, estimating and then sweeping the effects out of the differences. To get a two-factor interaction, get the two-way table of difference means. Because we have already subtracted out the grand mean and main effects, these means are the two-factor interaction effects. Continue by computing two-way means and sweeping the effects out of the differences. Proceed up through higher order interactions. As long as we proceed in a hierarchical fashion, we will obtain the desired estimated effects.

Example 8.7 Faulting in Concrete Pavement

Faulting is a difference in height on two sides of a concrete pavement joint. Many factors can influence the amount of faulting as the pavement ages. This study examines four factors: the thickness of the base layer (5, 10, or 15 inches), the resilience of the subgrade layer (5,000, 10,000, or 20,000 psi), the ratio of water to cement in the mix (.33, .42, .55), and the presence of dowel pins between successive pavement slabs (yes or no). The re-

Table 8.9: Concrete faulting in inches after 20 year design life. Factors are subgrade resilience modulus (5,000, 10,000, or 20,000 psi); base thickness (5, 10, or 15 inches); water/cement ratio (.33, .42, .55); and dowels in joints (yes or no). Simulated data from N. Funk; data set `ConcreteFaulting`.

| | | | Subgrade modulus/Base | | | | | | | |
| | | 5,000 | | | 10,000 | | | 20,000 | | |
Ratio	Dowels	5	10	15	5	10	15	5	10	15
.33	No	.222	.189	.183	.230	.204	.196	.243	.220	.201
		.220	.189	.183	.225	.202	.197	.239	.218	.210
.33	Yes	.178	.142	.137	.184	.156	.149	.196	.172	.163
		.173	.140	.136	.184	.153	.149	.195	.171	.161
.42	No	.230	.198	.192	.238	.213	.205	.252	.229	.219
		.230	.197	.195	.236	.213	.205	.249	.226	.219
.42	Yes	.189	.153	.148	.196	.167	.161	.208	.184	.175
		.182	.151	.146	.193	.164	.159	.204	.182	.175
.55	No	.239	.208	.202	.249	.224	.217	.263	.242	.232
		.234	.207	.204	.248	.221	.216	.261	.241	.231
.55	Yes	.199	.163	.158	.206	.178	.171	.219	.194	.185
		.195	.162	.160	.201	.176	.170	.218	.192	.185

sponse is faulting (in inches) after a 20 year life. There are two runs for each factor-level combination, resulting in 108 observations; the data are shown in Table 8.9.

We begin by fitting the four-factor model in line 1.
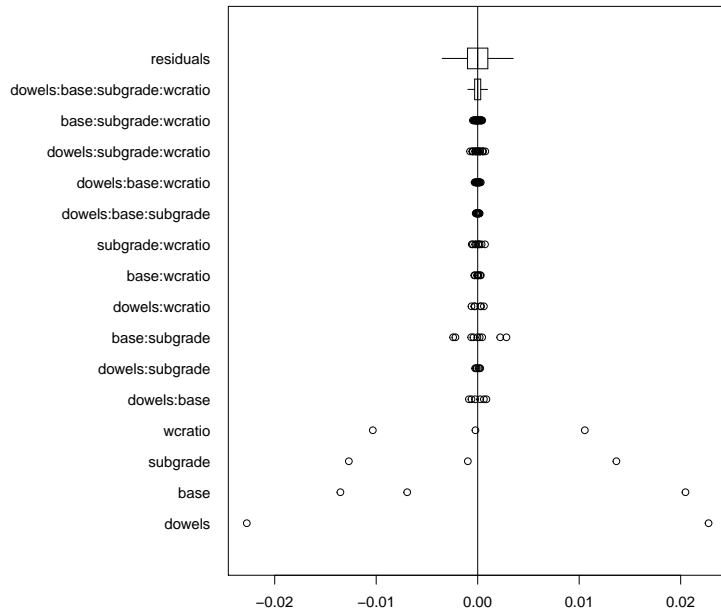
Figure 8.5: Side-by-side plot for concrete faulting data.

```
1 > fit <- lm(faulting~dowels*base*subgrade*wcratio,data=ConcreteFaulting)
2 > plot(fit)
3 > anova(fit)
 Analysis of Variance Table

 Response: faulting
                              Df    Sum Sq  Mean Sq    F value     Pr(>F)
 dowels                        1 0.055897 0.055897 18128.6757 < 2.2e-16 ***
 base                          2 0.023412 0.011706  3796.5676 < 2.2e-16 ***
 subgrade                      2 0.012559 0.006280  2036.6577 < 2.2e-16 ***
 wcratio                       2 0.007857 0.003928  1274.0901 < 2.2e-16 ***
 dowels:base                   2 0.000042 0.000021     6.8739   0.00219 **
 dowels:subgrade               2 0.000003 0.000002     0.5676   0.57025
 base:subgrade                 4 0.000293 0.000073    23.7793 2.239e-11 ***
 dowels:wcratio                2 0.000020 0.000010     3.2793   0.04528 *
 base:wcratio                  4 0.000004 0.000001     0.3468   0.84507
 subgrade:wcratio              4 0.000016 0.000004     1.3288   0.27108
 dowels:base:subgrade          4 0.000001 0.000000     0.1036   0.98077
 dowels:base:wcratio           4 0.000003 0.000001     0.2207   0.92574
 dowels:subgrade:wcratio       4 0.000018 0.000004     1.4279   0.23726
 base:subgrade:wcratio         8 0.000006 0.000001     0.2545   0.97753
 dowels:base:subgrade:wcratio  8 0.000019 0.000002     0.7815   0.62072
 Residuals                    54 0.000167 0.000003
 ---
 Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
4 > sidebyside(fit,left.margin=13)
```

We then check the residuals visually in line 2. There are no problems apparent, so we do not show the plots. However, the normal plot of residuals does give a hint that these are simulated data; can you see it? Having decided assumptions are reasonably well met, we do an analysis of variance in line 3. We see that most of the interactions are not significant. Only the base by subgrade interaction is highly significant, and the dowels by base and dowels by ratio interactions are marginally significant We can visualize the relative size of these effects in a side-by-side plot, created in line 4 and displayed in Figure 8.5. Only the main effects and one interaction look big compared to error variation (although keep in mind that something can look small relative to error variation and still be statistically significant, you just need a big enough sample size).

Figure 8.6 shows interaction plots created in lines 5–7 for the three interactions that were (potentially) significant.

```
 5 > with(ConcreteFaulting,interactplot(wcratio,dowels,faulting,confidence=.95,
      sigma2=.000003,df=54))
 6 > with(ConcreteFaulting,interactplot(base,subgrade,faulting,confidence=.95,
      sigma2=.000003,df=54))
 7 > with(ConcreteFaulting,interactplot(base,dowels,faulting,confidence=.95,
      sigma2=.000003,df=54))
 8 > effects::effect("base:subgrade",fit)
   NOTE: base:subgrade is not a high-order term in the model
    base*subgrade effect
       subgrade
   base      5000     10000     20000
     5  0.2075833 0.2158333 0.2289167
    10 0.1749167 0.1892500 0.2059167
    15 0.1703333 0.1829167 0.1970833
```

Note that these plots include confidence intervals, with error mean square and degrees of freedom copied from the ANOVA table. These intervals are smaller than the numbers labeling the traces. Visually, the interaction appears mostly as the differences across subgrade being smaller for base 5 than for the other bases. This is verified in the effects from line 8.

It looks like the lowest faulting will be achieved with dowels, 15 inch base, 5,000 psi resilience, and .33 ratio, but are there any other combinations that work as well?

```
 9 > myCF <- within(ConcreteFaulting,
   {combined=conf.design::join(dowels,base,subgrade,wcratio)})
10 > fit2 <- lm(faulting~combined,data=myCF)
11 > compare.to.best(fit2,combined,lowisbest=TRUE)
                                    difference    allowance
...
* yes:15:10000:0.33 - yes:15:5000 :0.33     0.0125 0.005251383
* yes:15:5000 :0.42 - yes:15:5000 :0.33     0.0105 0.005251383
  yes:10:5000 :0.33 - yes:15:5000 :0.33     0.0045 0.005251383
best is yes:15:5000 :0.33                   0.0000         NA
```

Line 9 creates a new factor that joins all 54 factor-level combinations into a single factor, and line 10 refits the model using that factor. Line 11 finds all the factor-level combinations that are not significantly different from the best
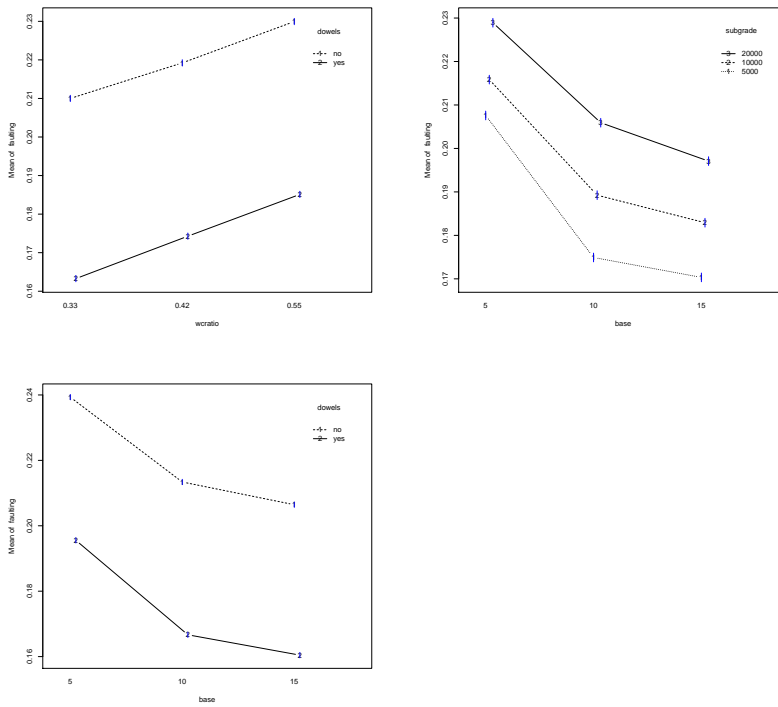
Figure 8.6: Interaction plots for concrete faulting data: ratio by dowels, base by subgrade, and base by dowels.

(lowest). In this case, only one other combination, replacing the 15 inch base with a 10 inch base, cannot be distinguished from the best.

Example 8.8 Faulting in Concrete Pavement, a Bayesian glimpse

Bayesians use interaction plots and residual plots in similar ways to frequentists, but fit and choose models differently. `BayesFactor::anovaBF` can sift through many factorial submodels quickly.

```
12 > modelBFs <- BayesFactor::anovaBF(faulting~base*subgrade*dowels*wcratio,
      data=ConcreteFaulting,whichModels="withmain")
13 > sort(modelBFs)
   ...
   [164] dowels + base + dowels:base + subgrade + base:subgrade + wcratio +
      dowels:wcratio + subgrade:wcratio: 2.716193e+113 2.81%
   [165] dowels + base + dowels:base + subgrade + base:subgrade + wcratio:
      3.477098e+113 4.46%
   [166] dowels + base + dowels:base + subgrade + base:subgrade + wcratio +
      dowels:wcratio: 6.563809e+113 2.17%

   Against denominator:
     Intercept only
   ---
   Bayes factor type: BFlinearModel, JZS
```

Line 12 shows the call to `anovaBF`. Using `whichModels="withmain"` says to start with the intercept-only model, and gradually add term by term, making sure that lower order terms are always included. It returns the Bayes factor compared to the intercept-only model for each of these models. Even with only four factors, there are 166 potential models to consider. Line 13 sorts these Bayes factors, and we have included output for the three highest. Here again, the model with the highest Bayes factor is the model with main effects and the three two-factor interactions we discussed before.

Example 8.9 Faulting in Concrete Pavement, a predictive approach

We have discussed the use of AIC for model selection, and we can employ AIC to select a good model for the concrete faulting data. We want to compare AIC for hierarchical models (models where the presence of an interaction implies the presence of all terms contained in the interaction, see Section 8.11) and choose the one with lowest AIC. The function `step` in **R** starts with the full model, and then step by step removes the term that most decreases AIC, with the provisos that it can only remove terms that do not break hierarchy and that it stops removing terms if removing a term would increase the AIC.

```
14 > fit <- lm(faulting~dowels*base*subgrade*wcratio,ConcreteFaulting)
15 > step(fit)
Start:  AIC=-1337.33
faulting ~ dowels * base * subgrade * wcratio

                                Df  Sum of Sq          RSS      AIC
- dowels:base:subgrade:wcratio   8 1.9278e-05 0.00018578 -1341.5
<none>                                        0.00016650 -1337.3

Step:  AIC=-1341.49
faulting ~ dowels + base + subgrade + wcratio + dowels:base +
    dowels:subgrade + base:subgrade + dowels:wcratio + base:wcratio +
    subgrade:wcratio + dowels:base:subgrade + dowels:base:wcratio +
    dowels:subgrade:wcratio + base:subgrade:wcratio

                             Df  Sum of Sq         RSS      AIC
- base:subgrade:wcratio       8 6.2778e-06 0.00019206 -1353.9
- dowels:base:subgrade        4 1.2778e-06 0.00018706 -1348.8
- dowels:base:wcratio         4 2.7222e-06 0.00018850 -1347.9
<none>                                     0.00018578 -1341.5
- dowels:subgrade:wcratio     4 1.7611e-05 0.00020339 -1339.7
```

The output from `step` in line 15 is voluminous, and we only show part. The first step is the full model with AIC of -1337.33, and the only term which can be removed and maintain hierarchy is the four-factor interaction, which yields an AIC of -1341.5. This term is removed, and we consider all potential terms to remove next, which are the four three-factor interactions. Removing one of them increases the AIC, and removing any one of the other three reduces the AIC, with `base:subgrade:wcratio` yielding the best AIC of -1353.9. Not shown here, but we proceed step by step removing in turn `dowels:base:subgrade`, `dowels:base:wcratio`, and `base:wcratio`, where we arrive at the bottom of the output below line 15.

```
Step:  AIC=-1373.35
 faulting ~ dowels + base + subgrade + wcratio + dowels:base +
    dowels:subgrade + base:subgrade + dowels:wcratio + subgrade:wcratio +
    dowels:subgrade:wcratio

                           Df  Sum of Sq         RSS      AIC
<none>                                   0.00020033 -1373.3
- dowels:subgrade:wcratio   4 1.7611e-05 0.00021794 -1372.2
- dowels:base               2 4.2389e-05 0.00024272 -1356.6
- base:subgrade             4 2.9328e-04 0.00049361 -1284.0
```

At this point, removing any of the potential terms increases the AIC, and we stop with the selected model. AIC is generally more liberal than testing, and we see this in the additional terms selected for the model.

BIC (Bayesian Information Criterion) looks like AIC but uses $\log(N)$ as the multiplier in the penalty for number of parameters instead of 2. Asymptotically as the number of data points increases, BIC will choose the same model as the Bayes factor, although they need not agree for any finite N. In these data we have $N = 108$, and we can employ BIC as shown in line 16.

BIC

```
16 > step(fit,k=log(108))
Start:  AIC=-1192.49
faulting ~ dowels * base * subgrade * wcratio

                                  Df  Sum of Sq        RSS      AIC
- dowels:base:subgrade:wcratio  8 1.9278e-05 0.00018578 -1218.1
<none>                                       0.00016650 -1192.5


...
Step:  AIC=-1332.45
faulting ~ dowels + base + subgrade + wcratio + dowels:base +
    base:subgrade

                  Df Sum of Sq        RSS      AIC
<none>                         0.0002581 -1332.5
- dowels:base     2 0.0000424 0.0003004 -1325.4
- base:subgrade   4 0.0002933 0.0005513 -1269.2
- wcratio         2 0.0078569 0.0081149  -969.4
```

BIC chooses the model with all main effects plus `dowels:base` and `base:subgrade`. This is the model with the second largest Bayes factor in line 13, and its Bayes factor relative to the best model is .53, making them reasonably equivalent in that regard.

8.8 Pooling Terms into Error

Pooling is the practice of adding sums of squares and degrees of freedom for nonsignificant model terms to those of error to form a new (pooled together) error term for further testing. In statistical software, this is usually done by computing the ANOVA for a model that does not include the terms to be pooled into error. I do *not* recommend pooling as standard practice, because pooling may lead to biased estimates of the error. This bias could be negative or positive depending on how you select terms to be pooled into error.

Pooling leads to biased estimates of error

Pooling may be advantageous if there are very few error degrees of freedom. In that case, the loss of power from possible overestimation of the error may be offset by the increase in error degrees of freedom. Only consider pooling a term into error if

Rules for pooling

1. There are 10 or fewer error degrees of freedom, and

2. The term under consideration for pooling has an F-ratio less than 2.

Otherwise, do not pool.

For unbalanced factorials or factorials using polynomial terms, refitting with a model that only includes significant/required terms may be important. This is because in those situations coefficients for terms in a model can depend on what other terms are in the model. Thus while one should test using the full model error variance, you may need to refit with a reduced model to get estimates of the coefficients you need to retain. See Chapter 9.

Coefficients can depend on terms in model

8.9 Assumptions and Transformations

The validity of our inference procedures still depends on the accuracy of our assumptions; factorial structure does not change that. We still need to check for normality, constant variance, and independence and take corrective action as required, just as we did in single-factor models. Corrective action could be a transformation of the response, or it could be use of a GLM or some other model instead of the standard Gaussian model.

<div style="text-align: right">Check assumptions</div>

One wrinkle that occurs for factorial data is that violations of assumptions may sometimes follow the factorial structure. For example, we may find that error variance is constant within a given level of factor B, but differs among levels of factor B.

A second wrinkle with factorials is that the appropriate model for the mean structure can depend on the scale in which we are analyzing the data. Specifically, interaction terms may appear to be needed on one scale but not on another. This is easily seen in the following example. Suppose that the means for the factor level combinations follow the model

<div style="text-align: right">Transformation affects mean structure</div>

$$\mu_{ij} = M \exp \alpha_i \exp \beta_j \ \ .$$

This model is multiplicative in the sense that changing levels of factor A or B rescales the response by multiplying rather than adding to the response. If we fit the usual factorial model to such data, we will need the interaction term, because an additive model won't fit multiplicative data well. For log-transformed data, however, the mean structure is

$$\log\left(\mu_{ij}\right) = \log\left(M\right) + \alpha_i + \beta_j \ \ .$$

Multiplicative data look additive after log transformation; no interaction term is needed. Serendipitously, log transformations often fix nonconstant variance at the same time.

Some people find this confusing at first, and it begs the question of what do we mean by interaction. How can the data have interaction on one scale but not on another? Our use of the term "interaction" is not science of a particular situation; instead, it reflects a particular formulation of the model for the means. Data are interactive *when analyzed on a particular scale* if the main-effects-only model is inadequate and one or more interaction terms are required. Whether or not interaction terms are needed can depend on the scale of the response.

<div style="text-align: right">Interaction depends on scale</div>

Example 8.10 Transmission of laser light through polyvinyl chloride.

The first section of this chapter introduced the laser transmission experiment, where we measure the amount of laser light passing through a piece of clear polyvinyl chloride. The PVC can have one of six thicknesses and one of three surface treatments, for a total of eighteen treatments. Table 8.10 shows the data.

Table 8.10: Percent of laser light transmitted through polyvinyl chloride. PVC has three different thicknesses (mm) and two different surface treatments (none, front side sanded, both sides sanded). Data from J. Van de Ven; data set `LaserTransmission`.

Thickness	Sanding	Transmission (%)				
1.57	none	92.38	92.14	92.24	92.48	92.09
	front	88.77	87.08	86.84	87.03	87.66
	both	86.31	84.23	84.47	83.66	79.94
3.18	none	91.03	90.55	90.41	89.97	90.94
	front	85.87	86.45	88.14	87.51	82.79
	both	84.52	84.67	80.09	78.45	84.96
4.78	none	89.82	89.68	90.55	89.73	90.21
	front	84.47	86.26	85.82	84.52	84.09
	both	83.03	78.11	83.46	78.11	81.63
6.35	none	88.57	89.20	88.24	88.57	89.15
	front	85.15	83.94	82.93	82.69	84.28
	both	82.02	79.94	82.31	81.97	79.51
9.53	none	87.08	87.42	86.84	87.42	87.51
	front	83.80	80.86	80.62	80.67	80.18
	both	78.35	78.06	81.87	80.28	78.98
12.70	none	85.87	86.16	86.89	87.27	87.13
	front	78.30	83.51	82.84	80.81	79.36
	none	77.77	75.65	78.30	78.93	75.55

We begin this extended example by fitting the full factorial model as shown on line 1:

```
1 > fit <- lm(transmission~thickness*sanding,data=LaserTransmission)
2 > plot(fit,which=1:2)
3 > with(LaserTransmission,interactplot(thickness,sanding,transmission,
    confidence=.95,pooled=FALSE))
4 > with(LaserTransmission,interactplot(thickness,sanding,-1/(100-transmission)^1.5,
    confidence=.95,pooled=FALSE))
5 > fitw <- gls(transmission~thickness*sanding,data=LaserTransmission,
    weights=varIdent(form=~1|sanding))
6 > plot(fitw)
7 > qqnorm(residuals(fitw,type="pearson"))
```

On line 2 we create residual plots shown in the first two panels of Figure 8.7. These plots indicate variance that decreases with the mean. Decreasing variance is unusual, but it arises here because the percentage is capped at 100. The mean can only approach 100 when all of the data cluster near 100; thus the variability decreases. Line 3 creates an interaction plot with unpooled confidence intervals (panel four of Figure 8.7. This also shows the decreasing variance.

The ratio of maximum to minimum for transmission is about 1.22, so no reasonable power transformation will fix this non-constant variance. Instead,
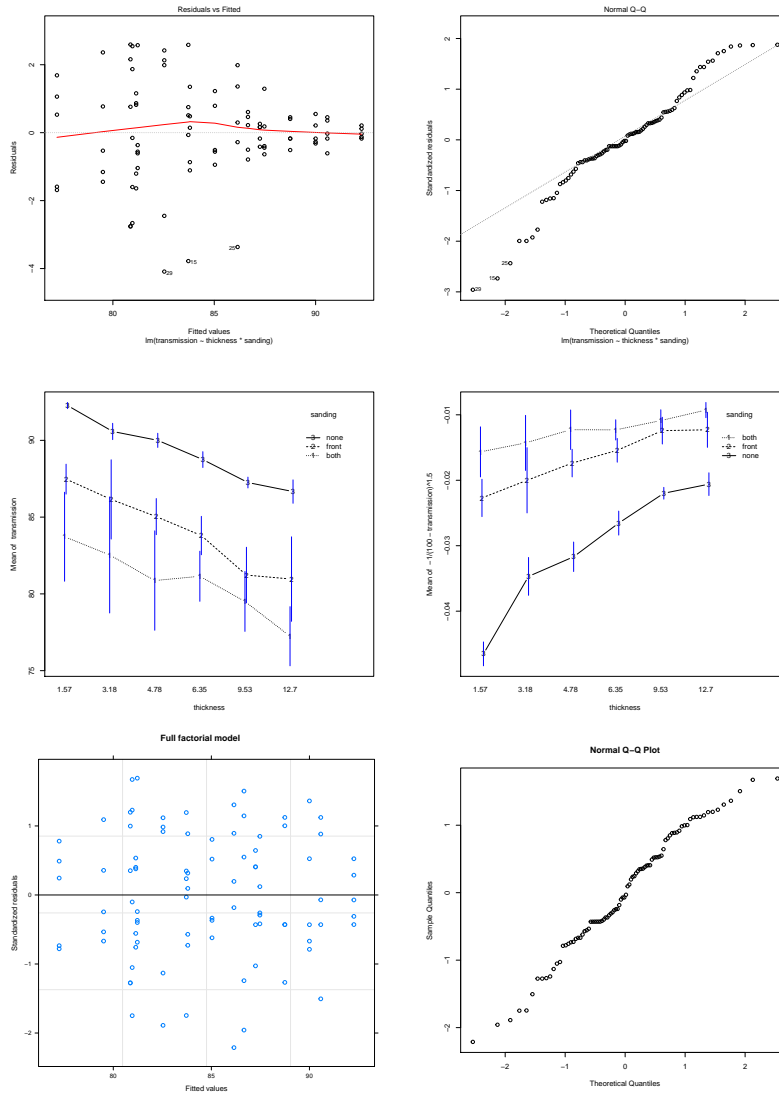
Figure 8.7: Residual and interaction plots for the laser transmission data. Panels 1–3: residual plots for the (unweighted) full factorial model. Panel 4: interaction plot with individual treatment confidence intervals. Panels 5–6: residual plots for the full factorial linear model with separate variances fit by level of sanding.

we could look at reflectance as 100 minus the transmission. The ratio of maximum to minimum for reflectance is about 3.25, so a power transformation has a chance to fix things.

A Box-Cox analysis of the full factorial model for reflectance suggests a

power of –1.5. Line 4 produces the interaction plot on this scale, as shown in panel 4 of Figure 8.7. Comparing panels 3 and 4, we note two things. First, the variance is much more stable with the new response. Second, what looked to be a potentially additive model on the original scale in panel 3 is now most definitely has interaction on the new scale in panel 4. We fixed the non-constant variance, but the new response and the interaction make the model more difficult to interpret.

A second interpretation of panel 3 is that error variability differs by levels of the sanding factor. Instead of transforming the response, explore a model that fits separate error variances by levels of sanding, as done in line 5. Lines 6–7 produce diagnostic plots for the weighted (unequal variance) model, shown in panels 5–6 of Figure 8.7. These plots look much better than what we saw previously.

One can generate an "analysis of variance" for models like `fitw` with fitted weights as shown in line 8 below:

```
8 > anova(fitw)
  Denom. DF: 72
                    numDF   F-value p-value
  (Intercept)         1 1598579.0  <.0001
  thickness           5     156.4  <.0001
  sanding             2     362.2  <.0001
  thickness:sanding  10       0.6  0.7664
```

Analysis of variance is quoted here for a couple of reasons. First, it is not really partitioning variability in the data as we have discussed. However, the tests it generates will be equivalent to F-tests in an ANOVA in the situation where all of the weights in the `gls` call are equal. Second, the tests are computed assuming that the (relative) weights of data in the fit were fixed and known ahead of time. This is not true, so this procedure can underestimate the variability of the parameter estimates and lead to p-values that are smaller than they should be. For these data, however, the results seem fairly clear: on this scale, the interaction is not significant, and both main effects are significant.

The summary of the weighted fit is long (even as abbreviated here), but contains much useful information.

```
) > summary(fitw)
Generalized least squares fit by REML
  Model: transmission ~ thickness * sanding
  Data: LaserTransmission
       AIC      BIC    logLik
  312.9271 360.7371 -135.4636

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | sanding
 Parameter estimates:
    none     front      both
1.000000 3.768814 5.361904

Coefficients:
                         Value Std.Error  t-value p-value
(Intercept)           84.73213 0.1629106 520.1142  0.0000
thickness1             3.08853 0.3642792   8.4785  0.0000
thickness2             1.69027 0.3642792   4.6400  0.0000
thickness3             0.56840 0.3642792   1.5603  0.1231
thickness4            -0.16773 0.3642792  -0.4605  0.6466
thickness5            -2.07060 0.3642792  -5.6841  0.0000
sanding1              -3.89477 0.2803900 -13.8905  0.0000
sanding2              -0.62417 0.2286250  -2.7301  0.0080
thickness1:sanding1   -0.20370 0.6269711  -0.3249  0.7462
thickness2:sanding1    0.00857 0.6269711   0.0137  0.9891
thickness3:sanding1   -0.53777 0.6269711  -0.8577  0.3939
thickness4:sanding1    0.47777 0.6269711   0.7620  0.4485
thickness5:sanding1    0.74163 0.6269711   1.1829  0.2407
thickness1:sanding2    0.27710 0.5112209   0.5420  0.5895
thickness2:sanding2    0.35417 0.5112209   0.6928  0.4907
thickness3:sanding2    0.35723 0.5112209   0.6988  0.4869
thickness4:sanding2   -0.14083 0.5112209  -0.2755  0.7837
thickness5:sanding2   -0.81297 0.5112209  -1.5902  0.1162
...
Residual standard error: 0.4037683
Degrees of freedom: 90 total; 72 residual
```

The output reminds us the model was fit using REML. The estimated error standard deviations by levels of sanding have ratios 1:3.8:5.4, with estimated error standard deviation .40 in the no surface treatment group. We can see that the standard deviations of the estimated effects differ across levels of sanding; this is to be expected with the non-constant error variability. Finally, we had already seen that the 10 degree of freedom interaction effect was not significant, but here we see that none of the interaction effects in this parameterization is individually significant. It is sometimes the case that some individual effects in a term can look significant even when the term as a whole is not significant, particularly in terms with many degrees of freedom.

Thickness is quantitative, so we can consider polynomial (or other functional) models of the factorial mean. These can be very useful, but interpretation of interaction can be challenging. For interactions of a categorical factor and a polynomial term, the interaction coefficients are adjustments to the overall polynomial term. For the interaction of two quantitative factors, we literally get the product of (powers of) the quantitative factors, leading to cross-product terms in a polynomial equation.

```
10 > with(LaserTransmission,interactplot(thickness,sanding,transmission,
       confidence=.95,pooled=FALSE,at=sort(unique(thickness.z))))
11 > fitq5i <- gls(transmission~poly(thickness.z,5)*sanding,data=LaserTransmission,
       weights=varIdent(form=~1|sanding),method='ML')
12 > summary(fitq5i)
   Generalized least squares fit by maximum likelihood
     Model: transmission ~ poly(thickness.z, 5) * sanding
     Data: LaserTransmission
         AIC      BIC     logLik
     294.4463 346.9423 -126.2231

   Variance function:
    Structure: Different standard deviations per stratum
    Formula: ~1 | sanding
    Parameter estimates:
       none     front      both
    1.000000 3.768813 5.361904

   Coefficients:
                                   Value Std.Error  t-value p-value
    (Intercept)                  84.73213 0.1629106 520.1142  0.0000
    poly(thickness.z, 5)1       -19.79399 1.5455059 -12.8074  0.0000
    poly(thickness.z, 5)2         2.80514 1.5455059   1.8150  0.0737
    poly(thickness.z, 5)3        -0.00772 1.5455059  -0.0050  0.9960
    poly(thickness.z, 5)4         0.81376 1.5455059   0.5265  0.6001
    poly(thickness.z, 5)5         0.40277 1.5455059   0.2606  0.7951
    sanding1                     -3.89477 0.2803900 -13.8905  0.0000
    sanding2                     -0.62417 0.2286249  -2.7301  0.0080
    poly(thickness.z, 5)1:sanding1   0.44319 2.6600134   0.1666  0.8681
    poly(thickness.z, 5)2:sanding1  -2.56537 2.6600134  -0.9644  0.3381
    poly(thickness.z, 5)3:sanding1  -2.88885 2.6600134  -1.0860  0.2811
    poly(thickness.z, 5)4:sanding1  -0.37078 2.6600134  -0.1394  0.8895
    poly(thickness.z, 5)5:sanding1   2.21932 2.6600134   0.8343  0.4069
    poly(thickness.z, 5)1:sanding2  -2.43086 2.1689266  -1.1208  0.2661
    poly(thickness.z, 5)2:sanding2   1.40709 2.1689266   0.6487  0.5186
    poly(thickness.z, 5)3:sanding2   2.64156 2.1689266   1.2179  0.2272
    poly(thickness.z, 5)4:sanding2   0.28761 2.1689266   0.1326  0.8949
    poly(thickness.z, 5)5:sanding2  -0.47669 2.1689266  -0.2198  0.8267
    ...
    Residual standard error: 0.3611414
    Degrees of freedom: 90 total; 72 residual
```

Line 10 above does an interaction plot with a twist. In line 10, we request
that the horizontal factor be positioned at its actual values rather than simply
equally spaced in order. Panel 1 of Figure 8.8 shows this plot. Positioned in
this way, the traces show a little bit of positive curvature (convexity). Thus
we should not be surprised if we find that we need quadratic or higher order
terms to describe the response.

There are six levels of thickness, so we can fit polynomials up to order
5 in thickness (`thickness.z` is a non-factor quantitative version of thick-
ness). We do this in line 11. Note that we asked to use maximum likelihood
rather than REML. We need ML to compare different mean structures. Line
12 gives us summary information. Again, there are several things to note.
First, the variance ratios are the same in line 12 as in line 9; the parameter-
ization of the mean did not change our estimate of error variance. Second,
the estimate of error standard deviation is lower in line 12 (.3611) than in

line 9 (.4038). We expect REML standard errors to be larger, because REML adjusts for degrees of freedom in the mean model and ML does not. In fact, $\sqrt{90/72.3611} = .4038$. Third, none of the interaction coefficients is statistically significant in this parameterization either. Finally, it does not look like we need any powers higher than 2.

Given that interaction never seems to be significant, let's refit a model that leaves out interaction. We do this in line 13, with the summary information in line 14. We see that the error standard deviation ratios change, which is perhaps not surprising given the different mean structure with 10 fewer degrees of freedom. What might be surprising is that the polynomial coefficients changed from line 12 to line 14. If we were using `lm` (that is, equal variances) on these balanced data, then these coefficients would not have changed when the interaction was removed. However, using `gls` (and thus unequal variances) makes the model behave as if the data were unbalanced.

One of the side effects of taking out the interaction is that the p-values for both linear and quadratic are tiny, while those of the higher orders are large. This suggests another refit using only linear and quadratic, as is done in line 15 with summary information in line 16. Note that even though we have orthogonal polynomials, the error variance ratios change slightly when we changed the mean structure, leading to slightly changed polynomial coefficients.

```
13 > fitq5 <- gls(transmission~poly(thickness.z,5)+sanding,data=LaserTransmission,
      weights=varIdent(form=~1|sanding),method='ML')
14 > summary(fitq5)
 Generalized least squares fit by maximum likelihood
   Model: transmission ~ poly(thickness.z, 5) + sanding
   Data: LaserTransmission
        AIC       BIC    logLik
   282.0999 309.5978 -130.0499

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | sanding
 Parameter estimates:
    none    front     both
 1.00000 4.00653 5.63385

Coefficients:
                          Value Std.Error  t-value p-value
(Intercept)             84.73213 0.1617469 523.8562  0.0000
poly(thickness.z, 5)1 -18.10246 0.6301332 -28.7280  0.0000
poly(thickness.z, 5)2   3.87033 0.6301332   6.1421  0.0000
poly(thickness.z, 5)3   0.28561 0.6301332   0.4532  0.6516
poly(thickness.z, 5)4   0.89550 0.6301332   1.4211  0.1591
poly(thickness.z, 5)5  -1.15364 0.6301332  -1.8308  0.0708
sanding1                -3.89477 0.2778821 -14.0159  0.0000
sanding2                -0.62417 0.2279984  -2.7376  0.0076
...
Residual standard error: 0.3631841
Degrees of freedom: 90 total; 82 residual
```

Line 17 does model comparison for our full factorial, additive factorial, and additive second order factorial. We can do this because we used ML. Neither

larger model is a statistically significant improvement on the additive second order model.

The orthogonal polynomials are numerically and statistically stable, but we typically want our polynomial in the original variables for ease of interpretation. We do this on line 18, with summary information on line 19.

```
15 > fitq2 <- gls(transmission~poly(thickness.z,2)+sanding,data=LaserTransmission,
     weights=varIdent(form=~1|sanding),method='ML')
16 > summary(fitq2)
   Generalized least squares fit by maximum likelihood
     Model: transmission ~ poly(thickness.z, 2) + sanding
     Data: LaserTransmission
          AIC      BIC    logLik
     281.5636 301.5621 -132.7818

   Variance function:
    Structure: Different standard deviations per stratum
    Formula: ~1 | sanding
    Parameter estimates:
       none    front     both
   1.000000 3.630014 5.027434

   Coefficients:
                          Value Std.Error  t-value p-value
   (Intercept)          84.73213 0.1580692 536.0445  0.0000
   poly(thickness.z, 2)1 -18.16171 0.6781548 -26.7811  0.0000
   poly(thickness.z, 2)2   3.84827 0.6781548   5.6746  0.0000
   sanding1              -3.89477 0.2701993 -14.4144  0.0000
   sanding2              -0.62417 0.2236554  -2.7908  0.0065
   ...
   Residual standard error: 0.401867
   Degrees of freedom: 90 total; 85 residual
17 > anova(fitq2,fitq5,fitq5i)
        Model df      AIC      BIC    logLik    Test  L.Ratio p-value
   fitq2      1  8 281.5636 301.5621 -132.7818
   fitq5      2 11 282.0999 309.5978 -130.0499 1 vs 2 5.463704  0.1408
   fitq5i     3 21 294.4463 346.9423 -126.2231 2 vs 3 7.653618  0.6626
```

We are fitting the same mean structure as in line 15, but with a different parameterization. Thus the coefficients, including the intercept, can be quite different, as we see here.
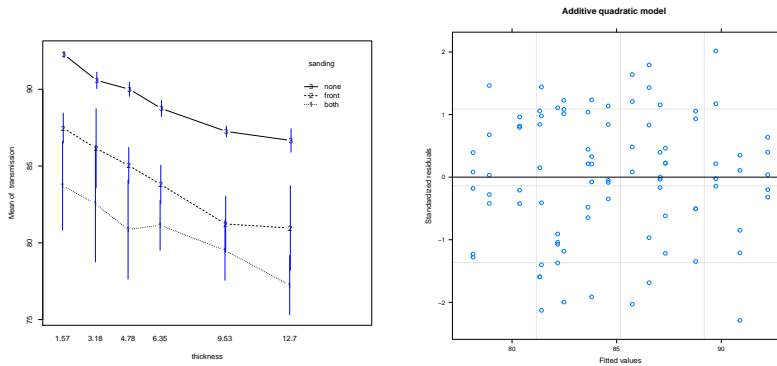
Figure 8.8: Residual and interaction plots for the laser transmission data. Panel 1: interaction plot with horizontal axis placed quantitatively. Panel 2: residual plot for the additive quadratic (weighted) model.

```
8 > fitq2b <- gls(transmission~thickness.z+I(thickness.z^2)+sanding,
    data=LaserTransmission,weights=varIdent(form=~1|sanding),method='ML')
9 > summary(fitq2b)
 Generalized least squares fit by maximum likelihood
   Model: transmission ~ thickness.z + I(thickness.z^2) + sanding
   Data: LaserTransmission
        AIC      BIC    logLik
   281.5636 301.5621 -132.7818

 Variance function:
  Structure: Different standard deviations per stratum
  Formula: ~1 | sanding
  Parameter estimates:
     none    front     both
 1.000000 3.630014 5.027434

 Coefficients:
                   Value  Std.Error   t-value p-value
 (Intercept)     89.16622 0.29242020 304.92497  0.0000
 thickness.z     -0.98436 0.08634724 -11.40005  0.0000
 I(thickness.z^2)  0.03328 0.00586384   5.67462  0.0000
 sanding1        -3.89477 0.27019933 -14.41442  0.0000
 sanding2        -0.62417 0.22365539  -2.79075  0.0065
 ...
 Residual standard error: 0.401867
 Degrees of freedom: 90 total; 85 residual
0 > plot(fitq2b)
```

As a check, we add a residual plot in line 20, which appears in panel 2 of Figure 8.8. Residuals still look good.

Example 8.11 Sprouting of barley seeds under water and time treatments.

Recall the sprouting barley data of Table 8.1. These were the counts of sprouting barley seeds from batches of 100 that were aged a certain number of weeks post-harvest and treated with one of two amounts of water. These are simple counts with a two-way factorial structure.

Lines 1–2 below produce interaction plots for these data, without and with pooled confidence limits, shown in panels 1–2 of Figure 8.9.

```
1 > with(SproutingBarley,interactplot(weeks.z,water,sprouting))
2 > with(SproutingBarley,interactplot(weeks.z,water,sprouting,confidence=.95))
3 > fit.lm <- lm(sprouting~weeks*water,data=SproutingBarley)
4 > plot(fit.lm,which=1)
5 > fit.bin <- glm(cbind(sprouting,100-sprouting)~weeks*water,data=SproutingBarley,
    family=binomial())
6 > plot(fit.bin,which=1)
7 > anova(fit.bin)
  ...
  Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev
  NULL                        29     302.042
  weeks        4  106.491    25     195.551
  water        1  103.401    24      92.150
  weeks:water  4    6.247    20      85.902
```

Even though there is considerable variation, these plots seem to indicate a decreasing effect of water, and increasing effect of time (weeks), and possibly an interaction. Line 3 fits the factorial linear model to these data, with line 4 producing the residuals versus fitted plot (panel 3). There is clear increasing variability.

A little thought reminds us that the response should be a binomial count (X out of 100 sprouting); binomial counts have non-constant variance, so we should not have been surprised. Line 5 refits with a binomial GLM, and line 6 shows the residual plot (panel 4), which is much improved. Feeling good about this model, we look at the "anova" in line 7; this produces the analysis of deviance. We see that adding weeks to a null model uses 4 degrees of freedom to decrease the deviance by 106.5 (highly significant); adding water uses 1 degree of freedom to decrease the deviance by 103.4 (highly significant); and adding the interaction term uses 4 degrees of freedom to decrease the deviance by only 6.2 (not significant at all).

However, look at that residual deviance of 85.9 on 20 degrees of freedom. A chi-square with 20 degrees of freedom is rarely above 40 or 45, so 85.9 is much too large. This indicates that the variability in the data is greater than binomial variability, so a quasibinomial model would be appropriate. We fit this model in line 8.

```
8 > fit.qbin <- glm(cbind(sprouting,100-sprouting)~weeks*water,data=SproutingBarley,
     family=quasibinomial())
9 > summary(fit.qbin)
...
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.02946    0.13013 -15.595 1.17e-12 ***
weeks1         -0.65652    0.29700  -2.210 0.038889 *
weeks2         -0.54070    0.30326  -1.783 0.089781 .
weeks3          0.12274    0.24052   0.510 0.615417
weeks4          0.16948    0.24554   0.690 0.497976
water1          0.57027    0.13013   4.382 0.000288 ***
weeks1:water1  -0.23932    0.29700  -0.806 0.429847
weeks2:water1   0.12809    0.30326   0.422 0.677251
weeks3:water1   0.01153    0.24052   0.048 0.962247
weeks4:water1   0.20880    0.24554   0.850 0.405175
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasibinomial family taken to be 4.060115)
...
```

Line 9 shows the summary. The dispersion parameter is about 4, so the standard errors for model effects are about twice what they were in the binomial model.

Weeks is a quantitative variable (so is water, but there are only two levels of water so it makes little difference whether you consider water to be quantitative or qualitative: it's the same degree of freedom either way), so we can refit with a polynomial model as shown in line 10.

```
10 > fit.qbin <- glm(cbind(sprouting,100-sprouting)~poly(weeks.z,4)*water,
     data=SproutingBarley,family=quasibinomial())
11 > summary(fit.qbin)
...
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -2.02946    0.13013 -15.595 1.17e-12 ***
poly(weeks.z, 4)1          2.99205    0.69676   4.294 0.000354 ***
poly(weeks.z, 4)2          0.21977    0.66995   0.328 0.746293
poly(weeks.z, 4)3          0.27173    0.75205   0.361 0.721654
poly(weeks.z, 4)4          0.66536    0.72954   0.912 0.372619
water1                     0.57027    0.13013   4.382 0.000288 ***
poly(weeks.z, 4)1:water1   0.21636    0.69676   0.311 0.759374
poly(weeks.z, 4)2:water1  -0.67632    0.66995  -1.010 0.324786
poly(weeks.z, 4)3:water1   0.01281    0.75205   0.017 0.986577
poly(weeks.z, 4)4:water1  -0.52107    0.72954  -0.714 0.483331
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasibinomial family taken to be 4.060115)
...
```

The summary in line 11 shows an effect of water, a linear effect of weeks, and nothing else. It's actually a very simple model in the end.

The interaction plots appeared to show some interaction and some curvature, but neither of these appears in our final model. Why? The linear model is in the logit scale, and that gets transformed back to the probability scale
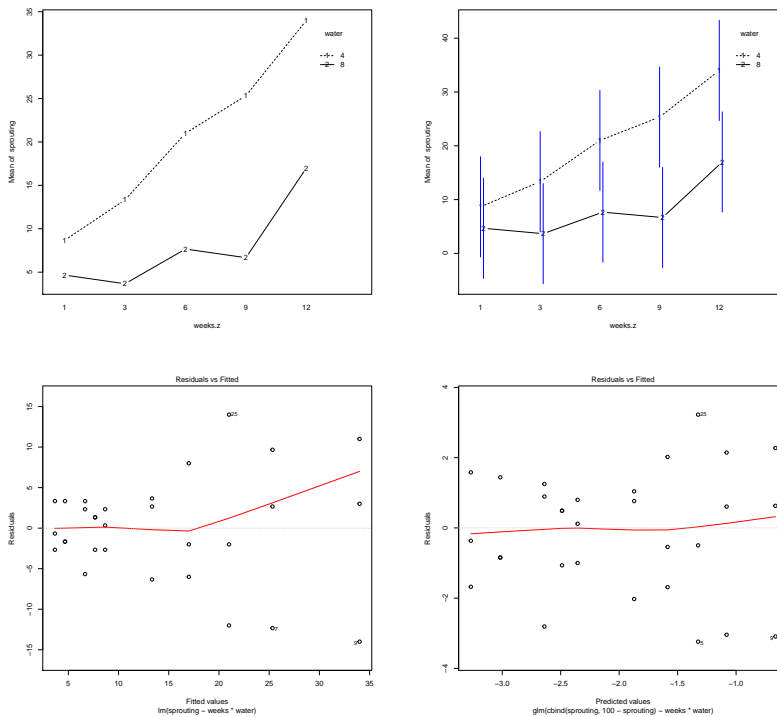
Figure 8.9: Residual and interaction plots for the sprouting barley data. Panels 1–2: residual versus fitted plots for the linear and binomial GLM models. Panels 3–4: interaction plot with and without (pooled) standard errors.

via $p = \exp(\ell)/(1 + \exp(\ell))$. For a probability less than .25 ($\ell < -1$), going from the linear predictor scale to the probability scale is essentially just exponentiation, because the denominator $1 + \exp(\ell)$ does not change much over that range. Thus linear on the ℓ scale looks curved on the p scale, and additive on the ℓ scale looks multiplicative on the p scale. Scale matters.

8.10   Single Replicates

Some factorial experiments are run with only one unit at each factor-level combination ($n = 1$). Clearly, this will lead to trouble, because we have no degrees of freedom for estimating error. What can we do? At this point, analysis of factorials becomes art as well as science, because you must choose among several approaches and variations on the approaches. None of these approaches is guaranteed to work, because none provides the estimate of pure experimental error that we can get from replication. If we use an approach that has an error estimate that is biased upwards, then we will have a conser-

No estimate of
pure error in
single replicates

vative procedure. Conservative in this context means that the p-value that we compute is generally larger than the true p-value; thus we reject null hypotheses less often than we should and wind up with models with fewer terms than might be appropriate. On the other hand, if we use a procedure with an error estimate that is biased downwards, then we will have a liberal procedure. Liberal means that the computed p-value is generally smaller than the true p-value; thus we reject null hypotheses too often and wind up with models with too many terms.

<div style="text-align: right">Biased estimates of error lead to biased tests</div>

The most common approach is to pool one or more high-order interaction mean squares into an estimate of error; that is, select one or more interaction terms and add their sums of squares and degrees of freedom to get a surrogate error sum of squares and degrees of freedom. If the underlying true interactions are null (zero), then the surrogate error mean square is an unbiased estimate of error. If any of these interactions is nonnull, then the surrogate error mean square tends on average to be a little bigger than error. Thus, if we use a surrogate error mean square as an estimate of error and make tests on other effects, we will have tests that range from valid (when interaction is absent) to conservative (when interaction is present).

<div style="text-align: right">High-order interactions can estimate error</div>

This valid to conservative range for surrogate errors assumes that you haven't peeked at the data. It is very tempting to look at interaction mean squares, decide that the small ones must be error and the large ones must be genuine effects. However, this approach tends to give you error estimates that are too small, leading to a liberal test. It is generally safer to choose the mean squares to use as error before looking at the data, although you can use the pooling rules above as a fallback.

<div style="text-align: right">Data snooping makes MS_E too small</div>

A second approach to single replicates is to use an external estimate of error. That is, we may have run similar experiments before, and we know what the size of the random errors was in those experiments. Thus we might use an MS_E from a similar experiment in place of an MS_E from this experiment. This *might* work, but it is a risky way of proceeding. The reason it is risky is that we need to be sure that the external estimate of error is really estimating the error that we incurred during this experiment. If the size of the random errors is not stable, that is, if the size of the random errors changes from experiment to experiment or depends on the conditions under which the experiment is run, then an external estimate of error will likely be estimating something other than the error of this experiment.

<div style="text-align: right">External estimates of error are possible but risky</div>

A final approach is to use one of the models for interaction described in the next chapter. These interaction models often allow us to fit the bulk of an interaction with relatively few degrees of freedom, leaving the other degrees of freedom for interaction available as potential estimates of error.

<div style="text-align: right">Model interaction</div>

Example 8.12 CPU page faults

This is an old, but fun, data set. Some computers divide memory into pages. When a program runs, it is allocated a certain number of pages of RAM. The program itself may require more pages than were allocated. When this is the case, currently unused pages are stored on disk. From time to time,

Table 8.11: Page faults for a CPU experiment. Data set
`PageFaults`.

			Allocation		
Algorithm	Sequence	Size	1	2	3
1	1	1	32	48	538
		2	53	81	1901
		3	142	197	5689
	2	1	52	244	998
		2	112	776	3621
		3	262	2625	10012
	3	1	59	536	1348
		2	121	1879	4637
		3	980	5698	12880
2	1	1	49	67	789
		2	100	134	3152
		3	233	350	9100
	2	1	79	390	1373
		2	164	1255	4912
		3	458	3688	13531
	3	1	85	814	1693
		2	206	3394	5838
		3	1633	10022	17117

a page stored on disk is needed; this is called a *page fault*. When a page fault
occurs, one of the currently active pages must be moved to disk in order to
make room for the page that must be brought in from disk. The trick is to
choose a "good" page to send out to disk, where "good" means a page that
will not be used soon.

The experiment consists of running different programs on a computer
under different configurations and counting the number of page faults. There
were two paging algorithms to study, and this is the factor of primary interest.
A second factor with three levels was the sequence in which system routines
were initialized. Factor three was the size of the program (small, medium,
or large memory requirements), and factor four was the amount of RAM
memory allocated (large, medium, or small). We expect size, sequence, and
allocation to all affect the number of page faults and probably to interact as
well. We want to learn about the algorithm. Table 8.11 shows the number of
page faults that occurred for each of the 54 combinations.

Before fitting any models, look at the data. There is no replication, so
there is no estimate of pure error, and we will need to use some of the inter-
actions as experimental error. There are two sensible ways to move forward.
First, one can fit all main effects, two-factor, and three-factor interactions
in the model and use the four-factor interaction as a surrogate error. That
has 8 degrees of freedom, which is on the low end of acceptable. A sec-

ond approach would be pool into error all three- and four-factor interactions that involve algorithm, giving us 20 degrees of freedom for surrogate error. This leaves the three-way interaction size:sequence:ram in the model, as our introductory information implied it was likely to be important.

The second thing to notice is that the data range over several orders of magnitude and just look multiplicative. Increasing the program size or changing the allocation seems to double or triple the number of page faults, rather than just adding a constant number. This suggests that a log transformation of the response is advisable, as it will turn a multiplicative data set into a more additive data set.

Transforming the data is what we will do in the end, but let's assume that we had not noticed the multiplicative nature of the data and started with surrogate errors on the original scale.

```
1 > fit2fi <- lm(faults~(alg+ram+size+init)^2+ram:size:init,data=PageFaults)
2 > plot(fit2fi,which=1)
3 > fit3fi <- lm(faults~(alg+ram+size+init)^3,data=PageFaults)
4 > plot(fit3fi,which=1)
5 > boxCox(fit2fi)
6 > boxCox(fit3fi)
7 > fit2fil <- lm(log(faults)~(alg+ram+size+init)^2+ram:size:init,data=PageFaults)
8 > plot(fit2fil,which=1)
9 > fit3fil <- lm(log(faults)~(alg+ram+size+init)^3,data=PageFaults)
0 > plot(fit3fil,which=1)
```

Line 1 fits the model with two-factor interactions plus the three-way interaction ram:size:init, and line 2 plots the residuals versus fitted values plot. Lines 3–4 do the same for the three-factor interaction model, and the two plots are the first two panels of Figure 8.10.

Panel 1 shows a form I call the "flopping fish." Residuals follow a U-shaped pattern, being high on the left and right and low in the middle. On the left (the tail), the variation is smaller, and on the right (the head), the variation is larger. You can only get this shape when you use a model smaller than the full factorial model. The shape could indicate that one of the terms left out of the model is important, but it probably means that we need to transform the response.

You cannot really see the flopping fish in panel 2. A large number of residuals that comprise few degrees of freedom can be surprisingly well behaved, and the flopping fish indication of a needed transformation is often more visible in slightly under-fit models. Box-Cox analysis (not shown) suggests that we should transform to a power slightly less than 0 for the smaller model, and slightly greater than 0 for the full three-factor interaction model. The logarithm seems like a good compromise.

We fit the models to log data and plot the residuals in lines 7–10; these residual plots are panels 3–4 of Figure 8.10. They look quite good for the smaller model and seem to indicate we over-transformed for the 3-way interaction model.

Line 11 produces an ANOVA for the smaller model fit to log data.

```
11 > anova(fit2fil)
   Response: log(faults)
                Df Sum Sq Mean Sq    F value    Pr(>F)
   alg           1  2.502   2.502   742.7233 < 2.2e-16 ***
   ram           2 92.697  46.349 13759.5769 < 2.2e-16 ***
   size          2 41.692  20.846  6188.5247 < 2.2e-16 ***
   init          2 24.639  12.320  3657.3417 < 2.2e-16 ***
   alg:ram       2  0.060   0.030     8.9120  0.001709 **
   alg:size      2  0.022   0.011     3.2974  0.057853 .
   alg:init      2  0.018   0.009     2.6179  0.097749 .
   ram:size      4  0.504   0.126    37.4284 5.074e-09 ***
   ram:init      4  9.510   2.378   705.8462 < 2.2e-16 ***
   size:init     4  0.829   0.207    61.5234 5.897e-11 ***
   ram:size:init 8  1.052   0.132    39.0431 1.505e-10 ***
   Residuals    20  0.067   0.003
```

Most terms are extremely significant, the alg:ram interaction is highly significant, and the other two-way interactions with algorithm are not significant. Line 12 shows the ANOVA for the three-way model.

```
12 > anova(fit3fil)
   Response: log(faults)
                Df Sum Sq Mean Sq    F value    Pr(>F)
   alg           1  2.502   2.502   877.9857 1.824e-09 ***
   ram           2 92.697  46.349 16265.4282 3.654e-15 ***
   size          2 41.692  20.846  7315.5596 8.919e-14 ***
   init          2 24.639  12.320  4323.4054 7.300e-13 ***
   alg:ram       2  0.060   0.030    10.5350  0.005736 **
   alg:size      2  0.022   0.011     3.8979  0.065794 .
   alg:init      2  0.018   0.009     3.0947  0.101042
   ram:size      4  0.504   0.126    44.2447 1.689e-05 ***
   ram:init      4  9.510   2.378   834.3927 1.632e-10 ***
   size:init     4  0.829   0.207    72.7278 2.511e-06 ***
   alg:ram:size  4  0.004   0.001     0.3511  0.836454
   alg:ram:init  4  0.026   0.007     2.2818  0.149073
   alg:size:init 4  0.015   0.004     1.2778  0.354761
   ram:size:init 8  1.052   0.132    46.1535 6.726e-06 ***
   Residuals     8  0.023   0.003
```

Results here are similar to what we say in line 11, with none of the three-way interactions involving algorithm significant. In fact, we only have 8 degrees of freedom for error in this model, and two of the three-way interactions meet our criteria to be pooled into error.

```
13 > fit3filb <- lm(log(faults)~(alg+ram+size+init)^3-alg:ram:size-alg:size:init,
      data=PageFaults)
14 > anova(fit3filb)
 Response: log(faults)
                Df Sum Sq Mean Sq   F value     Pr(>F)
 alg             1  2.502   2.502   967.7896 9.683e-16 ***
 ram             2 92.697  46.349 17929.1216 < 2.2e-16 ***
 size            2 41.692  20.846  8063.8244 < 2.2e-16 ***
 init            2 24.639  12.320  4765.6206 < 2.2e-16 ***
 alg:ram         2  0.060   0.030    11.6126 0.0007664 ***
 alg:size        2  0.022   0.011     4.2966 0.0320944 *
 alg:init        2  0.018   0.009     3.4113 0.0583516 .
 ram:size        4  0.504   0.126    48.7702 9.148e-09 ***
 ram:init        4  9.510   2.378   919.7377 < 2.2e-16 ***
 size:init       4  0.829   0.207    80.1667 2.243e-10 ***
 alg:ram:init    4  0.026   0.007     2.5151 0.0825350 .
 ram:size:init   8  1.052   0.132    50.8743 6.241e-10 ***
 Residuals      16  0.041   0.003
```

Line 13 fits a model with these two terms pooled into error, and line 14 shows the corresponding ANOVA. It is nearly identical to that of line 11, because the models differ by only a single, non-significant term.

We have said that the log transformation made our data more additive. The side-by-side plots produced in lines 15–16 are for original scale and log scale data respectively; the plots are shown in Figure 8.11.

```
15 > sidebyside(fit2fi)
16 > sidebyside(fit2fil)
17 > effects::effect("alg:ram",fit2fil)
 alg*ram effect
     ram
 alg    large   medium    small
   1 4.710091 6.227057 7.989746
   2 5.190324 6.702067 8.325972
```

In general, the interaction coefficients are smaller relative to the main effects on the log scale, implying that the data are more additive. In fact, if you look at the fraction of explained variance in the model that is due to main effects, it is 93% on the log scale and only 68% on the original scale.

Finally, to address the original question about algorithms, the only significant interaction involving algorithm is alg:ram, so we should look at its effects, as shown in line 17. Algorithm 2 produces more page faults, with factors of $\exp(5.19 - 4.71) = 1.62$, $\exp(6.70 - 6.23) = 1.60$, and $\exp(8.33 - 7.99) = 1.41$ across large, medium, and small ram. Thus, algorithm 2 is always worse, but algorithm 1's advantage is not as large when the ram size is small.

8.11   Hierarchy

A factorial model for data is called *hierarchical* if the presence of any term in the model implies the presence of all lower order terms. For example, a

Hierarchical models don't skip terms
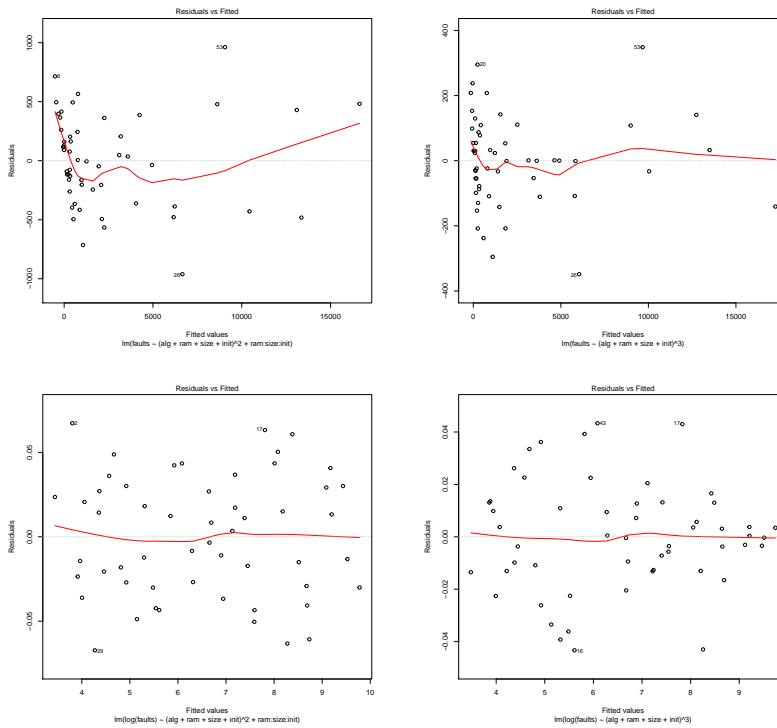
Figure 8.10: Residual plots for the page fault data. Panels 1–2: residuals versus fitted values for data on the original scale using the two-factor interaction plus ram:size:init model and three-factor interaction models. Panels 3–4: residuals versus fitted values for data on the log scale using the two-factor interaction plus ram:size:init model and three-factor interaction models.
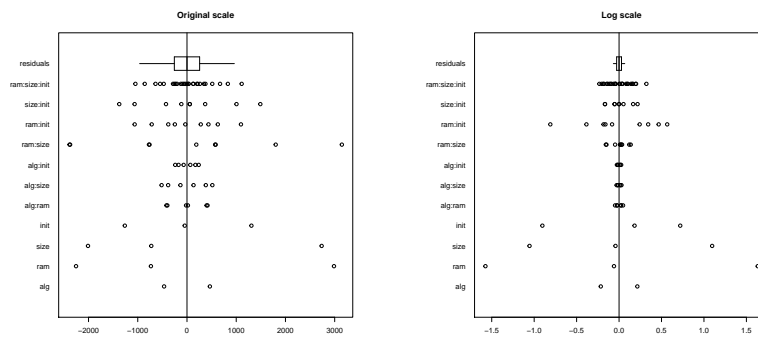
Figure 8.11: Side-by-side plots for the page fault data using the two-factor interaction plus ram:size:init model. Panel 1: original scale; panel 2: log scale.

hierarchical model including the AB interaction must include the A and B main effects, and a hierarchical model including the BCD interaction must include the B, C, and D main effects and the BC, BD, and CD interactions. One potential source of confusion is that lower-order terms occur earlier in a model and thus appear above higher-order terms in the ANOVA table; lower-order terms are above.

One view of data analysis for factorial treatment structure is the selection of an appropriate model for the data; that is, determining which terms are needed, and which terms can be eliminated without loss of explanatory ability. We don't necessarily care how those means are parameterized, we are instead interested in what structures of means are needed. From this point of view, always use hierarchical models when modeling factorial data. Do not automatically test terms above (that is, lower-order to) a needed interaction. If factors A and B interact, conclude that A and B act jointly to influence the response; there is no need to test the A and B main effects.

Choose among hierarchical models

From another point of view, the F-test allows us to test whether any set of parameters is zero, even the main effects of A when the AB interaction is in the model. Why should we not test these lower-order terms, and possibly break hierarchy, when we have the ability to do so? The distinction is one between generic modeling of how the response depends on factors and interactions, and testing specific hypotheses about specific parameters (equivalently, testing specific contrasts in the treatment means). Tests of main effects are tests that certain very specific contrasts are zero. If those specific contrasts are genuinely of interest, then testing main effects is appropriate, even if interactions exist. Thus I only consider nonhierarchical models when I know that the main-effects contrasts, and thus the nonhierarchical model, make sense in the experimental context.

Building a model versus testing hypotheses

The problem with breaking hierarchy is that parameters depend on how we have defined our parameters. We have defined main effect and interaction parameters in a sensible way, but it is also a completely arbitrary way. There are other sensible ways to define these parameters that lead to different values. Thus for the same set of data, I might conclude that there is a main effect of A, and you might conclude that there is not a main effect of A, and we could both be correct if we are using different definitions of main effects.

We have chosen to define main effects and interactions using equally weighted averages of treatment means, but we could instead define main effects and interactions using unequally weighted averages. This new set of main effects and interactions is just as valid mathematically as our usual set, but one set may have zero main effects and the other set have nonzero main effects. Which do we want to test? We need to know the appropriate set of weights, or equivalently, the appropriate contrast coefficients, for the problem at hand.

Are equally weighted averages appropriate?

Example 8.13 Unequal weights

Suppose that we have a three by two factorial design testing two antibiotics against three strains of bacteria. The response is the number of rats (out

Table 8.12: Number of rats that died after exposure to three strains of bacteria and treatment with one of two antibiotics, and factorial decompositions using equal weighting and 1,2,1 weighting of rows.

Means		Equal Weights			Row Weighted		
120	168	-24	24	-8	-21	21	-9
144	168	-12	12	4	-9	9	3
192	120	36	-36	4	39	-39	3
		0	0	152	-3	3	153

of 500) that die from the given infection when treated with the given antibiotic. Our goal is to find the antibiotic with the lower death rate. Table 8.12 gives hypothetical data and two ways to decompose the means into grand mean, row effects, column effects, and interaction effects.

The first decomposition in Table 8.12 (labeled equal weights) is our usual factorial decomposition. The row effects and column effects add to zero, and the interaction effects add to zero across any row or column. With this standard factorial decomposition, the column (antibiotic) effects are zero, so there is no average difference between the antibiotics.

On the other hand, suppose that we knew that strain 2 of bacteria was twice as prevalent as the other two strains. Then we would probably want to weight row 2 twice as heavily as the other rows in all averages that we make. The second decomposition uses 1,2,1 row weights; all these factorial effects are different from the equally weighted effects. In particular, the antibiotic effects change, and with this weighting antibiotic 1 has a mean response 6 units lower on average than antibiotic 2 and is thus preferred to antibiotic 2. The test of no antibiotic effect in the equally weighted example is a test of

$$\mu_{11} + \mu_{21} + \mu_{31} - \mu_{12} - \mu_{22} - \mu_{32} = 0 \quad,$$

and the test of no antibiotic effect in the row-weighted example is a test of

$$\mu_{11} + 2\mu_{21} + \mu_{31} - \mu_{12} - 2\mu_{22} - \mu_{32} = 0.$$

Either, or neither, of these could be what you really want, but the answer to whether there are non-zero column effects clearly depends on how you do the decomposition.

Analogous examples have zero column effects for weighted averages and nonzero column effects in the usual decomposition. Note in the weighted decomposition that column effects add to zero and the interactions add to zero across columns, but row effects and interaction effects down columns only add to zero with 1,2,1 weights.

If factors A and B do not interact, then the A and B main effects are the same regardless of how we weight the means. In the absence of AB interaction, testing the main effects of A and B computed using our equally weighted averages gives the same results as for any other weighting. Similarly, if there is no ABC interaction, then testing AB, AC, or BC using the standard ANOVA gives the same results as for any weighting.

Weighting matters due to interaction

Table 8.13: Amylase specific activity (IU), for two varieties of sprouted maize under different growth and analysis temperatures (degrees C). Data from Orman (1986); data set `AmylaseActivity.`

GT	Var.	Analysis Temperature							
		40	35	30	25	20	15	13	10
25	B73	391.8	427.7	486.6	469.2	383.1	338.9	283.7	269.3
		311.8	388.1	426.6	436.8	408.8	355.5	309.4	278.7
		367.4	468.1	499.8	444.0	429.0	304.5	309.9	313.0
	O43	301.3	352.9	376.3	373.6	377.5	308.8	234.3	197.1
		271.4	296.4	393.0	364.8	364.3	279.0	255.4	198.3
		300.3	346.7	334.7	386.6	329.2	261.3	239.4	216.7
13	B73	292.7	422.6	443.5	438.5	350.6	305.9	319.9	286.7
		283.3	359.5	431.2	398.9	383.9	342.8	283.2	266.5
		348.1	381.9	388.3	413.7	408.4	332.2	287.9	259.8
	O43	269.7	380.9	389.4	400.3	340.5	288.6	260.9	221.9
		284.0	357.1	420.2	412.8	309.5	271.8	253.6	254.4
		235.3	339.0	453.4	371.9	313.0	333.7	289.5	246.7

Factorial effects are only defined in the context of a particular weighting scheme for averages. As long as we are comparing hierarchical models, we know that the parameter tests make sense for any weighting. When we test lower-order terms in the presence of an including interaction, we must use the correct weighting.

Use correct weighting

R is fairly fanatical about enforcing hierarchy when fitting factorial models. For example, if you fit the model `y~ A+A:B`, then the `A:B` term has a(b-1) degrees of freedom instead of (a-1)(b-1), because **R** will automatically subsume the B main effect into A:B if B is not explicitly in the model. Thus what you get as $\widehat{\alpha\beta}_{ij}$ in the non-hierarchical model is actually equal to $\widehat{\beta}_j + \widehat{\alpha\beta}_{ij}$ from the full hierarchical version of the model.

R enforces hierarchy

Example 8.14 Amylase activity

Orman (1986) studied germinating maize. One of his experiments looked at the amylase specific activity of sprouted maize under 32 different treatment conditions. These treatment conditions were the factorial combinations of analysis temperature (eight levels, 40, 35, 30, 25, 20, 15, 13, and 10 degrees C), growth temperature of the sprouts (25 or 13 degrees C), and variety of maize (B73 or Oh43). There were 96 units assigned at random to these 32 treatments. Table 8.13 gives the amylase specific activities in International Units.

This is an eight by two by two factorial with replication, so we fit the full
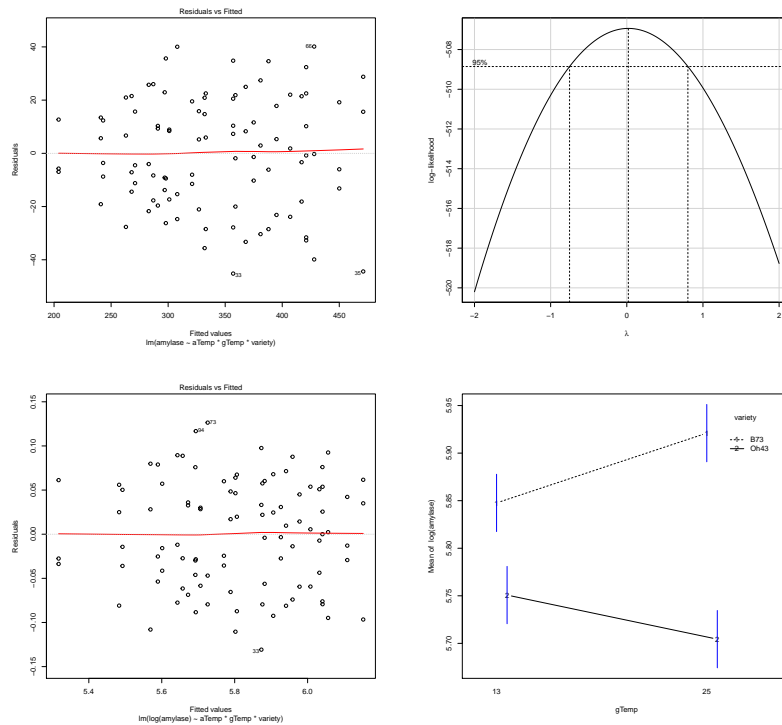
Figure 8.12: Plots for amylase activity data. Panel 1: residuals versus fitted on the original scale. Panel 2: Box-Cox profile for the factorial model fit on the original scale. Panel 3: residuals versus fitted on the log scale. Panel 4: interaction plot of growth temperature by variety.

factorial model.

```
1 > fit <- lm(amylase~aTemp*gTemp*variety,data=AmylaseActivity)
2 > plot(fit,which=1)
3 > boxCox(fit)
4 > fit2 <- lm(log(amylase)~aTemp*gTemp*variety,data=AmylaseActivity)
5 > plot(fit2,which=1)
6 > anova(fit2)
  Response: log(amylase)
                       Df  Sum Sq Mean Sq  F value      Pr(>F)
  aTemp                 7 3.01613 0.43088  78.8628 < 2.2e-16 ***
  gTemp                 1 0.00438 0.00438   0.8016 0.3739757
  variety               1 0.58957 0.58957 107.9085 2.305e-15 ***
  aTemp:gTemp           7 0.08106 0.01158   2.1195 0.0539203 .
  aTemp:variety         7 0.02758 0.00394   0.7212 0.6543993
  gTemp:variety         1 0.08599 0.08599  15.7392 0.0001863 ***
  aTemp:gTemp:variety   7 0.04764 0.00681   1.2457 0.2916176
  Residuals            64 0.34967 0.00546
7 > with(AmylaseActivity,interactplot(gTemp,variety,log(amylase),confidence=.95,
     sigma2=.00546,df=64))
```

Line 1 fits the factorial model on the original scale, and line 2 shows that there is increasing variance (panel 1 of Figure 8.12). The Box-Cox profile (panel 2) suggests a log transformation, and the residual variability look better after the transformation (line 5, panel 3). (Not shown here, but the normal probability plot of residuals looks better on the original scale.)

Analyzing on the log scale, the ANOVA is shown on line 6. The growth temperature by variety interaction is highly significant, but the main effect of growth temperature is not significant. Nevertheless, we retain the main effect of growth temperature to maintain hierarchy.

What does an interaction with no main effect look like? The interaction plot from line 7 is shown in panel 4. The change in response going from growth temperature 13 to 15 is positive for variety 1 and negative for variety 2 (the interaction). However, the positive and negative changes cancel each other, so the means are roughly the same at the two levels of growth temperature (no main effect).

8.12   Problems

Diet affects weight gain. We wish to compare nine diets; these diets are the factor-level combinations of protein source (beef, pork, and grain) and number of calories (low, medium, and high). There are eighteen test animals that were randomly assigned to the nine diets, two animals per diet. The mean responses (weight gain) are:

Exercise 8.1

	Calories		
Source	Low	Medium	High
Beef	76.0	86.8	101.8
Pork	83.3	89.5	98.2
Grain	83.8	83.5	86.2

also given in data set `WeightGain`. The mean square for error was 8.75. Analyze these data to determine an appropriate model.

An experiment was conducted to determine the effect of germination time (in days) and temperature (degrees C) on the free alpha amino nitrogen (FAN) content of rice malt. The values shown in the following are the treatment means of FAN with $n = 2$ (data from Aniche and Okafor 1989, data set `RiceMalt`).

Exercise 8.2

| | Temperature | | | | |
Days	22	24	26	28	Row Means
1	39.4	49.9	55.1	59.5	50.98
2	56.4	68.0	76.4	88.8	72.40
3	70.2	81.5	95.6	99.6	86.72
Column Means	55.33	66.47	75.70	82.63	
Grand Mean	70.03				

The total sum of squares was 8097. Draw an interaction plot for these data. Compute an ANOVA table and determine which terms are needed to describe the means.

Manufacturing integrated circuits is an enormously complicated task, as there are many process variables that can be manipulated (thickness of this, width of that, doping level of something else, etc). Generally speaking, many copies of a circuit are put onto a single wafer, which is made all at once. We have an experiment where we are varying two factors, each at two levels. We have 20 wafers and assign each of the four factor/level combinations to five wafers at random and make the wafer. We then choose three circuits at random on each wafer and measure the performance of the circuit. Construct a skeleton ANOVA table (that is, just the sources and degrees of freedom).

Exercise 8.3

We have a 2^2 factorial design with $N = 17$. The (corrected) total sum of squares is 100. Below are three different ANOVAs for these data. The first pools the main effects and interaction into a single 3 degree of freedom term. The others are sequential. Fill in the missing sums of squares.

Exercise 8.4

Source	DF	SS
A.B	3	75
Error	13	?

Source	DF	SS
A	1	20
B	1	?
A.B	1	20
Error	13	?

Source	DF	SS
B	1	35
A	1	?
A.B	1	?
Error	13	?

Problem 8.1 Particleboard is made from wood chips and resins. An experiment is conducted to study the effect of using slash chips (waste wood chips) along with standard chips. The researchers make eighteen boards by varying the target density (42 or 48 lb/ft^3), the amount of resin (6, 9, or 12%), and the fraction of slash (0, 25, or 50%). The response is the actual density of the boards produced (lb/ft^3, data from Boehner 1975, data set `ParticleBoard`). Analyze these data to determine the effects of the factors on particleboard density and how the density differs from target.

	42 Target			48 Target		
Resin	0%	25%	50%	0%	25%	50%
6	40.9	41.9	42.0	44.4	46.2	48.4
9	42.8	43.9	44.8	48.2	48.6	50.7
12	45.4	46.0	46.2	49.9	50.8	50.3

Problem 8.2 We have data from a four by three factorial with 24 units. Below are ANOVA tables for the data and log-transformed data, plus residual versus predicted plots for both. What would you conclude about interaction in the data?

Original data:

```
Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
r          3 1136.65  378.88 35.6135 2.971e-06 ***
c          2  445.80  222.90 20.9518 0.0001217 ***
r:c        6  215.90   35.98  3.3823 0.0343903 *
Residuals 12  127.66   10.64
```

```
Response: log(y)
          Df Sum Sq Mean Sq F value    Pr(>F)
r          3 7.1393 2.37977 39.2914 1.751e-06 ***
c          2 3.4291 1.71454 28.3081 2.861e-05 ***
r:c        6 0.3290 0.05483  0.9052    0.5224
Residuals 12 0.7268 0.06057
```
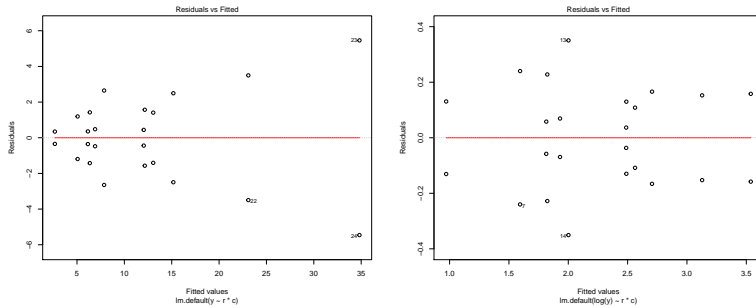
Implantable heart pacemakers contain small circuit boards called substrates. These substrates are assembled, cut to shape, and fired. Some of the substrates will separate, or delaminate, making them useless. The purpose of this experiment was to study the effects of three factors on the rate of delamination. The factors were A: firing profile time, 8 versus 13 hours with the theory suggesting 13 hours is better; B: furnace airflow, low versus high, with theory suggesting high is better; and C: laser, old versus new, with theory suggesting new cutting lasers are better.

Problem 8.3

A large number of raw, assembled substrates are divided into sixteen groups. These sixteen groups are assigned at random to the eight factor-level combinations of the three factors, two groups per combination. The substrates are then processed, and the response is the fraction of substrates that delaminate. Data from Todd Kerkow, data set `PacemakerDelamination`.

	8 hrs		13 hrs	
	Low	High	Low	High
Old	.83	.68	.18	.25
	.78	.90	.16	.20
New	.86	.72	.30	.10
	.67	.81	.23	.14

Analyze these data to determine how the treatments affect delamination.

Pine oleoresin is obtained by tapping the trunks of pine trees. Tapping is done by cutting a hole in the bark and collecting the resin that oozes out. This experiment compares four shapes for the holes and the efficacy of acid treating the holes. Twenty-four pine trees are selected at random from a plantation, and the 24 trees are assigned at random to the eight combinations of hole shape (circular, diagonal slash, check, rectangular) and acid treatment (yes or no). The response is total grams of resin collected from the hole (data from Low and Bin Mohd. Ali 1985, data set `PineOleoresin`).

Problem 8.4

| | | Shape | | |
Treatment	Circ.	Diag.	Check	Rect.
Control	9	43	60	77
	13	48	65	70
	12	57	70	91
Acid	15	66	75	97
	13	58	78	108
	20	73	90	99

Analyze these data to determine how the treatments affect resin yield.

Problem 8.5

A study looked into the management of various tropical grasses for improved production, measured as dry matter yield in hundreds of pounds per acre over a 54-week study period. The management variables were height of cut (1, 3, or 6 inches), the cutting interval (1, 3, 6, or 9 weeks), and amount of nitrogen fertilizer (0, 8, 16, or 32 hundred pounds of ammonium sulfate per acre per year). Forty-eight plots were assigned in completely randomized fashion to the 48 factor-level combinations. Dry matter yields for the plots are shown in the table below (data from Richards 1965, data set `TropicalGrasses`). Analyze these data and write your conclusions in a report of at most two pages.

| | | | Interval | | |
Ht.	Fert.	1	3	6	9
1	0	74.1	65.4	96.7	147.1
	8	87.4	117.7	190.2	188.6
	16	96.5	122.2	197.9	232.0
	32	107.6	140.5	241.3	192.0
3	0	61.7	83.7	88.8	155.6
	8	112.5	129.4	145.0	208.1
	16	102.3	137.8	173.6	203.2
	32	115.3	154.3	211.2	245.2
6	0	49.9	72.7	113.9	143.4
	8	92.9	126.4	175.5	207.5
	16	100.8	153.5	184.5	194.2
	32	115.8	160.0	224.8	197.5

Problem 8.6

Big sagebrush is often planted in range restoration projects. An experiment is performed to determine the effects of storage length and relative humidity on the viability of seeds. Sixty-three batches of 300 seeds each are randomly divided into 21 groups of three. These 21 groups each receive a different treatment, namely the combinations of storage length (0, 60, 120, 180, 240, 300, or 360 days) and storage relative humidity (0, 32, or 45%). After the storage time, the seeds are planted, and the response is the percentage of seeds that sprout (data from Welch 1996, data set `BigSagebrush`). Analyze these data for the effects of the factors on viability.

Humidity	0	60	120	Days 180	240	300	360
0%	82.1	78.6	79.8	82.3	81.7	85.0	82.7
	79.0	80.8	79.1	75.5	80.1	87.9	84.6
	81.9	80.5	78.2	79.1	81.1	82.1	81.7
32%	83.1	78.1	80.4	77.8	83.8	82.0	81.0
	80.5	83.6	81.8	80.4	83.7	77.6	78.9
	82.4	78.3	83.8	78.8	81.5	80.3	83.1
45%	83.1	66.5	52.9	52.9	52.2	38.6	25.2
	78.9	61.4	58.9	54.3	51.9	37.9	25.8
	81.0	61.2	59.3	48.7	48.8	40.6	21.0

Problem 8.7

Everyone likes microwave popcorn, but nobody likes unpopped kernels. Thirty-six 3 oz. bags of microwave popcorn are popped, and the number of unpopped kernels in each bag is recorded. The 36 runs are randomly assigned to the twelve combinations of microwave wattage (500, 700, or 1000) and brand ("P", "A", "J", or "O"). Ovens are allowed to cool for 10 minutes with the door open between runs, and popping continues until there is a three second gap between consecutive pops. Data from P. Stenberg (data set `UnpoppedKernels`).

Brand	1000			700			500		
A	40	33	35	11	8	5	32	23	23
J	36	19	24	5	9	7	24	37	24
O	36	40	21	6	9	8	22	26	12
P	19	20	22	7	9	8	13	19	15

Analyze these data to determine the effects of brand and wattage. What surprising effect do you find?

Problem 8.8

Everyone likes old-style non-microwave popcorn, but what is the best recipe to get that light, fluffy popcorn? This experiment looks at the volume ratio of popped popcorn to unpopped kernels. Thirty-two batches of popcorn are produced, two each for the combinations of popcorn kernel amount (1/8 or 1/4 cup), popcorn type (generic or gourmet), oil type (canola or "popcorn" oil), and oil amount (1 or 2 tablespoons). The following table shows the results (data from J. McLaren, pers. comm., data set `PopcornRatios`).

Pop. Type	Pop. Amt.	Oil Amt.	Oil Type Canola		Popcorn	
Generic	1/8	1	24.5	24.5	28	21
		2	21.5	20	22.5	17.5
	1/4	1	21.5	23.5	23	24
		2	22.25	24.5	24.75	22.5
Gourmet	1/8	1	17.5	20.5	20.5	17.5
		2	18	18.5	16	17
	1/4	1	14.5	16.25	19.25	22.75
		2	21	20.25	18.25	19.25

What factors and/or interactions influence the ratios? Analyze these data and report your results.

Problem 8.9

In what ways can we analyze the page fault data from Example 8.12 without transformation and still have a valid analysis? How do the conclusions from these other analyses compare with those from the example?

Problem 8.10

Tablets delivering medicines orally must dissolve for the dose to be delivered. Generally, we want tablets that will be hard (so they do not break) but dissolve quickly. Pharmaceutical companies can vary the pressure used to compress the ingredients into a tablet (with higher pressures assumed to create harder tablets) and can add a "disintegrant" to the formulation to shorten the time until dissolution (a disintegrant is supposed to help the tablet break up into smaller pieces when it encounters water).

In this study, 16 batches of tablets will be manufactured, with two batches assigned at random to the factor level combinations of pressure (10 or 20 foot pounds) and disintegrant (2, 4, 6, or 8% by weight). After manufacture, six tablets are selected at random from each batch, and the response is the time until dissolution when the table is placed in a Vanderkamp Disintegration Tester (basically, it swishes the tablet around in body-temperature water). In total, we have 96 dissolution time measurements.

Create a "skeleton ANOVA" for this experiment. This is an abbreviated ANOVA table including only the sources and degrees of freedom.

Problem 8.11

Bacterial resistance to antibiotics is a concern, and bacteria may be adapting to non-zero levels of antibiotics in sewage. We will take some sewage, treat it with some kind of antibiotic, grow it on a plate, and then count the bacteria that survive. Forty-eight preparations were made, three each for the combinations of origin of sample (activated sludge or effluent), growth medium (LB or CAGY), and antibiotic (Amoxicillin, Tetracycline, Tylosin, or none). Bacterial counts (per μL of sample) are in the following table (data adapted from S. Ghosh, pers. comm., data set `SewageBacteria`).

Origin	Antibiotic	LB			CAGY		
Sludge	Amox.	760000	440000	330000	153000	188000	182000
	Tetra.	17000	11000	21000	72000	65000	67000
	Tylosin	620000	1380000	540000	600000	400000	860000
	None	2150000	1680000	1660000	306000	273000	213000
Effluent	Amox.	141	162	168	118	123	150
	Tetra.	15.6	8.3	7.6	10.6	13.2	11.1
	Tylosin	210	220	260	112	153	131
	None	2900	1420	1440	249	290	286

(a) Analyze these data to determine the effects of the factors on bacterial counts.
(b) There is something truly bizarre in these data; what is it?

Problem 8.12

The dye Rhodomine 6G can be adsorbed by activated carbon beads incorporated with calcium alginate. The experiment studies how three factors

affect the percentage of dye adsorbed. The factors are initial concentration of dye (100, 200, or 300 mg/l), pH (7, 8, or 9), and temperature (30 or 60 degrees C). Eighteen units were randomly assigned to the factor/level combinations, and the adsorption is shown in the table below (data originally from Annadurai, Juang, and Lee 2002 via Lye 2019, data set `DyeAdsorption`).

	Temperature/pH					
	30°			30°		
dye	7	8	9	7	8	9
100	98.5	98.7	99.2	99.9	100.0	100.0
200	96.7	97.0	97.3	98.2	98.6	98.2
300	94.8	95.3	95.6	96.4	96.7	97.1

Analyze these data to determine the effects of the factors.

Problem 8.13

Using a cell phone while driving affects reaction time, but is it just using a cell phone, or do other factors affect reaction time? In this study, 54 male adults aged 22–24 years were asked to drive in a simulator under various conditions, and their reaction times would be measured. At some point during the simulation, the drivers begin talking on the cell phone. After a certain time on the phone has elapsed, the "car" ahead of the driver hits the brakes, and the time between the leading car hitting the brakes and the driver in the simulator hitting the brakes is the measured response.

The 54 individuals were randomized to 18 conditions, which are the factor/level combinations of: meters (10, 15, or 20; the trailing distance behind the leading car), conditions (day or night driving), and duration (30, 60, or 90 seconds of cell phone conversation before the leading car hits the brakes). The response is the reaction time (in milli seconds). The data are in the table below (adapted from Al-Darrab, Khan, and Ishrat 2009 via Lye 2019, data set `ReactionTimes`).

		Distance								
Cond.	Dur.	10			15			20		
Day	30	90	250	230	200	150	220	70	90	180
	60	120	180	460	150	180	630	170	210	900
	90	80	90	200	120	60	150	120	70	130
Night	30	350	300	190	120	150	710	140	120	760
	60	100	180	180	90	210	170	110	110	120
	90	200	90	290	270	120	590	120	150	710

Analyze these data to determine the effects of the factors on response time.

Problem 8.14

The dye methylene blue may be removed from aqueous solution via an oxidation reaction with persulfate. This experiment studies how the percentage of dye removed varies with reaction time (5, 10, 15, 20, or 25 minutes), persulfate concentration (355, 710, or 1065 mg/l), the initial dye concentration (10, 15, or 20 mg/l), and the process temperature (60 or 70 degrees C). The response is the color removal efficiency (CRE in precentage). There

are 90 factor/level combinations, each observed once (you may assume a completely randomized design). The data are in the table below (data from Kordkandi and Forouzesh 2014 via Lye 2019, data set `DyeRemoval`).

		Persulfate/Dye								
		355			710			1065		
Temp	Time	10	15	20	10	15	20	10	15	20
60	5	14.6	12.4	10.6	17.8	14.1	12.0	20.5	16.0	10.5
	10	26.1	18.8	17.3	28.7	22.4	16.6	33.0	27.2	20.0
	15	34.3	25.9	21.4	37.1	29.3	22.1	43.1	35.1	25.7
	20	39.6	27.6	23.8	45.3	35.0	25.7	51.7	42.8	31.6
	25	44.0	30.4	26.1	50.0	38.3	29.5	60.0	49.9	36.3
70	5	33.5	20.4	19.5	40.6	29.3	27.1	48.6	33.5	30.6
	10	48.5	33.3	29.2	66.2	43.3	39.0	74.4	56.4	51.6
	15	59.4	38.8	33.7	81.1	52.2	51.8	89.7	71.8	65.2
	20	65.4	43.2	38.5	88.5	60.1	59.3	95.9	82.4	74.6
	25	70.6	46.8	44.1	92.6	67.1	66.1	99.1	88.6	81.5

Analyze these data to determine the influential effects on dye removal.

Question 8.1

Consider a balanced four by three factorial. Show that orthogonal contrasts in row means (ignoring factor B) are also orthogonal contrasts for all twelve treatments when the contrast coefficients have been repeated across rows ($w_{ij} = w_i$). Show that a contrast in the row means and the analogous contrast in all twelve treatment means have the same sums of squares.

Question 8.2

In a two-way factorial, we have defined $\widehat{\mu}$ as the grand mean of the data, $\widehat{\mu} + \widehat{\alpha}_i$ as the mean of the responses for the ith level of factor A, $\widehat{\mu} + \widehat{\beta}_j$ as the mean of the responses for the jth level of factor B, and $\widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j + \widehat{\alpha\beta}_{ij}$ as the mean of the ijth factor-level combination. Show that this implies our zero-sum restrictions on the estimated effects.

Suppose that we use the same idea, but instead of ordinary averages we use weighted averages with v_{ij} as the weight for the ijth factor-level combination. Derive the new zero-sum restrictions for these weighted averages.

Chapter 9

Further Topics in Factorials

This chapter addresses some additional key topics that users of factorials should know. Some of these are simple generalizations of ideas from non-factorial data, but others are new. Here we look at:

- Power and sample size for factorials.
- Analysis of unbalanced factorials.
- Contrasts and multiple comparisons for factorials.
- Models for interaction.
- Tools for two-series factorials.

9.1   Power and Sample Size

Chapter 7 described the computation of power and sample size for completely randomized designs. If we ignore the factorial structure and consider our treatments simply as g treatments, then we can use the methods of Chapter 7 to compute power and sample size for the overall null hypothesis of no model effects. Recall that power depends on the Type I error rate $\mathcal{E}_I$, numerator and denominator degrees of freedom, and the effects, sample sizes, and error variance through the noncentrality parameter.

For factorial data, we usually test null hypotheses about main effects or interactions in addition to the overall null hypothesis of no model effects. Power for these tests again depends on the Type I error rate $\mathcal{E}_I$, numerator and denominator degrees of freedom, and the effects, sample sizes, and error variance through the noncentrality parameter, so we can do the same kinds of power and sample size computations for factorial effects once we identify the degrees of freedom and noncentrality parameters.

We will address power and sample size only for balanced data, because most factorial experiments are designed to be balanced, and simple formulae for noncentrality parameters exist only for balanced data. For concreteness,

we present the formulae in terms of a three-factor design; the generalization to more factors is straightforward. In a factorial, main effects and interactions are tested separately, so we can perform a separate power analysis for each main effect and interaction. The numerator degrees of freedom are simply the degrees of freedom for the factorial effect: for example, $(b-1)(c-1)$ for the BC interaction. Error degrees of freedom $(N - abc)$ are the denominator degrees of freedom for all our tests.

The noncentrality parameter depends on the factorial parameters, sample size, and error variance. The algorithm for a noncentrality parameter in a balanced design is

1. Square the factorial effects and sum them,

2. Multiply this sum by the total number of data in the design divided by the number of levels in the effect, and

3. Divide that product by the error variance.

Noncentrality parameter

For the AB interaction, this noncentrality parameter is

$$\frac{\frac{N}{ab} \sum_{ij} \alpha\beta_{ij}^2}{\sigma^2} = \frac{nc \sum_{ij} \alpha\beta_{ij}^2}{\sigma^2} \, .$$

The factor in step 2 equals the number of data values observed at each level of the given effect. For the AB interaction, there are n values in each treatment, and c treatments with the same ij levels, for a total of nc observations in each ij combination.

As in Chapter 7, minimum sample sizes to achieve a given power can be found iteratively, literally by trying different sample sizes and finding the smallest one that achieves the required power.

Example 9.1 Power for zinc retention

Recall the zinc retention design of Example 8.4. The treatments have a 4 (meal protein) by 2 (meal zinc) by 2 (diet zinc) structure. Assume that we will test at the .01 level and have the following design criteria:

1. We need power .9 for detecting the situation where any two individual levels of a main effect differ by 20.

2. We need power .8 for detecting the situation where any two individual two-factor interaction effects differ by 10.

3. We need power .6 for detecting the situation where any two individual three-factor interaction effects differ by 10.

We believe that the error variance is 50. The smallest sum of squared effects where two main effects differ by 20 is 200 $((-10)^2 + 10^2 + 0)$; the smallest sum of squared effects where two two-way interaction effects differ by 10 is 100 $((-5)^2 + 5^2 + 5^2 + (-5)^2 + 0)$; and the smallest sum of squared effects

where two three-way interaction effects differ by 10 is 200 $((-5)^2 + 5^2 + 5^2 + (-5)^2 + (-5)^2 + 5^2 + 5^2 + (-5)^2 + 0)$.

We can use the function `mixed.power` to compute power in a factorial and thence choose our sample size.

```
1 > mixed.power(~mp*mz*dz,c(4,2,2,2),list(Error=50,mp=200/3,mz=200/1,dz=200/1,
      "mp:mz"=100/3,"mp:dz"=100/3,"mz:dz"=100/1,"mp:mz:dz"=200/3),alpha=.01)
               num.ev den.ev num.df den.df power
  Intercept    50.0000    50      1     16  0.01
  mp          583.3333    50      3     16  0.94
  mz         3250.0000    50      1     16  1.00
  dz         3250.0000    50      1     16  1.00
  mp:mz       183.3333    50      3     16  0.26
  mp:dz       183.3333    50      3     16  0.26
  mz:dz       850.0000    50      1     16  0.84
  mp:mz:dz    183.3333    50      3     16  0.26
```

The arguments to `mixed.power` are a one-sided model (we have no data to put on the left hand side) giving variable names and structure, in this case a full factorial. The second argument is the number of levels for for each factor (4, 2, 2) and the amount of replication. We start with $n = 2$. The third argument is a list, where the name of the component indicates the term in the model, and the value of the component is the sum of squared effects for the term divided by term degrees of freedom or the variance in the case of Error. Any of these except Error could be zero, but we have specified non-null values for each of our terms. The last argument is simply $\mathcal{E}_I$.

The power for main effects already meets our requirement at $n = 2$, as does the `mz:dz` interaction, but other three interactions fall short of their goal.

```
2 mixed.power(~mp*mz*dz,c(4,2,2,5),list(Error=50,mp=200/3,mz=200/1,dz=200/1,
      "mp:mz"=100/3,"mp:dz"=100/3,"mz:dz"=100/1,"mp:mz:dz"=200/3),alpha=.01)
               num.ev den.ev num.df den.df power
  ...
  mp:mz       383.3333    50      3     64  0.88
  mp:dz       383.3333    50      3     64  0.88
  mz:dz      2050.0000    50      1     64  1.00
  mp:mz:dz    383.3333    50      3     64  0.88
```

In fact, we need to raise n to 5 to meet the power goals.

It is clear that the interaction power goals are driving the sample size, at least when we are using this conservative value for non-centrality. Two things contribute to this. First, our conservative lower bound has the fewest possible interaction effects non-zero. If all of the interaction effects had absolute value 5, the power would be greater. Second, the multiplier for the of squared effects in the non-centrality parameter is smaller for interaction effects.

```
3 > mixed.power(˜mp*mz*dz,c(4,2,2,5),list(Error=100,mp=200/3,mz=200/1,dz=200/1,
    "mp:mz"=100/3,"mp:dz"=100/3,"mz:dz"=100/1,"mp:mz:dz"=200/3),alpha=.01)
             num.ev den.ev num.df den.df power
Intercept  100.0000    100      1     64  0.01
mp        1433.3333    100      3     64  1.00
mz        8100.0000    100      1     64  1.00
dz        8100.0000    100      1     64  1.00
mp:mz      433.3333    100      3     64  0.49
mp:dz      433.3333    100      3     64  0.49
mz:dz     2100.0000    100      1     64  0.96
mp:mz:dz   433.3333    100      3     64  0.49
4 > power.f.test(ncp=5*200/100,df1=3,df2=64,alpha=.01)
 [1] 0.4901104
```

If the error variance were 100 instead of 50, then the power values all decrease, although only three interactions fail to meet our goals. The last line gives an alternative method of computing power for the three-way interaction with $n = 5$, relating `mixed.power` to our previous `power.f.test`.

9.2   Unbalanced Data

Our discussion of factorials to this point has assumed *balance;* that is, that all factor-level combinations have the same amount of replication. When this is not true, the data are said to be *unbalanced*. The analysis of unbalanced data is more complicated, in part because there are no simple formulae for the quantities of interest, but also because it is not as clear what the appropriate quantities should be.

Balanced versus unbalanced data

The root cause of these complications has to do with orthogonality, or rather the lack of it. When the data are balanced, a contrast for one main effect or interaction is orthogonal to a contrast for any other main effect or interaction. One consequence of this orthogonality is that we can estimate effects and compute sums of squares one term at a time, and the results for that term do not depend on what other terms are in the model. When the data are unbalanced, the results we get for one term depend on what other terms are in the model, so we must to some extent do all the computations simultaneously.

Imbalance destroys orthogonality

The questions we want to answer do not change because the data are unbalanced. We still want to determine which terms are required to model the response adequately, and we may wish to test specific null hypotheses about model parameters. We made this distinction for balanced data in Section 8.11, even though the test statistics for comparing models or testing hypotheses are the same. For unbalanced data, this distinction actually leads to different tests.

Build models and/or test hypotheses

Our discussion will be divided into two parts: building models and testing hypotheses about parameters. We will consider only exact approaches for computing sums of squares and doing tests. There are approximate methods for unbalanced factorials that were popular before the easy availability

Use exact methods
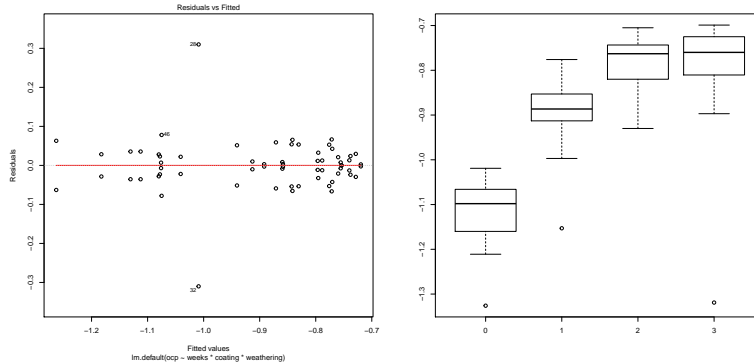
Figure 9.1: Residuals versus predicted plot and box plot by weeks for corrosion data.

of computers for doing all the hard computations. But when you have the computational horsepower, you might as well use it to get exact results.

9.2.1 Sums of squares in unbalanced data

We have formulated the sum of squares for a term in a balanced ANOVA model as the difference in error sum of squares for a reduced model that excludes the term of interest, and that same model with the term of interest included. The term of interest is said to have been "adjusted for" the terms in the reduced model. We also presented simple formulae for these sums of squares. When the data are unbalanced, we still compute the sum of squares for a term as a difference in error sums of squares for two models, but there are no simple formulae to accomplish that task. Furthermore, precisely which two models are used doesn't matter in balanced data so long as they only differ by the term of interest, but which models are used *does* matter for unbalanced data.

> SS adjusted for terms in reduced model

> Terms in model affect SS

Models are usually specified as a sequence of terms. For example, in a three-factor design we might specify (1, A, B, C) for main effects, or (1, A, B, AB, C) for main effects and the AB interaction. The "1" denotes the overall grand mean μ that is included in all models. The sum of squares for a term is the difference in error sums of squares for two models that differ only by that term. For example, if we look at the two models (1, A, C) and (1, A, B, C), then the difference in error sums of squares will be the sum of squares for B adjusted for 1, A, and C. We write this as $SS(B|1, A, C)$.

> $SS(B|1, A)$ is SS of B adjusted for 1 and A

Example 9.2 Anti-corrosion coatings

Aluminum corrosion is a major issue in the aerospace industry, and a great deal of work has gone into finding coatings that will prevent, or at least delay, corrosion. Electrochemistry states that only the more reactive of two

Table 9.1: Open circuit potential (volts) for 64 coated and weathered aluminum samples. Data from T. Chen; data set `Corrosion`.

Coating	Weathering	0	1	2	3
			Weeks		
30%	B117	-1.019	-0.776	-0.763	-0.762
		-1.063	-0.801	-0.748	-0.714
	D5894	-1.052	-0.777	-0.763	-0.728
		-1.108	-0.884	-0.828	-0.813
40%	B117	-1.069	-0.868	-0.781	-0.829
		-1.083	-0.851	-0.739	-0.723
	D5894	-1.154	-0.895	-0.754	-0.699
		-1.211	-0.889	-0.754	-1.319
45%	B117	-1.095	-0.889	-0.776	-0.789
		-1.166	-0.992	-0.907	-0.897
	D5894	-1.326	-1.153	-0.812	-0.785
		-1.200	-0.997	-0.930	-0.808
Commercial	B117	-1.148	-0.923	-0.705	-0.727
		-1.077	-0.903	-0.838	-0.753
	D5894	-1.101	-0.855	-0.722	-0.699
		-1.055	-0.862	-0.717	-0.758

metals will react when two joined metals are in a corrosive environment. For this reason, coatings that are magnesium rich continue to be developed for aluminum. These are called "sacrificial coatings," because they are literally consumed while protecting the aluminum underneath.

This experiment compares a commercially available coating to three experimental coatings, when weathered according to two different protocols for corrosive conditions (B117 or D5894), maintained for four different lengths of time (0, 1, 2, or 3 weeks). The three experimental coatings differ in the concentration of magnesium pigment: 30%, 40%, or 45%. (Too high a Mg concentration can degrade other desirable properties such as adhesion or durability.)

Sixty-four aluminum samples are randomly assigned to the thirty-two combinations of coating, weathering, and time. After treatment, each sample is tested for its "open circuit potential" (OCP). Lower (more negative) values are better. Data are shown in Table 9.1.

We begin by fitting the three-factor model to the complete, balanced dataset.

```
1 > fit <- lm(ocp~weeks*coating*weathering,data=Corrosion)
2 > plot(fit,which=1)
3 > with(Corrosion,boxplot(ocp~weeks))
```

The residual plot (panel one of Figure 9.1) shows two outliers. Recall that there are two units per treatment in these data, so one true outlier can look like

two. The boxplot produced in line 3 (panel two of Figure 9.1) shows that the outlier is an incredibly low (good) value of OCP observed after three weeks of weathering. This value is better than all but one of the other responses, including those for units that were not weathered at all. It thus seems to be physically implausible. We will analyze without this data point and request that the experimenter determine whether there was a transcription error (or some other issue) on this response.

Removing one outlier makes this dataset unbalanced.

```
 4 > anova(tmpfit <- lm(ocp~coating,data=Corrosion,subset=(1:64)!=32))
           Df  Sum Sq  Mean Sq F value Pr(>F)
   coating   3 0.13783 0.045944  1.8688 0.1447
   Residuals 59 1.45052 0.024585
 5 > model.effects(tmpfit,"coating")
         30          40          45 commercial
   0.0430250   0.0063625  -0.0771625   0.0277750
 6 > anova(tmpfit <- lm(ocp~weeks+coating,data=Corrosion,subset=(1:64)!=32))
           Df  Sum Sq Mean Sq F value    Pr(>F)
   weeks     3 1.26278 0.42093 127.202 < 2.2e-16 ***
   coating   3 0.14026 0.04675  14.129 5.696e-07 ***
   Residuals 56 0.18531 0.00331
 7 > model.effects(lm(tmpfit,"coating")
         30          40          45  commercial
   0.04092325  0.01266776 -0.07926425  0.02567325
 8 > anova(tmpfit <- lm(ocp~weathering+coating,data=Corrosion,subset=(1:64)!=32))
              Df  Sum Sq  Mean Sq F value Pr(>F)
   weathering  1 0.01037 0.010367  0.4175 0.5207
   coating     3 0.13768 0.045892  1.8480 0.1485
   Residuals  58 1.44031 0.024833
 9 > model.effects(tmpfit,"coating")
          30          40           45   commercial
   0.043237288  0.005725636 -0.076950212  0.027987288
10 > anova(tmpfit <- lm(ocp~weeks+weathering+coating,data=Corrosion,
     subset=(1:64)!=32))
              Df  Sum Sq Mean Sq  F value    Pr(>F)
   weeks       3 1.26278 0.42093 129.8914 < 2.2e-16 ***
   weathering  1 0.00736 0.00736   2.2713    0.1375
   coating     3 0.13998 0.04666  14.3986 4.858e-07 ***
   Residuals  55 0.17823 0.00324
11 > model.effects(tmpfit,"coating")
         30          40          45  commercial
   0.04110938  0.01210937 -0.07907813  0.02585937
```

Lines 4–11 show ANOVA tables and effects for coating for various other terms preceding coating in the model. Note that all of these sums of squares for coating differ, and all of the fitted effects for coating differ. This is a result of the lack of balance.

```
12 > anova(tmpfit <- lm(ocp~weeks*weathering+coating,data=Corrosion,
     subset=(1:64)!=32))
                    Df  Sum Sq Mean Sq  F value     Pr(>F)
  weeks              3 1.26278 0.42093 134.3364 < 2.2e-16 ***
  weathering         1 0.00736 0.00736   2.3490    0.1314
  coating            3 0.13998 0.04666  14.8913 4.033e-07 ***
  weeks:weathering   3 0.01530 0.00510   1.6274    0.1943
  Residuals         52 0.16294 0.00313
13 > model.effects(tmpfit,"coating")
          30          40          45  commercial
  0.04074410  0.01320519 -0.07944340  0.02549410
14 > anova(tmpfit <- lm(terms(ocp~weeks*weathering+coating,keep.order=TRUE),
     data=Corrosion,subset=(1:64)!=32))
                    Df  Sum Sq Mean Sq  F value     Pr(>F)
  weeks              3 1.26278 0.42093 134.3364 < 2.2e-16 ***
  weathering         1 0.00736 0.00736   2.3490    0.1314
  weeks:weathering   3 0.01475 0.00492   1.5687    0.2081
  coating            3 0.14053 0.04684  14.9500  3.85e-07 ***
  Residuals         52 0.16294 0.00313
15 > model.effects(tmpfit,"coating")
          30          40          45  commercial
  0.04074410  0.01320519 -0.07944340  0.02549410
```

Line 12 shows that **R** reorders terms in the model so that main effects enter first (then two-factor interactions, and so on). That means that you need to tell **R** that you insist on your chosen order for the terms if you want to have an interaction in the model before a main effect. Line 14 shows how to do that using the `terms` function with `keep.order=TRUE`. Again, the sum of squares for coating differs from the previous quantities calculated. Note, however, that the coefficients in lines 13 and 15 are the same; coefficients depend on what terms are in the model, but not on the order in which the terms were entered.

The simplest choice for a sum of squares is *sequential* sums of squares. This is called Type I in SAS, and that terminology is also widely used. For sequential sums of squares, we specify a model and the sum of squares for any term is adjusted for those terms that precede it in the model. If the model is (1, A, B, AB, C), then the sequential sums of squares are $SS(A|1)$, $SS(B|1, A)$, $SS(AB|1, A, B)$, and $SS(C|1, A, B, AB)$. Notice that if you specify the terms in a different order, you get different sums of squares; the sequential sums of squares for (1, A, B, C, AB) are $SS(A|1)$, $SS(B|1, A)$, $SS(C|1, A, B)$, and $SS(AB|1, A, B, C)$. The `anova` function in **R** produces sequential sums of squares.

Type I SS is sequential

Type I SS depends on order of terms

Two models that include the same terms in different orders will have the same estimated treatment effects and interactions. However, models that include different terms may have different estimated effects for the terms they have in common. Thus (1, A, B, AB, C) and (1, A, B, C, AB) will have the same $\widehat{\alpha}_i$'s, but (1, A, B, AB, C) and (1, A, B, C) may have different $\widehat{\alpha}_i$'s. We saw this in Example 9.2.

Estimated effects don't depend on order of terms

9.2.2 Building models

Building models means deciding which main effects and interactions are needed to describe the data adequately. I build hierarchical models. In a hierarchical model, the inclusion of any interaction in a model implies the inclusion of any term that is "above" it, where we say that a factorial term U is above a factorial term V if every factor in term U is also in term V. The goal is to find the hierarchical model that includes all terms that must be included, but does not include any unnecessary terms.

Compare hierarchical models

Our approach to computing sums of squares when model-building is to use as the reduced model for term U the largest hierarchical model M that does not contain U. This is called Type II in SAS, and the SAS terminology is widely used. In two-factor models, this might be called "Yates' fitting constants" or "each adjusted for the other."

Type II SS or Yates' fitting constants

Consider computing Type II sums of squares for all the terms in a three-factor model. The largest hierarchical models not including ABC, BC, and C are (1, A, B, C, AB, AC, BC), (1, A, B, C, AC, AB), and (1, A, B, AB), respectively. Thus for Type II sums of squares, the three-factor interaction is adjusted for all main effects and two-factor interactions, a two-factor interaction is adjusted for all main effects and the other two-factor interactions, and a main effect is adjusted for the other main effects and their interactions, or $\text{SS}(ABC|1, A, B, C, AB, AC, BC)$, $\text{SS}(BC|1, A, B, C, AB, AC)$, and $\text{SS}(C|1, A, B, AB)$. In Example 9.2, the Type II sum of squares for coating is .14053.

Type II adjusts for largest hierarchal model not including term

It is important to point out that the denominator mean square used for testing is MS_E from the full model. We do not pool "unused" terms into error. Thus, the Type II SS for C is $\text{SS}(C|1, A, B, AB)$, but the error mean square for testing is from the model (1, A, B, C, AB, AC, BC, ABC).

Use MS_E from full model

Example 9.3 Anti-corrosion coatings, continued

While it is mildly instructive to generate Type II sums of squares by forcing the proper ordering of terms in a model, it is also unnecessarily tedious.

```
16 > fit2 <- lm(ocp~weeks*coating*weathering,data=Corrosion,
     subset=(1:64)!=32)
17 > car::Anova(fit2,type=2)
  Anova Table (Type II tests)

  Response: ocp
                        Sum Sq Df  F value    Pr(>F)
  weeks                 1.26103  3 147.6457 < 2.2e-16 ***
  coating               0.14053  3  16.4539 1.406e-06 ***
  weathering            0.00706  1   2.4812    0.1254
  weeks:coating         0.02912  9   1.1363    0.3681
  weeks:weathering      0.01538  3   1.8009    0.1676
  coating:weathering    0.02210  3   2.5871    0.0708 .
  weeks:coating:weathering 0.02347  9   0.9162    0.5242
  Residuals             0.08826 31
```

Line 16 fits the full model with the outlier removed, and we can get the full

Table 9.2: A highly unbalanced two by two factorial.

A		Low				High				
Low	2.7	7.9	26.3	-1.9	30.6		21.5			
	3.8	27.2	20.9	20.6	14.6					
High		26.1				41.1	46.7	57.8	38	39.3

Type II ANOVA table by using the `Anova` function from the `car` package with the argument `type=2`, as in line 17. Note that the sum of squares for coating (.14053) is the same as what we found on line 14, with coating adjusted for weeks, weathering, and weeks:weathering.

We begin by asking whether we need the three factor interaction. It is not significant, so we can now look at the two-factor interactions. None of them is significant, so we now look at the main effects. Weeks and coating are both significant, but weather is not. We would thus include main effects of weeks and coating in our final model, but not other terms.

Note that the coating:weathering interaction is approaching significance. If it had been significant, we would have decided to include coating, weathering, and coating:weathering without ever testing coating and weathering individually. That is, we would maintain hierarchy and not even consider eliminating an included term for a significant interaction.

Type I sums of squares for the terms in a model will sum to the overall model sum of squares with $g - 1$ degrees of freedom. This is not true for Type II sums of squares, as can be seen in Line 17; the model sum of squares is 1.5001, but the Type II sums of squares add to 1.4987.

The Type II approach to model building is not foolproof. The following example shows that in some situations the overall model can be highly significant, even though none of the individual terms in the model is significant.

Example 9.4 Unbalanced data puzzle

Consider the data in Table 9.2. These data are *highly* unbalanced.

```
1 > fit <- lm(y~A*B,data=HighlyUnbalanced)
2 > car::Anova(fit,type=2)
nova Table (Type II tests)
          Sum Sq Df F value  Pr(>F)
          485.29  1  4.3187 0.05807 .
          254.63  1  2.2660 0.15614
:B         65.24  1  0.5806 0.45967
esiduals 1460.79 13
3 > anova(lm(y~A:B,data=HighlyUnbalanced))
         Df Sum Sq Mean Sq F value   Pr(>F)
:B        3 2876.9  958.96  8.5341 0.002164 **
esiduals 13 1460.8  112.37
```

Line 1 fits the two-way model for these data, and line 2 shows the Type II ANOVA. Nothing appears significant here, although the main effect of A has

a p-value under .06. However, look at the ANOVA for the model in line 3. This model lumps all three between-treatments degrees of freedom into a single term (analogous to a four-level factor), and this three degree of freedom term is very significant. We thus have a model that is highly significant, but none of the terms seems to be significant (at least not according to Type II). That is a little disturbing.

What has actually happened in these data is that either A or B alone explains a large amount of variation, but they are in some sense explaining the same variation. This can be seen in lines 4 and 5.

```
4 > anova(lm(y~A*B,data=HighlyUnbalanced))
            Df  Sum Sq Mean Sq F value     Pr(>F)
  A          1 2557.00 2557.00 22.7555 0.0003658 ***
  B          1  254.63  254.63  2.2660 0.1561423
  A:B        1   65.24   65.24  0.5806 0.4596671
  Residuals 13 1460.79  112.37
5 > anova(lm(y~B*A,data=HighlyUnbalanced))
            Df  Sum Sq Mean Sq F value     Pr(>F)
  B          1 2326.35 2326.35 20.7029 0.0005451 ***
  A          1  485.29  485.29  4.3187 0.0580671 .
  B:A        1   65.24   65.24  0.5806 0.4596671
  Residuals 13 1460.79  112.37
```

Thus B is not needed if A is already present, A is not needed if B is already present, and the interaction is never needed. But we need *something*!

In summary, Type II is usually a good way of building a model, but you should also check on the total predictive capacity of the model to ensure that a good model is not being hidden due to high correlation between factor effects.

Test full model too

9.2.3 Testing hypotheses

In some situations we may wish to test specific hypotheses about treatment means rather than building a model to describe the means. Many of these hypotheses can be expressed in terms of the factorial parameters, but recall that the parameters we use in our factorial decomposition carry a certain amount of arbitrariness in that they assume equally weighted averages. When the hypotheses of interest correspond to our usual, equally weighted factorial parameters, testing is reasonably straightforward; otherwise, special purpose contrasts must be used.

Standard tests are for equally weighted factorial parameters

Let's review how means and parameters correspond in the two-factor situation. Let μ_{ij} be the mean of the ijth treatment:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$$

with

$$0 = \sum_i \alpha_i = \sum_j \beta_j = \sum_i \alpha\beta_{ij} = \sum_j \alpha\beta_{ij} \ .$$

Let n_{ij} be the number of observations in the ijth treatment. Form row and column averages of treatment means using equal weights for the treatment

means:

$$\mu_{i\bullet} = \sum_{j=1}^{b} \mu_{ij}/b$$
$$= \mu + \alpha_i \ ,$$
$$\mu_{\bullet j} = \sum_{i=1}^{a} \mu_{ij}/a$$
$$= \mu + \beta_j \ .$$

The null hypothesis that the main effects of factor A are all zero ($\alpha_i \equiv 0$) is the same as the null hypothesis that all the row averages of the treatment means are equal ($\mu_{1\bullet} = \mu_{2\bullet} = \cdots = \mu_{a\bullet}$). This is also the same as the null hypothesis that all factor A main-effects contrasts evaluate to zero.

Testing the null hypothesis that the main effects of factor A are all zero ($\alpha_i \equiv 0$) is accomplished with an F-test. We compute the sum of squares for this hypothesis by taking the difference in error sum of squares for two models: the full model with all factors and interactions, and that model with the main effect of factor A deleted, or $SS(A|1, B, C, AB, AC, BC, ABC)$ in a three-factor model. This reduced model is not hierarchical; it includes interactions with A but not the main effect of A. Similarly, we compute a sum of squares for any other hypothesis that a set of factorial effects is all zero by comparing the sum of squares for the full model with the sum of squares for the model with that effect removed. This may be called "standard parametric," "Yates' weighted squares of means," or "fully adjusted"; in SAS it is called Type III.

Example 9.5 Unbalanced data puzzle, continued

Let us continue Example 9.4. If we wish to test the null hypothesis that $\alpha_i \equiv 0$ or $\beta_j \equiv 0$, we need to use Type III tests, as shown in line 6.

```
6 > car::Anova(fit,type=3)
 Anova Table (Type III tests)
             Sum Sq Df F value     Pr(>F)
(Intercept) 5019.8  1 44.6726 1.508e-05 ***
A            500.0  1  4.4492   0.05488 .
B            265.5  1  2.3625   0.14826
A:B           65.2  1  0.5806   0.45967
Residuals   1460.8 13
```

None of the null hypotheses about main effects or interaction is anywhere near as significant as the overall model; all have p-values greater than .05.

How can this be so when we know that there are large differences between treatment means in the data? Consider for a moment the test for factor A main effects. The null hypothesis is that the factor A main effects are zero, but no constraint is placed on factor B main effects or the interactions. We can fit the data fairly well with the α_i's equal to zero, so long as we can manipulate the β_j's and $\alpha\beta_{ij}$'s to take up the slack. Similarly, when testing

factor B, no constraint is placed on factor A main effects or AB interactions. These three tests of A, B, and AB do not test that all three null hypotheses are true simultaneously. For that we need to test the overall model with 3 degrees of freedom, and that test is highly significant.

When we test the null hypothesis that a contrast in treatment effects is zero, we are testing the null hypothesis that a particular linear combination of treatment means is zero with no other restrictions on the cell means. This is equivalent to testing that the single degree of freedom represented by the contrast can be removed from the full model, so the contrast has been adjusted for all other effects in the model. Thus the sum of squares for any contrast is a Type III sum of squares.

Contrast SS are
Type III

Example 9.6 Anti-corrosion coatings, continued

Continuing Example 9.2, the Type III ANOVA can be found at line 18.

```
18 > car::Anova(fit2,type=3)
   Anova Table (Type III tests)
                          Sum Sq Df   F value     Pr(>F)
   (Intercept)           49.162  1 17268.0770 < 2.2e-16 ***
   weeks                  1.265  3   148.1638 < 2.2e-16 ***
   coating                0.141  3    16.5663 1.319e-06 ***
   weathering             0.006  1     2.0064   0.16661
   weeks:coating          0.029  9     1.1499   0.35970
   weeks:weathering       0.017  3     1.9485   0.14233
   coating:weathering     0.022  3     2.6005   0.06978 .
   weeks:coating:weathering 0.023 9     0.9162   0.52424
   Residuals              0.088 31
19 > linear.contrast(fit2,weathering,c(-1,1))
      estimates        se   t-value  p-value    lower-ci    upper-ci
   1 -0.0191875 0.01354607 -1.416463 0.166611 -0.04681489 0.008439889
```

The Type III sum of squares for weathering (which has 1 degree of freedom) is .006, different from both Types I and II. The F is 2.0064 with a p-value of .16661. Line 19 computes a linear contrast for weathering (it has a single degree of freedom, so there is really only one possible contrast). Notice that it has exactly the same p-value as we see in the Type III ANOVA, and the ANOVA F for weathering is the square of the contrast t-statistic.

In this example, Type II and Type III tests are giving the same results: only main effects of weeks and coating are needed.

9.2.4 Empty cells

The problems of unbalanced data are immensely increased when one or more of the cells are empty, that is, when there are no data for some factor-level combinations. The model-building/Type II approach to analysis doesn't really change. We can just keep comparing hierarchical models. However, anything depending on the parameters, including estimation and the hypothesis testing/Type III approach, becomes very problematic, because the parameters

Empty cells make
factorial effects
ambiguous

Table 9.3: A three-by-two table of means with one empty cell, and two different decompositions of the means into "grand mean", row, column, and interaction effects.

		156.0	-23.0	23.0	133.0	.0	.0
196	124	4.0	59.0	-59.0	27.0	36.0	-36.0
156	309	76.5	-53.5	53.5	99.5	-76.5	76.5
47		-80.5	-5.5	5.5	-126.5	40.5	-40.5

about which we are making inference are no longer uniquely defined, even when we are sure we want to work with equal weighting.

When there are empty cells, there are infinitely many different sets of factorial effects that fit the observed treatment means exactly; these different sets of effects disagree on what they fit for the empty cells. Consider the three-by-two table of means with one empty value, and two different factorial decompositions of the means into grand mean, row, column, and interaction effects shown in Figure 9.3. Both of these factorial decompositions meet the usual zero-sum requirements, and both add together to match the table of means exactly. The first is what would be obtained if the empty cell had mean 104, and the second if the empty cell had mean -34.

Because the factorial effects are ambiguous, it makes no sense to test hypotheses about the factorial model parameters. For example, are the column effects above zero or nonzero? What does make sense is to look at simple effects and to set up contrasts that make factorial-like comparisons where possible. For example, levels 1 and 2 of factor A are complete, so we can compare those two levels with a contrast. Note that the difference of row means is 72.5, and $\alpha_2 - \alpha_1$ is 72.5 in both decompositions. We might also want to compare level 1 of factor B with level 2 of factor B for the two levels of factor A that are complete. There are many potential ways to choose interesting contrasts for designs with empty cells.

9.3 Contrasts and Multiple Comparisons for Factorial Data

Contrasts allow us to examine particular ways in which treatments differ. With factorial data, we can use contrasts to look at how specific main effects differ and to see patterns in interactions. Indeed, we have seen that the usual factorial ANOVA can be built from sets of contrasts. Chapters 4 and 5 discussed contrasts and multiple comparisons in the context of single factor analysis. These procedures carry over to factorial treatment structures with little or no modification.

In this section we will discuss contrasts in the context of a three-way factorial; generalization to other numbers of factors is straightforward. The factors in our example experiment are drug (one standard drug and two new drugs), dose (four levels, equally spaced), and administration time (morning

Expected value	$\displaystyle\sum_{ijk} w_{ijk}\,\mu_{ijk}$
Variance	$\displaystyle\sigma^2 \sum_{ijk} \frac{w_{ijk}^2}{n_{ijk}}$
Sum of squares	$\displaystyle\frac{\left(\sum_{ijk} w_{ijk}\,\overline{y}_{ijk\bullet}\right)^2}{\sum_{ijk} w_{ijk}^2/n_{ijk}}$
Confidence interval	$\displaystyle\sum_{ijk} w_{ijk}\,\overline{y}_{ijk\bullet} \pm t_{\mathcal{E}/2,\,N-abc}$ $\times \sqrt{\mathrm{MS_E}\ \textstyle\sum_{ijk} w_{ijk}^2/n_{ijk}}$
F-test	$\displaystyle\frac{\left(\sum_{ijk} w_{ijk}\,\overline{y}_{ijk\bullet}\right)^2}{\mathrm{MS_E}\sum_{ijk} w_{ijk}^2/n_{ijk}}$

Display 9.1: Contrast formulae for a three-way factorial.

or evening). We will usually assume balanced data, because contrasts for balanced factorial data have simpler orthogonality relationships.

We saw in one-way analysis that the arithmetic of contrasts is not too hard; the big issue was finding contrast coefficients that address an interesting question. The same is true for factorials. Suppose that we have a set of contrast coefficients w_{ijk}. We can work with this contrast for a factorial just as we did with contrasts in the one-way case using the formulae in Display 9.1. These formulae are nothing new, merely the application of our usual contrast formulae to the design with $g = abc$ treatments. We still need to find meaningful contrast coefficients.

Inference for contrasts remains the same

Pairwise comparisons are differences between two treatments, ignoring the factorial structure. We might compare the standard drug at the lowest dose with morning administration to the first new drug at the lowest dose with evening administration. As we have seen previously with pairwise comparisons, there may be a multiple testing issue to consider, and our pairwise multiple comparisons procedures (for example, HSD) carry over directly to the factorial setting.

Pairwise comparisons

A *simple effect* is a particular kind of pairwise comparison. A simple effect is a difference between two treatments that have the same levels of all factors but one. A comparison between the standard drug at the lowest dose with morning administration and the standard drug at the lowest dose with evening administration is a simple effect. Differences of main effects are averages of simple effects.

Simple effects are pairwise differences that vary just one factor

The structure of a factorial design suggests that we should also consider

contrasts that reflect the design, namely main-effect contrasts and interaction contrasts. In general, we use contrasts with coefficient patterns that mimic those of factorial effects. A *main-effect contrast* is one where the coefficients w_{ijk} depend only on a single index; for example, k for a factor C contrast. That is, two contrast coefficients are equal if they have the same k index. These coefficients will add to zero across k for any i and j. For *interaction* contrasts, the coefficients depend only on the indices of factors in the interaction in question and satisfy the same zero-sum restrictions as their corresponding model terms. Thus a BC interaction contrast has coefficients w_{ijk} that depend only on j and k and add to zero across j or k when the other subscript is kept constant. For an ABC contrast, the coefficients w_{ijk} must add to zero across any subscript.

Main-effect and interaction contrasts examine factorial components

We can use pairwise multiple comparisons procedures such as HSD for marginal means. Thus to compare all levels of factor B using HSD, we treat the means $\overline{y}_{\bullet j \bullet \bullet}$ as b treatment means each with sample size acn and do multiple comparisons with $abc(n-1)$ degrees of freedom for error. The same approach works for two-way and higher marginal tables of means. For example, treat $\overline{y}_{\bullet jk\bullet}$ as bc treatment means each with sample size an and $abc(n-1)$ degrees of freedom for error. Pairwise multiple comparisons procedures also work when applied to main effects—for example, $\widehat{\beta}_j$—but most do not work for interaction effects due to the additional zero sum restrictions. (Bonferroni does work.)

Pairwise multiple comparisons work for marginal means

Please note: simple-effects, main-effects, and interaction contrasts are examples of contrasts that are frequently useful in analysis of factorial data; there are many other kinds of contrasts.

> Use contrasts that address your questions. Don't be put off if a contrast that makes sense to you does not fit into one of these neat categories.

Example 9.7 Factorial contrasts

Let's look at some factorial contrasts for our three-way drug test example. Coefficients w_{ijk} for these contrasts are shown in Table 9.4. Suppose that we want to compare morning and evening administration times averaged across all drugs and doses. The first contrast in Table 9.4 has coefficients -1 for evening and 1 for morning and thus makes the desired comparison. This is a main-effect contrast (coefficients only depend on administration time, factor C). We can get the same information by using a contrast with coefficients (1, -1) and the means $\overline{y}_{\bullet\bullet k\bullet}$ or effects $\widehat{\gamma}_k$.

The response presumably changes with drug dose (factor B), so it makes sense to examine dose as a quantitative effect. To determine the linear effect of dose, use a main-effect contrast with coefficients -3, -1, 1, and 3 for doses 1 through 4 (Appendix Table C.6); this is the second contrast in Table 9.4. As with the first example, we could again get the same information from a contrast in the means $\overline{y}_{\bullet j\bullet\bullet}$ or effects $\widehat{\beta}_j$ using the same coefficients. The

Table 9.4: Example contrasts.

Morning versus Evening

	Morning/Dose					Evening/Dose			
Drug	1	2	3	4	Drug	1	2	3	4
1	1	1	1	1	1	-1	-1	-1	-1
2	1	1	1	1	2	-1	-1	-1	-1
3	1	1	1	1	3	-1	-1	-1	-1

Linear in Dose

	Morning/Dose					Evening/Dose			
Drug	1	2	3	4	Drug	1	2	3	4
1	-3	-1	1	3	1	-3	-1	1	3
2	-3	-1	1	3	2	-3	-1	1	3
3	-3	-1	1	3	3	-3	-1	1	3

Linear in Dose by Morning versus Evening

	Morning/Dose					Evening/Dose			
Drug	1	2	3	4	Drug	1	2	3	4
1	-3	-1	1	3	1	3	1	-1	-3
2	-3	-1	1	3	2	3	1	-1	-3
3	-3	-1	1	3	3	3	1	-1	-3

Linear in Dose by Morning versus Evening
by Drug 2 versus Drug 3

	Morning/Dose					Evening/Dose			
Drug	1	2	3	4	Drug	1	2	3	4
1	0	0	0	0	1	0	0	0	0
2	-3	-1	1	3	2	3	1	-1	-3
3	3	1	-1	-3	3	-3	-1	1	3

Linear in Dose for Drug 1

	Morning/Dose					Evening/Dose			
Drug	1	2	3	4	Drug	1	2	3	4
1	-3	-1	1	3	1	-3	-1	1	3
2	0	0	0	0	2	0	0	0	0
3	0	0	0	0	2	0	0	0	0

simple coefficients -3, -1, 1, and 3 are applicable here because the doses are equally spaced and balance gives equal sample sizes.

A somewhat more complex question is whether the linear effect of dose is the same for the two administration times. To determine this, we compute the linear effect of dose from the morning data, and then subtract the linear effect of dose from the evening data. This is the third contrast in Table 9.4. This is a two-factor interaction contrast; the coefficients add to zero across dose or administration time. Note that this contrast is literally the elementwise product of the two corresponding main-effects contrasts.

A still more complex question is whether the dependence of the linear

effect of dose on administration times is the same for drugs 2 and 3. To determine this, we compute the linear in dose by administration time interaction contrast for drug 2, and then subtract the corresponding contrast for drug 3. This three-factor interaction contrast is the fourth contrast in Table 9.4. It is formed as the elementwise product of the linear in dose by administration time two-way contrast and a main-effect contrast between drugs 2 and 3.

Finally, the last contrast in Table 9.4 is an example of a useful contrast that is not a simple effect, main effect, or interaction contrast. This contrast examines the linear effect of dose for drug one, averaged across time.

The interaction contrasts in Example 9.7 illustrate an important special case of interaction contrasts, namely, products of main-effect contrasts. These products allow us to determine if an interesting contrast in one main effect varies systematically according to an interesting contrast in a second main effect.

We can reexpress a main-effect contrast in the individual treatment means $\overline{y}_{ijk\bullet}$ in terms of a contrast in the factor main effects or the factor marginal means. For example, a contrast in factor C can be reexpressed as

$$\sum_{ijk} w_{ijk}\, \overline{y}_{ijk\bullet} \;=\; \sum_{k}\left[w_{11k} \sum_{ij} \overline{y}_{ijk\bullet}\right]$$
$$=\; \sum_{k} w_k\, \overline{y}_{\bullet\bullet k\bullet}$$
$$=\; \sum_{k} w_k\, \widehat{\gamma}_k \;,$$

where $w_k = abw_{11k}$. Because scale is somewhat arbitrary for contrast coefficients, we could also use $w_k = w_{11k}$ and still get the same kind of information. For balanced data, two main-effect contrasts for the same factor with coefficients w_k and $w_k^\star$ are orthogonal if

$$\sum_{k} w_k\, w_k^\star = 0 \;\;.$$

We can also express an interaction contrast in the individual treatment means as a contrast in marginal means or interaction effects. For example, suppose w_{ijk} is a set of contrast coefficients for a BC interaction contrast. Then we can rewrite the contrast in terms of marginal means or interaction effects:

$$\sum_{ijk} w_{ijk}\, \overline{y}_{ijk\bullet} \;=\; \sum_{jk} w_{jk}\, \overline{y}_{\bullet jk\bullet}$$
$$=\; \sum_{jk} w_{jk}\, \widehat{\beta\gamma}_{jk}$$

where $aw_{1jk} = w_{jk}$. Two interaction contrasts for the same interaction with coefficients w_{jk} and $w_{jk}^\star$ are orthogonal if

$$\sum_{jk} w_{jk}\, w_{jk}^\star = 0 \ \ .$$

For balanced data, the formulae in Display 9.1 can be simplified by replacing the sample size n_{ijk} by the common sample size n. The formulae can be simplified even further for main-effect and interaction contrasts, because they can be rewritten in terms of the effects or marginal means of interest instead of using all treatment means. Consider a main-effect contrast in factor C with coefficients w_k; the number of observations at the kth level of factor C is abn. We have for the contrast $\sum_k w_k\, \overline{y}_{\bullet\bullet k\bullet}$:

Expected value	$\sum_k w_k\, \gamma_k$
Variance	$\sum_k w_k^2\, \sigma^2/(abn)$
Sum of squares	$\dfrac{(\sum_k w_k\, \overline{y}_{\bullet\bullet k\bullet})^2}{\sum_k w_k^2/(abn)}$
Confidence interval	$\sum_k w_k\, \overline{y}_{\bullet\bullet k\bullet} \pm$ $t_{\mathcal{E}/2,\,N-abc}\sqrt{\mathrm{MS_E}\, \sum_k w_k^2/(abn)}$
F-test	$\dfrac{(\sum_k w_k\, \overline{y}_{\bullet\bullet k\bullet})^2}{\mathrm{MS_E}\, \sum_k w_k^2/(abn)}$

The simplification is similar for interaction contrasts. For example, the BC interaction contrast $\sum_{jk} w_{jk}\, \overline{y}_{\bullet jk\bullet}$ has sum of squares

$$\frac{(\sum_{jk} w_{jk}\, \overline{y}_{\bullet jk\bullet})^2}{\sum_{jk} w_{jk}^2/(an)}$$

(an is the "sample size" at each jk combination).

The perceptive reader may have noticed that we can do a lot of F-tests in the analysis of a factorial, but we haven't been talking about multiple comparisons adjustments for the F-tests. Why this resounding silence, when we were so careful to describe and account for multiple testing for pairwise comparisons? I have no good answer; common statistical practice seems inconsistent in this regard. What common practice does is treat each main effect and interaction as a separate "family" of hypotheses and make multiple comparisons adjustments within a family but not between families.

We sometimes use an informal multiple comparisons correction when building hierarchical models. Suppose that we have a three-way factorial, and only the three-way interaction is significant, with a p-value of .04; the main-effects and two-factor interactions are not near significance. I would

probably conclude that the low p-value for the three-way interaction is due to chance rather than interaction effects. I conclude this because I usually expect main effects to be bigger than two-factor interactions, and two-factor interactions to be bigger than three-factor interactions. I thus interpret an isolated, marginally significant three-way interaction as a null result. I know that isolated three-way interaction can occur, but it seems less likely to me than chance occurrence of a moderately low p-value.

Somewhat more quantitatively, a significant interaction in a hierarchical model forces the inclusion of lower-order terms, so instead of merely testing an individual term U, we could consider testing the composite of U and all of the lower-order terms that are included in U that would not have been included in the model but for the significance of U. For example, consider a four factor model with apparently significant terms A, B, C, D, AB, and ABCD. If we do not include ABCD in the model, we only have five terms. If we include ABCD, we must also include AC, AD, BC, BD, CD, ABC, ABD, ACD, BCD as well as ABCD. Thus before I throw all 10 additional terms in the model simply because ABCD was significant, I can also test the composite of those 10 terms. This adds a "step down" aspect to our approach and helps guard against the random small p-value in higher-order terms that imply the inclusion of multiple additional terms. This helps protect the experimentwise error rate.

9.4 Modeling Interaction

Analysis of factorially structured data should be more than just an enumeration of which main effects and interactions are significant. We should look closely at the data to try to determine what the data are telling us by understanding the main effects and interactions in the data. For example, reporting that factor B only affects the response at the high level of factor A is more informative than reporting that factors A and B have significant main effects and interactions. One of my pet peeves is an analysis that just reports significant terms.

An interaction is a deviation from additivity. If the effect of going from dose 1 to dose 2 changes from drug 2 to drug 3, then there is an interaction between drug and dose. Similarly, if the interaction of drug and dose is different in morning and evening applications, then there is a three-factor interaction between drug, dose, and time. Try to understand and model any interaction that may be present in your data. This is often not easy, but when it can be done it leads to much greater insight into what the data have to say. As Tolstoy should have said, "Additive data sets are all alike; every non-additive data set is non-additive in its own way." This section discusses several specific models for interaction; there are many others.

9.4.1   One-cell interaction

A *one-cell interaction* is a common type of interaction where most of the experiment is additive, but one treatment deviates from the additive structure. The name "cell" comes from the idea that one cell in the table of treatment means does not follow the additive model. More generally, there may be one or a few cells that deviate from a relatively simple model. If the deviation from the simple model in these few cells is great enough, all the usual factorial interaction effects can be large and statistically significant.

Understanding one-cell interaction is easy: the data follow a simple model except for a single cell or a few cells. Finding a one-cell interaction is harder. It requires a careful study of the interaction effects or plots or a more sophisticated estimation technique than the least squares we have been using (see Daniel 1976 or Oehlert 1994). Be warned, large one-cell interactions can be masked or hidden by other large one-cell interactions.

One-cell interactions can sometimes be detected by examination of interaction effects. A table of interaction effects adds to zero across rows or columns. A one-cell interaction shows up in the effects as an entry with a large absolute value. The other entries in the same row and column are moderate and of the opposite sign, and the remaining entries are small and of the same sign as the interacting cell. For example, a three by four factorial with all responses 0 except for 12 in the (2,2) cell has interaction effects as follows:

$$\begin{array}{rrrr} 1 & -3 & 1 & 1 \\ -2 & 6 & -2 & -2 \\ 1 & -3 & 1 & 1 \end{array}$$

Rearranging the rows and columns to put the one-cell interaction in a corner emphasizes the pattern:

$$\begin{array}{rrrr} 6 & -2 & -2 & -2 \\ -3 & 1 & 1 & 1 \\ -3 & 1 & 1 & 1 \end{array}$$

Example 9.8 One-cell interaction

Consider the data in Table 9.5 (Table 1 of Oehlert 1994). These data are responses from an experiment with four factors, each at two levels labeled low and high, and replicated twice. A standard factorial ANOVA of these data shows that all main effects and interactions are highly significant, and analysis of the residuals reveals no problems.

Table 9.5: Data from a replicated four-factor experiment. All factors have two levels, labeled low and high. Data set `OneCell`.

			D			
A	B	C	low		high	
low	low	low	26.1	27.5	23.5	21.1
low	low	high	22.8	23.8	30.6	32.5
low	high	low	22.0	20.2	28.1	29.9
low	high	high	30.0	29.3	38.3	38.5
high	low	low	11.4	11.0	20.4	22.0
high	low	high	22.3	20.2	28.7	28.8
high	high	low	18.9	16.4	26.6	26.5
high	high	high	29.6	29.8	34.5	34.9

```
1 > fit <- lm(response~A*B*C*D,data=OneCellExample)
2 > anova(fit)
            Df Sum Sq Mean Sq  F value      Pr(>F)
A            1 120.90  120.90 117.4511 8.871e-09 ***
B            1 204.02  204.02 198.1979 1.970e-10 ***
C            1 472.78  472.78 459.2896 3.288e-13 ***
D            1 335.40  335.40 325.8336 4.621e-12 ***
A:B          1  18.00   18.00  17.4863 0.0007050 ***
A:C          1  24.85   24.85  24.1421 0.0001559 ***
B:C          1  27.38   27.38  26.5987 9.541e-05 ***
A:D          1  15.12   15.12  14.6934 0.0014664 **
B:D          1  10.81   10.81  10.5027 0.0051192 **
C:D          1   6.48    6.48   6.2951 0.0232492 *
A:B:C        1  11.52   11.52  11.1913 0.0041075 **
A:B:D        1  34.03   34.03  33.0601 2.985e-05 ***
A:C:D        1  50.00   50.00  48.5732 3.161e-06 ***
B:C:D        1  22.11   22.11  21.4803 0.0002754 ***
A:B:C:D      1  13.78   13.78  13.3880 0.0021183 **
Residuals 16  16.47    1.03
```

In an attempt to understand the interaction, we make an interaction plot on line 3, as shown in Figure 9.2. The lines look parallel, except the treatment with all factors low deviates from the pattern. Note that casual inspection of the data could have suggested that the treatment with mean 11.2 is the interacting cell, but that is incorrect. Line 4 creates a dummy (indicator) variable that is all zero except for the treatment with all factors low.

```
 3 > with(OneCellExample,interactplot(A:B,C:D,response))
 4 > all.low <- rep(0,32);alllow[1:2] <- 1
 5 > fit2 <- lm(response~A+B+C+D+all.low+A:B:C:D,data=OneCellExample)
 6 > anova(fit2)
             Df Sum Sq Mean Sq  F value     Pr(>F)
   A          1 120.90  120.90 117.4511 8.871e-09 ***
   B          1 204.02  204.02 198.1979 1.970e-10 ***
   C          1 472.78  472.78 459.2896 3.288e-13 ***
   D          1 335.40  335.40 325.8336 4.621e-12 ***
   all.low    1 217.35  217.35 211.1485 1.229e-10 ***
   A:B:C:D   10  16.74    1.67   1.6263    0.1861
   Residuals 16  16.47    1.03
```

Line 5 fits a model that includes main effects of the factor, the one-cell indicator variable, and all the rest of the interaction in the data rolled up into one term. Note that inclusion of the indicator variable makes the model nonorthogonal; the data are still balanced, but the model we fit is not the simple orthogonal effects model, and we need to treat it similarly to a situation where the data are unbalanced. The (Type I) ANOVA on line 6 shows that the dummy variable is highly significant, and there is no indication of other interaction remaining in the data.

If we want to look at coefficients, we need to refit the model using only the terms of interest, because inclusion of unneeded terms will affect the coefficients.

```
 7 > fit3 <- lm(response~A+B+C+D+all.low,data=OneCellExample)
 8 > summary(fit3)
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)  25.0330     0.2087 119.962  < 2e-16 ***
 A1            1.1580     0.2087   5.549 7.94e-06 ***
 B1           -3.3108     0.2087 -15.866 6.88e-15 ***
 C1           -4.6295     0.2087 -22.186  < 2e-16 ***
 D1           -4.0233     0.2087 -19.280  < 2e-16 ***
 all.low      12.5727     0.9638  13.045 6.45e-13 ***

 Residual standard error: 1.13 on 26 degrees of freedom
 Multiple R-squared:  0.976,Adjusted R-squared:  0.9714
 F-statistic: 211.5 on 5 and 26 DF,  p-value: < 2.2e-16
```

It looks like the interacting treatment is about 12.6 units higher than the additive model fit to the rest of the data would suggest.

9.4.2 Quantitative factors

A second type of interaction that can be easily modeled occurs when one or more of the factors have quantitative levels (doses). First consider the situation when the interacting factors are all quantitative. Suppose that the doses for factor A are z_{Ai}, and those for factor B are z_{Bj}. We can build a polynomial regression model for cell means as

Polynomial models for quantitative factors

$$\mu_{ij} = \theta_0 + \sum_{r=1}^{a-1} \theta_{Ar} z_{Ai}^r + \sum_{s=1}^{b-1} \theta_{Bs} z_{Bj}^s + \sum_{r=1}^{a-1}\sum_{s=1}^{b-1} \theta_{ArBs} z_{Ai}^r z_{Bj}^s \quad .$$
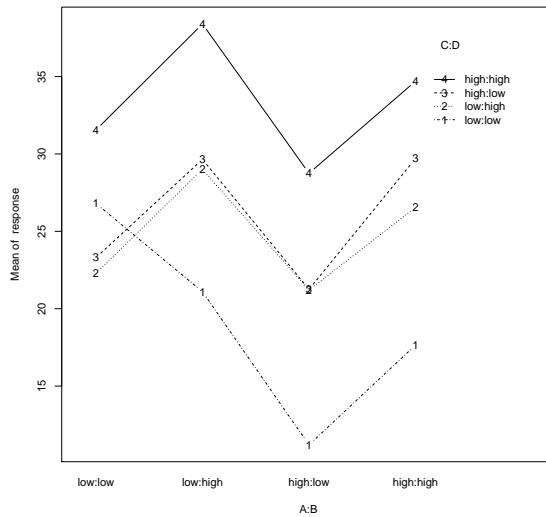
Figure 9.2: Interaction plot for data in Table 9.5.

Polynomial terms in z_{Ai} model the main effects of factor A, polynomial terms in z_{Bj} model the main effects of factor B, and cross product terms model the AB interaction. Models of this sort are most useful when relatively few of the polynomial terms are needed to provide an adequate description of the response.

A polynomial term $z_{Ai}^r z_{Bj}^s$ is characterized by its exponents (r, s). A term with exponents (r, s) is "above" a term with exponents (u, v) if $r \leq u$ *and* $s \leq v$; we also say that (u, v) is below (r, s). The mnemonic here is that in an ANOVA table, simpler terms (such as main effects) are above more complicated terms (such as interactions). This is a little confusing, because we also use the phrase *higher order* for the more complicated terms, but higher order terms appear below the simpler terms.

Lower powers are above higher powers

A term in this polynomial model is needed if its own sum of squares is large, or if it is above a term with a large sum of squares. This preserves a polynomial hierarchy. Non-hierarchical polynomial models only make sense when there is an unambiguous zero for the variable. For example, a non-hierarchical, quadratic-only model in degrees F becomes a model with linear and quadratic terms when re-expressed in degrees C. Even for things like mass or length where there is an unambiguous zero, we often need to recenter the variables to make the problem numerically stable. For example, in an experiment where the factor length ranges from 99.9 to 100.1 meters, we will be much better off subtracting 100 from the length and working with polynomials on the base factor ranging from –.1 to .1.

Use hierarchical polynomial models

The Type II sum of squares for a term is the difference in error sums of

squares for two models: the model with all terms that are not below the term of interest and the model of all terms that are not below the term of interest plus the term of interest. Thus, the sum of squares for the term $z_{Ai}^2 z_{Bi}^1$ is the error sum of squares for the model with terms z_{Ai}, z_{Ai}^2, z_{Bi} and $z_{Ai} z_{Bi}$, less the error sum of squares for the model with terms z_{Ai}, z_{Ai}^2, z_{Bi}, $z_{Ai} z_{Bi}$, and $z_{Ai}^2 z_{Bi}^1$.

Computing polynomial sums of squares

Computation of the polynomial sums of squares can usually be accomplished in statistical software with one command. Recall, however, that the polynomial coefficients θ depend on what other polynomial terms are in a given regression model. Thus if we determine that only linear and quadratic terms are needed, we must refit the model with just those terms to find their coefficients when the higher order terms are omitted. In particular, you should not use coefficients from the full model when predicting with a model with fewer terms. Use the full model MS_E for determining which terms to include, but use coefficients computed for a model including just your selected terms.

Compute polynomial coefficients for final model including only selected terms

Example 9.9 Amylase activity, continued

Recall the amylase specific activity data of Example 8.14. The three factors are analysis temperature, growth temperature, and variety. On the log scale, the analysis temperature by growth temperature interaction (both quantitative variables) was marginally significant. Let us explore the main effects and interactions using quantitative variables. We begin by fitting using orthogonal polynomials as shown on lines 1–2.

```
1 > fit <- lm(log(amylase)~poly(aTemp.z,7)*poly(gTemp.z,1)*variety,
    data=AmylaseActivity)
2 > summary(fit)
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   5.805966   0.007544 769.610  < 2e-16 ***
poly(aTemp.z, 7)1                             0.935612   0.073916  12.658  < 2e-16 ***
poly(aTemp.z, 7)2                            -1.445585   0.073916 -19.557  < 2e-16 ***
poly(aTemp.z, 7)3                            -0.204923   0.073916  -2.772 0.007280 **
poly(aTemp.z, 7)4                             0.053281   0.073916   0.721 0.473641
poly(aTemp.z, 7)5                            -0.001156   0.073916  -0.016 0.987566
poly(aTemp.z, 7)6                             0.058510   0.073916   0.792 0.431536
poly(aTemp.z, 7)7                            -0.052764   0.073916  -0.714 0.477924
poly(gTemp.z, 1)                              0.066178   0.073916   0.895 0.373976
variety1                                      0.078367   0.007544  10.388 2.31e-15 ***
poly(aTemp.z, 7)1:poly(gTemp.z, 1)            1.844241   0.724228   2.546 0.013298 *
poly(aTemp.z, 7)2:poly(gTemp.z, 1)            0.092453   0.724228   0.128 0.898820
poly(aTemp.z, 7)3:poly(gTemp.z, 1)            1.671746   0.724228   2.308 0.024224 *
poly(aTemp.z, 7)4:poly(gTemp.z, 1)            0.772194   0.724228   1.066 0.290325
poly(aTemp.z, 7)5:poly(gTemp.z, 1)           -0.813064   0.724228  -1.123 0.265775
poly(aTemp.z, 7)6:poly(gTemp.z, 1)            0.307444   0.724228   0.425 0.672615
poly(aTemp.z, 7)7:poly(gTemp.z, 1)            0.474707   0.724228   0.655 0.514517
poly(aTemp.z, 7)1:variety1                   -0.033144   0.073916  -0.448 0.655382
poly(aTemp.z, 7)2:variety1                    0.137003   0.073916   1.853 0.068422 .
poly(aTemp.z, 7)3:variety1                   -0.046678   0.073916  -0.632 0.529962
poly(aTemp.z, 7)4:variety1                    0.029442   0.073916   0.398 0.691721
poly(aTemp.z, 7)5:variety1                   -0.061100   0.073916  -0.827 0.411531
poly(aTemp.z, 7)6:variety1                    0.002074   0.073916   0.028 0.977707
poly(aTemp.z, 7)7:variety1                    0.030507   0.073916   0.413 0.681185
poly(gTemp.z, 1):variety1                     0.293245   0.073916   3.967 0.000186 ***
poly(aTemp.z, 7)1:poly(gTemp.z, 1):variety1 -0.011619   0.724228  -0.016 0.987250
poly(aTemp.z, 7)2:poly(gTemp.z, 1):variety1  0.165103   0.724228   0.228 0.820396
poly(aTemp.z, 7)3:poly(gTemp.z, 1):variety1 -1.976362   0.724228  -2.729 0.008195 **
poly(aTemp.z, 7)4:poly(gTemp.z, 1):variety1 -0.055415   0.724228  -0.077 0.939247
poly(aTemp.z, 7)5:poly(gTemp.z, 1):variety1  0.746643   0.724228   1.031 0.306445
poly(aTemp.z, 7)6:poly(gTemp.z, 1):variety1  0.017989   0.724228   0.025 0.980261
poly(aTemp.z, 7)7:poly(gTemp.z, 1):variety1 -0.281487   0.724228  -0.389 0.698809
```

Here are a few things to note.

1. None of the terms involving powers 4 or higher in analysis temperature is significant.

2. The cubic-in-analysis-temperature by linear-in-growth-temperature by variety term is significant.

3. No terms in the analysis-temperature by variety interaction are significant.

4. The linear-by-linear and cubic-by-linear terms in the analysis-temperature by growth-temperature interaction are modestly significant.

We can quantify item 1 by refitting using only the first three powers and comparing the models.

```
 3 > fit123 <- lm(log(amylase)~poly(aTemp.z,3)*poly(gTemp.z,1)*variety,
     data=AmylaseActivity)
 4 > anova(fit123,fit)
   Analysis of Variance Table

   Model 1: log(amylase) ~ poly(aTemp.z, 3) * poly(gTemp.z, 1) * variety
   Model 2: log(amylase) ~ poly(aTemp.z, 7) * poly(gTemp.z, 1) * variety
     Res.Df     RSS Df Sum of Sq     F Pr(>F)
   1     80 0.38735
   2     64 0.34967 16   0.03768 0.431 0.9683
```

Line 4 shows that the 16 degrees of freedom for powers 4–7 are not significant overall.

Item two brings a quandary. That cubic-in-analysis-temperature by linear-in-growth-temperature by variety term looks significant, but because we want to work with hierarchical polynomials, including that term means that we must also include five other (non-significant) terms: linear-in-analysis-temperature by linear-in-growth-temperature by variety, quadratic-in-analysis-temperature by linear-in-growth-temperature by variety, and linear-, quadratic-, and cubic-in-analysis-temperature by variety. We can test whether that one apparently significant term is sufficient to justify five other non-significant terms by fitting and comparing a reduced model without those six degrees of freedom.

```
 5 > fit123red1 <- lm(log(amylase)~poly(aTemp.z,3)*poly(gTemp.z,1)+
     variety*poly(gTemp.z,1),data=AmylaseActivity)
 6 > anova(fit123red1,fit123,fit)
   Analysis of Variance Table

   Model 1: log(amylase) ~ poly(aTemp.z, 3) * poly(gTemp.z, 1) + variety *
     poly(gTemp.z, 1)
   Model 2: log(amylase) ~ poly(aTemp.z, 3) * poly(gTemp.z, 1) * variety
   Model 3: log(amylase) ~ poly(aTemp.z, 7) * poly(gTemp.z, 1) * variety
     Res.Df     RSS Df Sum of Sq     F Pr(>F)
   1     86 0.45037
   2     80 0.38735  6   0.06302 1.9224 0.0906 .
   3     64 0.34967 16   0.03768 0.4310 0.9683
```

In this case, the combined six degree of freedom term is not significant, with a p-value of .09.

Item three is consistent with what we saw when we considered analysis temperature as an eight level categorical factor; no individual polynomial terms within it are significant.

Item four illustrates a bothersome phenomenon—the averaging involved in multi-degree-of-freedom mean squares can obscure some interesting effects in a cloud of uninteresting effects. The seven degree-of-freedom growth temperature by analysis temperature interaction is marginally significant with a p-value of .054, but two of the individual degrees of freedom in that seven degree-of-freedom bundle are rather more significant. We now test whether the linear-, quadratic-, and cubic by linear effects are significant when considered together as a three degree of freedom effect (recall that quadratic by linear is not significant by itself).

```
7 > fit123red2 <- lm(log(amylase)~poly(aTemp.z,3)+poly(gTemp.z,1)+
     variety*poly(gTemp.z,1),data=AmylaseActivity)
8 > anova(fit123red2,fit123red1,fit123,fit)
 Analysis of Variance Table

 Model 1: log(amylase) ~ poly(aTemp.z, 3) + poly(gTemp.z, 1) + variety *
     poly(gTemp.z, 1)
 Model 2: log(amylase) ~ poly(aTemp.z, 3) * poly(gTemp.z, 1) + variety *
     poly(gTemp.z, 1)
 Model 3: log(amylase) ~ poly(aTemp.z, 3) * poly(gTemp.z, 1) * variety
 Model 4: log(amylase) ~ poly(aTemp.z, 7) * poly(gTemp.z, 1) * variety
   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
 1     89 0.51500
 2     86 0.45037  3   0.06463 3.9431 0.01207 *
 3     80 0.38735  6   0.06302 1.9224 0.09060 .
 4     64 0.34967 16   0.03768 0.4310 0.96825
```

These three degrees of freedom are jointly significant with a p-value of about .01.

Orthogonal polynomials are a good choice for model selection, but they are less understandable once we want to look at model coefficients. For that, it makes sense to go back to ordinary polynomials.

```
9 > fit2 <- lm(log(amylase)~(aTemp.z+I(aTemp.z^2)+I(aTemp.z^3))*gTemp.z+
     variety*gTemp.z,data=AmylaseActivity)
0 > summary(fit2)
                       Estimate Std. Error t value Pr(>|t|)
 (Intercept)          6.105e+00  4.589e-01  13.306  < 2e-16 ***
 aTemp.z             -1.039e-01  6.559e-02  -1.585 0.116704
 I(aTemp.z^2)         6.847e-03  2.815e-03   2.432 0.017085 *
 I(aTemp.z^3)        -1.154e-04  3.722e-05  -3.101 0.002608 **
 gTemp.z             -5.603e-02  2.303e-02  -2.433 0.017051 *
 variety1            -1.641e-02  2.453e-02  -0.669 0.505274
 aTemp.z:gTemp.z      7.775e-03  3.292e-03   2.362 0.020438 *
 I(aTemp.z^2):gTemp.z -3.293e-04  1.413e-04  -2.331 0.022115 *
 I(aTemp.z^3):gTemp.z  4.404e-06  1.868e-06   2.358 0.020655 *
 gTemp.z:variety1     4.988e-03  1.231e-03   4.052 0.000111 ***
```

If we let z_A denote the level of analysis temperature, and we let z_B denote

the level of growth temperature, then our model for the mean is

$$
\begin{aligned}
\mu_{ij1} &= 6.105 - .1039z_A + .006847z_A^2 - .0001154z_A^3 - .05603z_B - \\
&\quad .05603z_B - .01641 + .007775z_Az_B - .0003293z_A^2z_B + \\
&\quad .000004404z_A^3z_B + .004988z_B \\
&= 6.089 - .1039z_A + .006847z_A^2 - .0001154z_A^3 - .05603z_B - \\
&\quad .06102z_B + .007775z_Az_B - .0003293z_A^2z_B + \\
&\quad .000004404z_A^3z_B
\end{aligned}
$$

$$
\begin{aligned}
\mu_{ij2} &= 6.105 - .1039z_A + .006847z_A^2 - .0001154z_A^3 - .05603z_B - \\
&\quad .05603z_B + .01641 + .007775z_Az_B - .0003293z_A^2z_B + \\
&\quad .000004404z_A^3z_B - .004988z_B \\
&= 6.121 - .1039z_A + .006847z_A^2 - .0001154z_A^3 - .05603z_B - \\
&\quad .05104z_B + .007775z_Az_B - .0003293z_A^2z_B + \\
&\quad .000004404z_A^3z_B
\end{aligned}
$$

When there is a combination of quantitative factors and categorical factors as in the preceding example, we will usually have a choice of how to parameterize the model. Typically this choice is between a "separate" polynomial model for each level of the categorial factor (or combination of levels of categorical factors) and a model that can be though of as a central polynomial model and deviations from the central model for each level of the categorial factor. For example, consider

$$
\mu_{ij} = \theta_j + \sum_{r=1}^{a-1} \theta_{Arj} z_{Ai}^r
$$

and

$$
\mu_{ij} = \theta_0 + \beta_j + \sum_{r=1}^{a-1} \theta_{Ar0} z_{Ai}^r + \sum_{r=1}^{a-1} \theta\beta_{Arj} z_{Ai}^r ,
$$

where $\theta_j = \theta_0 + \beta_j$, $\theta_{Arj} = \theta_{Ar0} + \theta\beta_{Arj}$, and the parameters have the zero sum restrictions $\sum_j \beta_j = 0$ and $\sum_j \theta\beta_{Arj} = 0$.

In both forms there is a separate polynomial of degree $a - 1$ in z_{Ai} for each level of factor B. The only difference between these models is how the regression coefficients are expressed. In the first version the constant terms of the model are expressed as θ_j; in the second version the constant terms are expressed as an overall constant θ_0 plus deviations β_j that depend on the qualitative factor. In the first version the coefficients for power r are expressed as θ_{Arj}; in the second version the coefficients for power r are expressed as an overall coefficient θ_{Ar0} plus deviations $\theta\beta_{Arj}$ that depend on the qualitative factor. These are analogous to having treatment means μ_i written as $\mu + \alpha_i$, an overall mean plus treatment effects.

Alternate forms
for regression
coefficients

In the preceding example, we wrote two formulae for μ_{ij1} and μ_{ij2}. In both cases, the first version was the central version, where we expressed the intercept and the slope for z_B as a central value and a deviation from the central value, and the second version was the combined version where we assemble the corresponding coefficients into a single value.

Example 9.10 Transmission of laser light through polyvinyl chloride, continued

Let's return briefly to the laser transmission data of Example 9.10. These data did not show an interaction on the original scale, but if instead of modeling the variance we had moved ahead by modeling transformed reflectance, we would have wound up with the model on line 1 and summarized on line 2.

```
1 > fitbc <- lm(-1/(100-transmission)^1.5~(thickness.z+I(thickness.z^2))*sanding,
    data=LaserTransmission)
2 > summary(fitbc)
...
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -3.232e-02  8.202e-04 -39.410  < 2e-16 ***
thickness.z                3.107e-03  2.764e-04  11.239  < 2e-16 ***
I(thickness.z^2)          -1.325e-04  1.877e-05  -7.058 5.21e-10 ***
sanding1                   1.545e-02  1.160e-03  13.322  < 2e-16 ***
sanding2                   5.984e-03  1.160e-03   5.159 1.73e-06 ***
thickness.z:sanding1      -2.174e-03  3.909e-04  -5.560 3.36e-07 ***
thickness.z:sanding2      -7.491e-04  3.909e-04  -1.916 0.058870 .
I(thickness.z^2):sanding1  1.050e-04  2.655e-05   3.953 0.000164 ***
I(thickness.z^2):sanding2  3.480e-05  2.655e-05   1.311 0.193667
...
```

You can rearrange this model to be a different quadratic curve for each level of sanding. For example, for sanding level 1 (both), the intercept is –.03232 + .01545 = –.01687; the linear coefficient is .003107 – .002174 = .000933; and the quadratic coefficient is –.0001325 + .0001050 = –.0000275. You can request this parameterization directly in **R** as shown on line 3–4.

```
3 > fitbc2 <- lm(-1/(100-transmission)^1.5~0+sanding+thickness.z:sanding+
    I(thickness.z^2):sanding,data=LaserTransmission)
4 > summary(fitbc2)
...
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
sandingboth                 -1.687e-02  1.421e-03 -11.876  < 2e-16 ***
sandingfront                -2.634e-02  1.421e-03 -18.541  < 2e-16 ***
sandingnone                 -5.376e-02  1.421e-03 -37.843  < 2e-16 ***
sandingboth:thickness.z      9.330e-04  4.788e-04   1.949  0.05479 .
sandingfront:thickness.z     2.358e-03  4.788e-04   4.924 4.40e-06 ***
sandingnone:thickness.z      6.029e-03  4.788e-04  12.593  < 2e-16 ***
sandingboth:I(thickness.z^2)  -2.755e-05  3.251e-05  -0.847  0.39931
sandingfront:I(thickness.z^2) -9.770e-05  3.251e-05  -3.005  0.00354 **
sandingnone:I(thickness.z^2)  -2.722e-04  3.251e-05  -8.373 1.38e-12 ***
```

Not only does this save you some computations, it also gives a standard error

for the combined coefficients. In this case, sanding level 1 does not look very quadratic.

I usually find it easier to choose my model using the regular main effects and interactions parameterization, but it is sometimes more interpretable if you can rearrange the model directly into intercepts and polynomial terms.

9.4.3 Tukey one-degree-of-freedom for nonadditivity

The *Tukey one-degree-of-freedom* model for interaction is also called *transformable nonadditivity,* because interaction of this kind can usually be reduced or even eliminated by transforming the response by an appropriate power. (Some care needs to be taken when using this kind of transformation, because the transformation to reduce interaction could introduce nonconstant variance.) The form of a Tukey interaction is similar to that of a linear by linear interaction, but the Tukey model can be used with non-quantitative factors.

Transformable nonadditivity is reduced on the correct scale

The Tukey model can be particularly useful in single replicates, where we have no estimate of pure error and generally must use high-order interactions as surrogate error. If we can transform to a scale that removes much of the interaction, then using high-order interactions as surrogate error is much more palatable.

In a two-factor model, Tukey interaction has the form $\alpha\beta_{ij} = \eta\alpha_i\beta_j/\mu$, for some multiplier η. If interaction is of this form, then transforming the responses with a power $1 - \eta$ will approximately remove the interaction. You may recall our earlier admonition that an interaction effect $\alpha\beta_{ij}$ was not the product of the main effects; well, the Tukey model of interaction for the two-factor model is a multiple of just that product. The Tukey model adds one additional parameter η, so it is a one degree of freedom model for nonadditivity. The form of the Tukey interaction for more general models is discussed in Section 9.6, but it is always a single degree of freedom scale factor times a combination of other model parameters.

Tukey interaction is a scaled product of main effects

There are several algorithms for fitting a Tukey interaction and testing its significance. The following algorithm is fairly general, though somewhat obscure.

Algorithm to fit a Tukey one-degree-of-freedom interaction

1. Fit a preliminary model; this will often be an additive model.

2. Get the predicted values from the preliminary model; square them and divide their squares by twice the mean of the data.

3. Fit the data with a model that includes the preliminary model and the rescaled squared predicted values as explanatory variables.

4. The improvement sum of squares going from the preliminary model to the model including the rescaled squared predicted values is the single degree of freedom sum of squares for the Tukey model.

5. Test for significance of a Tukey type interaction by dividing the Tukey sum of squares by the error mean square from the model including squared predicted terms.

6. The coefficient for the rescaled squared predicted values is $\widehat{\eta}$, an estimate of η. If Tukey interaction is present, transform the data to the power $1 - \widehat{\eta}$ to remove the Tukey interaction.

The transforming power $1 - \eta$ found in this way is approximate and can often be improved slightly.

Example 9.11 CPU page faults, continued

Recall the CPU page fault data from Example 8.12. We originally analyzed those data on the log scale because they simply looked multiplicative. Would we have reached the same conclusion via a Tukey interaction analysis?

On line 1 we fit the additive model as a preliminary model.

```
1 > fit <- lm(faults~alg+init+size+ram,data=PageFaults)
2 > rspv <- predict(fit)^2/mean(PageFaults$faults)/2
3 > fit2 <- lm(faults~alg+init+size+ram+rspv,data=PageFaults)
4 > anova(fit2)
            Df     Sum Sq   Mean Sq F value     Pr(>F)
  alg        1   11671500  11671500  13.106 0.0007432 ***
  init       2   59565822  29782911  33.444 1.258e-09 ***
  size       2  216880816 108440408 121.769 < 2.2e-16 ***
  ram        2  261546317 130773159 146.847 < 2.2e-16 ***
  rspv       1  215332859 215332859 241.801 < 2.2e-16 ***
  Residuals 45   40074226    890538
5 > coef(fit2)
    (Intercept)          alg1         init1         init2         size1
   -426.0884967   -47.0589702  -215.3516316   191.5639849  -208.0754407
          size2          ram1          ram2          rspv
    548.8711338  -269.9239726   693.5339301     0.8987778
```

Line 2 generates the rescaled squared predicted values, line 3 fits the model including the rescaled squared predicted values, and line 4 does an ANOVA. rspv is highly significant. We get its coefficient on line 5. One minus the coefficient is .1, which is close to the log transformation. Thus the Tukey procedure is giving us something similar to what we decided before.

Note that if your preliminary model has interactions, **R** will try to move rspv ahead of any interactions in the model. Thus you will either need to (a) use terms with keep.order set to true to force rspv to be the last term in the model, or (b) compute a Type II ANOVA to get rspv adjusted for all other terms, or (c) use anova to compare the models with and without rspv.

9.4.4   Hidden Additivity

The hidden additivity model of Franck, Nielsen, and Osborne (2013) is appropriate for two-factor models with a single replication. The idea is that the levels of one factor (A) can be partitioned into two sets. Within each subset of levels, there is an additive model for (the subset of levels of) A and B, but the B effects can differ between the two subsets. Thus there are row (A)

effects and two sets of column (B) effects, one set of B effects for each subset of rows. If we invent a pseudo-factor C that indicates the groups of rows in A, then the model is equivalent to `A+B+C:B`.

The trick, of course, is that we do not know what the subsets are, or even if such subsets exist. If this kind of interaction is present, it can usually be detected graphically in the interaction plot. The approach taken in Franck, Nielsen, and Osborne (2013) is to consider all partitions of the levels of A into two subsets and find the partition that yields the highest sum of squares for the interaction term. Test that sum of squares in the usual way, but do a Bonferroni correction to adjust for multiple comparisons.

Example 9.12 Big Sagebrush Seed Viability

Consider the big sagebrush seed viability data of Problem 8.6. Suppose that instead of having access to all the data, we are only given access to the treatment means shown in this table:

Humidity	0	60	120	180	240	300	360
0%	81	80	79	79	81	85	83
32%	82	80	82	79	83	80	81
45%	81	63	57	52	51	39	24

(Column header spanning Days)

We would, by default, use the two factor interaction term as a surrogate error.

```
 1 > sagemeans <- aggregate(viability~storage+humidity,BigSagebrush,mean)
 2 > anova(lm(viability~storage+humidity,sagemeans))
Analysis of Variance Table

Response: viability
          Df Sum Sq Mean Sq F value    Pr(>F)
storage    6  596.2   99.37  0.8612 0.5492679
humidity   2 3825.5 1912.74 16.5757 0.0003524 ***
Residuals 12 1384.7  115.39
```

Line 1 collects the treatment means for the data, and line 2 does an ANOVA with the interaction as error. Relative humidity is fairly significant, but storage time is not.

```
3 > sagemeanstab <- matrix(sagemeans[,3],nr=3,byrow=TRUE)
4 > sageout <- hiddenf::HiddenF(sagemeanstab)
5 > anova(sageout)
  The ACMIF test for the hidden additivity form of interaction
  Analysis of Variance Table

  Response: y
            Df Sum Sq Mean Sq   F value     Pr(>F)
  group      1 3825.4  3825.4 1064.1241 5.520e-08
  col        6  596.2    99.4   27.6432 0.0004036
  row        1    0.1     0.1    0.0186 0.8960723
  group:col  6 1363.2   227.2   63.1988 3.692e-05
  Residuals  6   21.6     3.6
  C.Total   20 5806.5
  (Pvalues in ANOVA table are NOT corrected for multiplicity.)
6 > sageout$adjpvalue
  [1] 0.0001107486
7 > plot(sageout)
```

The `HiddenF` function in the `hiddenf` package fits the hidden additivity
model, but it takes as input a matrix of means rather than variables indicating
rows and columns. Line 3 creates that matrix of means, line 4 fits the hidden
additivity model, and line 5 shows the ANOVA. The line for column (stor-
age) is the same in both ANOVAs, and the lines for row and group sum to
the SS and df for humidity. The hidden additivity interaction term accounts
for 1363.2 of the 1398.7 (98%) of the residual sum of squares in the additive
model. The residual mean square decreases from 115.4 in the additive model
to 3.6 in the hidden additivity model. With this decrease, both the overall
storage effect and interaction are highly significant.

Line 6 shows the p-value after Bonferroni adjustment; there are only three
potential groupings, so this just multiplies the interaction p-value by three.
This kind of interaction is highly significant even after adjustment. Finally,
line 7 plots the results (see Figure 9.3), showing how humidities 1 and 2 are
similar and very dissimilar to humidity 3.

9.5   Two-Series Factorials

A *two-series* factorial design is one in which all the factors have just two
levels. For k factors, we call this a 2^k design, because there are 2^k different
factor-level combinations. Similarly, a design with k factors, each with three
levels, is a three-series design and denoted by 3^k. Two-series designs are
somewhat special, because they are the smallest designs with k factors. They
are often used when screening many factors.

All factors have
exactly two levels
in two-series
factorials

Because two-series designs are so common, there are special notations
and techniques associated with them. The two levels for each factor are gen-
erally called *low* and *high*. These terms have clear meanings if the factors are
quantitative, but they are often used as labels even when the factors are not
quantitative. Note that "off" and "on" would work just as well, but low and
high are the usual terms.

Levels called low
and high

Hidden Additivity Plot
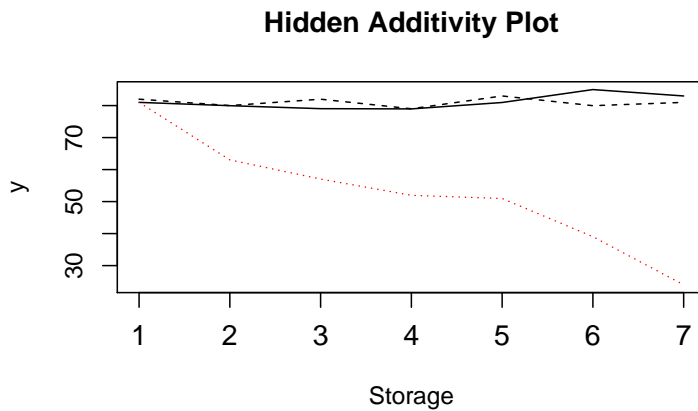
Figure 9.3: Hidden additivity plot for Big Sagebrush treatment means.

There are two methods for denoting a factor-level combination in a two-series design. The first uses letters and is probably the more common. Denote a factor-level combination by a string of lower-case letters: for example, bcd. We have been using these lower-case letters to denote the number of levels in different factors, but all factors in a two-series design have two levels, so there should be no confusion. Letters that are present correspond to factors at their high levels, and letters that are absent correspond to factors at their low levels. Thus ac is the combination where factors A and C are at their high levels and all other factors are at their low levels. Use the symbol (1) to denote the combination where all factors are at their low levels. Denote the mean response at a given factor-level combination by $\overline{y}$ with a subscript, for example $\overline{y}_{ab}$. Do not confuse the factor-level combination bc with the interaction BC; the former is a single treatment, and the latter is a contrast among treatments.

The second method uses numbers and generalizes to three-series and higher-order factorials as well. A factor-level combination is denoted by k binary digits, with one digit giving the level of each factor: a zero denotes a factor at its low level, and a one denotes a factor at its high level. Let x_A (either 0 or 1) be the level of factor A, x_B the level of factor B, and so on. The collection of digits indicates a treatment (factor-level combination). For three factors, describe the treatment as $x_C x_B x_A$. Thus 000 is all factors at low level, the same as (1), and 110 is factors B and C at high level, the same as bc. This generalizes to other factorials by using more digits. For example, we use the digits 0, 1, and 2 to denote the three levels of a three-series.

Note that you could just as easily use the digits in the order A, B, C; you just need to keep track of the order you are using. This order might seem more natural, but the order C, B, A has a slight advantage of interpretation

Table 9.6: Pluses and minuses for a 2^3 design.

	A	B	C
(1)	−	−	−
a	+	−	−
b	−	+	−
ab	+	+	−
c	−	−	+
ac	+	−	+
bc	−	+	+
abc	+	+	+

we will see shortly.

It is customary to arrange the factor-level combinations of a two-series factorial in *standard order*. Standard order will help us keep track of factor-level combinations when we later modify two-series designs. Standard order for a two-series design begins with (1). Then proceed through the remainder of the factor-level combinations with factor A varying fastest, then factor B, and so on. In standard order, factor A will repeat the pattern low, high; factor B will repeat the pattern low, low, high, high; factor C will repeat the pattern low, low, low, low, high, high, high, high; and so on though other factors. In general, the jth factor will repeat a pattern of 2^{j-1} lows followed by 2^{j-1} highs. For a 2^4, standard order is (1), a, b, ab, c, ac, bc, abc, d, ad, bd, abd, cd, acd, bcd, and $abcd$.

When using binary digits (in the order suggested above, for example, CBA) to indicate factor levels, we find that standard order is numerical order. That is, standard order for a 2^3 design is 000, 001, 010, 011, 100, 101, 110, 111. This is the reason for using the reverse order of the digits.

Two-series factorials form the basis of several designs we will consider later, and one of the tools we will use is a table of pluses and minuses. For a 2^k design, build a table with 2^k rows and k columns. The rows are labeled with factor-level combinations in standard order, and the columns are labeled with the k factors. In principle, the body of the table contains $+1$'s and $−1$'s, with $+1$ indicating a factor at a high level, and $−1$ indicating a factor at a low level. In practice, we use just plus and minus signs to denote the factor levels. Table 9.6 shows this table for a 2^3 design.

Table of + and −

9.5.1 Contrasts

One nice thing about a two-series design is that every main effect and interaction is just a single degree of freedom, so we may represent any main effect or interaction by a single contrast. For example, the main effect of factor A

in a 2^3 can be expressed as

$$
\begin{aligned}
\widehat{\alpha}_2 &= -\widehat{\alpha}_1 \\
&= (\bar{y}_{2\bullet\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet})/2 \\
&= \frac{1}{8}(\bar{y}_a + \bar{y}_{ab} + \bar{y}_{ac} + \bar{y}_{abc} - \bar{y}_{(1)} - \bar{y}_b - \bar{y}_c - \bar{y}_{bc}) \\
&= \frac{1}{8}(-\bar{y}_{(1)} + \bar{y}_a - \bar{y}_b + \bar{y}_{ab} - \bar{y}_c + \bar{y}_{ac} - \bar{y}_{bc} + \bar{y}_{abc}) \;,
\end{aligned}
$$

which is a contrast in the eight treatment means with plus signs where A is high and minus signs where A is low. Similarly, the sum of squares for A can be written

$$
\begin{aligned}
\text{SS}_\text{A} &= 4n\widehat{\alpha}_1{}^2 + 4n\widehat{\alpha}_2{}^2 \\
&= \frac{n}{8}(\bar{y}_a + \bar{y}_{ab} + \bar{y}_{ac} + \bar{y}_{abc} - \bar{y}_{(1)} - \bar{y}_b - \bar{y}_c - \bar{y}_{bc})^2 \\
&= \frac{n}{8}(-\bar{y}_{(1)} + \bar{y}_a - \bar{y}_b + \bar{y}_{ab} - \bar{y}_c + \bar{y}_{ac} - \bar{y}_{bc} + \bar{y}_{abc})^2 \;,
\end{aligned}
$$

which is the sum of squares for the contrast w_A with coefficients $+1$ where A is high and -1 where A is low (or .25 and $-.25$, or -17.321 and 17.321, as the sum of squares is unaffected by a nonzero multiplier for the contrast coefficients). Note that this contrast w_A has exactly the same pattern of pluses and minuses as the column for factor A in Table 9.6.

The difference

$$
\bar{y}_{2\bullet\bullet\bullet} - \bar{y}_{1\bullet\bullet\bullet} = \widehat{\alpha}_2 - \widehat{\alpha}_1 = 2\widehat{\alpha}_2
$$

is the *total effect* of factor A. The total effect is the average response where A is high, minus the average response where A is low, so we can also obtain the total effect of factor A by rescaling the contrast w_A

$$
\bar{y}_{2\bullet\bullet\bullet} - \bar{y}_{1\bullet\bullet\bullet} = \frac{1}{4}\sum_{ijk} w_{Aijk}\, \bar{y}_{ijk\bullet} \;,
$$

where the divisor of 4 is replaced by 2^{k-1} for a 2^k design.

The columns of Table 9.6 give us contrasts for the main effects. Interactions in the two-series are also single degrees of freedom, so there must be contrasts for them as well. We obtain these interaction contrasts by taking elementwise products of main-effects contrasts. For example, the coefficients in the contrast for the BC interaction are the products of the coefficients for the B and C contrasts. A three-way interaction contrast is the product of the three main-effects contrasts, and so on. This is most easily done by referring to the columns of Table 9.6, with $+$ and $-$ interpreted as $+1$ and -1. We show these contrasts for a 2^3 design in Table 9.7.

Table 9.7: All contrasts for a 2^3 design.

	A	B	C	AB	AC	BC	ABC
(1)	–	–	–	+	+	+	–
a	+	–	–	–	–	+	+
b	–	+	–	–	+	–	+
ab	+	+	–	+	–	–	–
c	–	–	+	+	–	–	+
ac	+	–	+	–	+	–	–
bc	–	+	+	–	–	+	–
abc	+	+	+	+	+	+	+

9.5.2 Single replicates

As with all factorials, a single replication in a two-series design means that we have no degrees of freedom for error. We can apply any of the usual methods for single replicates to a two-series design, but there are also methods developed especially for single replicate two-series. We describe three of these methods. The first is graphically based and is subjective; it does not provide p-values. The second is just slightly more complicated (it can be done by hand, if need be), but it does allow at least approximate testing; however, it assumes effect sparsity, as defined below. The third is computationally intensive, but it produces tests with good error control and no need for the effect sparsity assumption.

Single replicates need an estimate of error

The first two methods are based on the idea that if our original data are independent and normally distributed with constant variance, then the effects contrasts in Table 9.7 give us results that are also independent and normally distributed with constant variance. The expected value of any of these contrasts is zero if the corresponding null hypothesis of no main effect or interaction is true. If that null hypothesis is not true, then the expected value of the contrast is not zero. So, when we look at the results, contrasts corresponding to null effects should look like an independent sample from a normal distribution with mean zero and constant variance, and contrasts corresponding to non-null effects will also be independent with the same variance, but they will have different means and should look like outliers.

Effects are independent with constant variance

Significant effects are outliers

We now need a technique to identify outliers, and to do that we need to assume that most of the data are not outliers. That is, we need to assume that there are relatively few effects that are not null. This is *effect sparsity*. The first two techniques will work poorly if there are many non-null effects, because we won't have a good basis for deciding what null behavior is.

We assume effect sparsity

The first method is graphical and is attributed to Daniel (1959). Simply make a half-normal probability plot of the absolute values of the observed contrasts and look for outliers. We use absolute values, because don't care about the signs of the effects when determining which ones are outliers. A half-normal probability plot plots the sorted absolute values on the horizontal axis against the sorted expected scores from a half-normal distribution (that is, the expected value of ith smallest absolute value from a sample of size

Half-normal plot of effects

$2^k - 1$ from a normal distribution) on the vertical axis. If all the effects are null, this plot should look roughly linear. Non-null effects should appear as outliers to the right.

The choice of whether to put the data on the horizontal and normal scores on the vertical, or vice versa, is arbitrary, but we use this orientation to be analogous with a second graphical presentation we will see shortly. In addition, you can instead use a normal plot of effects instead of a half-normal plot of absolute effects, but I usually find the half-normal plots easier to interpret.

The second method computes a *pseudo-standard error* (PSE) for the contrasts, allowing us to do t-tests. Lenth (1989) computes the PSE in two steps. First, let s_0 be 1.5 times the median of the absolute values of the contrast results. Second, delete any contrasts results whose absolute values are greater than $2.5s_0$, and let the PSE be 1.5 times the median of the remaining absolute contrast results. Treat the PSE as a standard error for the contrasts with $(2^k - 1)/3$ degrees of freedom, and do t-tests. These can be individual tests, or you can do simultaneous tests using a Bonferroni correction.

The third method is the permutation (randomization) step up test of Basso and Salmaso (2006). Denote one of the factorial effects generically as $\hat{\beta}_i$, for $i = 1, ..., 2^k - 1$. Sort the squared effects into order: $\hat{\beta}^2_{(1)} \leq \hat{\beta}^2_{(2)} \leq \cdots \leq \hat{\beta}^2_{(2^k-1)}$. Form $2^k - 2$ test statistics:

$$
\begin{aligned}
h_2 &= \frac{\hat{\beta}^2_{(2)}}{\hat{\beta}^2_{(1)}} \\[2mm]
h_3 &= \frac{\hat{\beta}^2_{(3)}}{\hat{\beta}^2_{(1)} + \hat{\beta}^2_{(2)}} \\[2mm]
\vdots \quad &\quad \vdots \\[2mm]
h_{2^k-1} &= \frac{\hat{\beta}^2_{(2^k-1)}}{\hat{\beta}^2_{(1)} + \hat{\beta}^2_{(2)} + \cdots + \hat{\beta}^2_{(2^k-2)}}
\end{aligned}
$$

Start with h_2 and work your way toward h_{2^k-1}, testing in a step up fashion. That is, if you ever decide that $\hat{\beta}_{(i)}$ is non-zero, then you are deciding that all effects from $\hat{\beta}_{(i)}$ through $\hat{\beta}_{(2^k-1)}$ are also non-zero.

The null distribution for h_i is that created by randomly permuting the data and then computing h_i on the permuted data (call it $\tilde{h}_i$ to indicate that it is computed on permuted data). In general, we generate a large number of permutations, say 10,000, and compute $\tilde{h}_i$ on each of them. The p-value for h_i is the fraction of $\tilde{h}_i$ values that exceed h_i. We are potentially doing $2^k - 2$ tests, so we use a Bonferroni correction to control the experimentwise error rate. Given the step up nature of the procedure, we might expect that Basso-Salmaso also controls the strong familywise error rate; simulation studies suggest that it does. Note that this test involves simulation to get critical

Table 9.8: Performance index for ramp meters at one ramp on I-94. Minimum release rate, demand increment, and maximum release rate in vehicles per hour; maximum wait in minutes. Data adapted from A. Beegala; data set `RampMeters`.

| | | Max. wait/Min. release | | | |
| | | 3.5 | | 6 | |
Demand	Max. release	150	350	150	350
125	1650	6.9	11.5	36.8	69.0
	1800	6.9	17.3	29.9	69.0
250	1650	10.4	13.8	39.1	66.7
	1800	18.4	5.7	42.6	59.8

values, so it is possible to get different results on successive applications of the procedure if an effect is near the critical value.

One common presentation for the results of either the PSE or Basso-Salmaso results is to make a half-normal plot of the effects and then either (1) mark or label the points that correspond to significant effects or (2) indicate the individual and/or simultaneous PSE cutoffs for significance on the plot. Recognizing that with PSE or Basso-Salmaso we no longer need to visually seek outliers, a second common presentation is to use a Pareto plot instead of a half-normal plot. In the Pareto plot, the half-normal score plotting positions are replaced by the ranks of the (absolute) effects. The plot itself can be a bar chart of the effects or a set of horizontal line segments ending in the plotting point. We can add labels or cutoffs to the Pareto plot in the same way we do for a half-normal plot.

Half-normal or Pareto plot of results

Example 9.13 Ramp meters

Ramp meters are stop lights on the entrance ramps to freeways in urban settings. They are used to control the number of cars trying to merge onto the freeway at any given time. Proper use of a ramp meter can lead to shorter overall waiting times and fewer accidents. This experiment considers the effects of four factors on a performance index for the ramp. The experimental factors are minimum release rate (150 or 350 vehicles per hour), maximum allowed waiting time (3.5 or 6 minutes), maximum release rate on the ramp (1650 or 1800 vehicles per hour), and an increment in demand above normal traffic (125 or 250 vehicles/hour). The performance index is a combination of the number of vehicles that need to wait longer than the desired maximum, the total time waited above the desired maximum, and the total distance the queue on the ramp extends beyond the beginning of the ramp.

This experiment is not performed on the driving public, but rather is performed using a traffic simulator for the entire transportation system in a region. The simulator is extremely complex and includes many random inputs, so the response to adjusting parameters like the ramp meter parameters can only be determined by experiment. Data for this experiment are in Table 9.8.

We begin by fitting the full factorial model, as shown on line 1.

```
1 > fit <- lm(PI~minRel*maxWait*maxRel*Incr,data=RampMeters)
2 > TwoSeriesPlots(fit,pse=FALSE,alpha=.01,type="normal")
3 > TwoSeriesPlots(fit,pse=TRUE,alpha=.01,type="normal")
4 > TwoSeriesPlots(fit,pse=TRUE,alpha=.05,type="normal")
5 > TwoSeriesPlots(fit,pse=TRUE,alpha=.05,type="pareto")
```

The `TwoSeriesPlots` function can plot either half-normal or Pareto plots. Terms significant by the Basso-Salmaso test are marked with a solid dot. You may also request the Lenth PSE individual and simultaneous critical values be marked (by vertical lines), and any additional terms significant according to the individual PSE test be labeled (although they will not have solid dot markings).

Line 2 creates the half-normal plot at the .01 level with only the Basso-Salmaso test, as shown in panel one of Figure 9.4. Maximum wait time, minimum release rate, and their interaction are significant. The line matches the non-significant points. Line 3 adds PSE information to the plot, as shown in panel two. The dashed and dotted lines show the individual and simultaneous (Bonferroni) critical values for the Lenth procedure. Two additional terms are significant according to Lenth, and these are labeled on the plot, although they retain open dots. Line 4 creates an analogous plot for the .05 level (panel three); here the null terms fit the null line very well. Finally, line 5 creates the analogous Pareto plot instead of a half-normal plot (panel four).

Historically, subjective consideration of the half-normal plot was our first-line tool for analyzing single replicates of two-series designs. However, with the Basso-Salmaso and PSE procedures to select significant terms, the somewhat cleaner Pareto presentation is preferred by some.

An interesting feature of two-series factorials can be seen if you look at a data set consisting of all zeroes except for a single nonzero value. All factorial effects for such a data set are equal in absolute value, but some will be positive and some negative, depending on which data value is nonzero and the pattern of pluses and minuses. What this means is that a situation where many of the two-series effects are roughly equal in size and do not trend down toward zero probably indicates an outlier in the data, or possibly a strong one-cell interaction.

A single nonzero response yields effects equal in absolute value

With some careful study of the pattern of signs of the two-series effects and the pattern of pluses and minuses in the two-series contrasts, you can usually determine which point is the outlier. For example, suppose that c has a positive value and all other responses are zero. Looking at the row for c in Table 9.7, the effects for C, AB, and ABC should be positive, and the effects for A, B, AC, and BC should be negative. Similarly, if bc had a negative value and all other responses were zero, then the row for bc shows us that A, AB, AC, and ABC would be positive, and B, C, and BC would be negative. The patterns of positive and negative effects are unique for all combinations of which response is nonzero and whether the response is positive or negative.

Example 9.14 Ramp meters, continued.
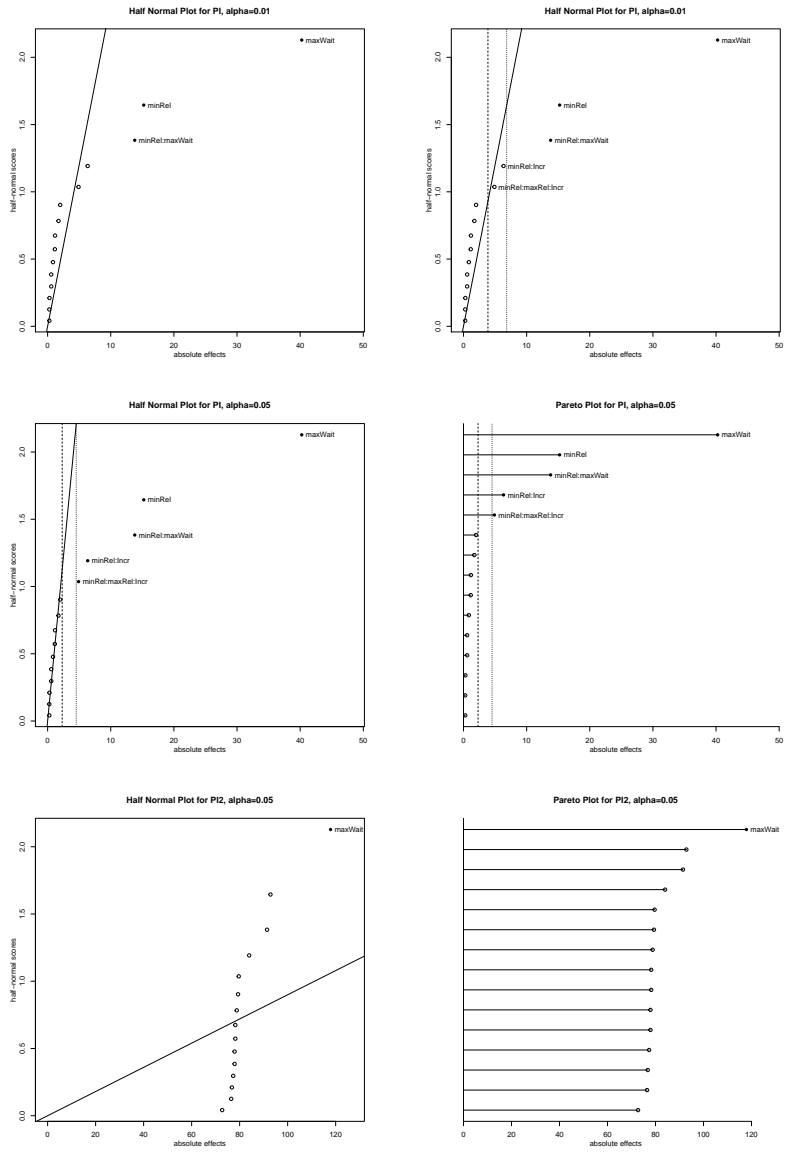
Figure 9.4: Half-normal and Pareto plots for ramp meter data and for the outlier contaminated ramp meter data.

Suppose that the fourth value of the ramp meter data had been recorded as 690 instead of 69.0. Line 6 refits the model using the contaminated response.

```
6 > fit2 <- lm(PI2~minRel*maxWait*maxRel*Incr,data=RampMeters)
7 > TwoSeriesPlots(fit2,pse=TRUE,alpha=.05,type="normal")
8 > TwoSeriesPlots(fit2,pse=TRUE,alpha=.05,type="pareto")
```

Lines 7 and 8 create the half-normal and Pareto plots for the contaminated data, as shown in panels five and six of Figure 9.4. The vertical banding of the effects is clear. Note, however, that this outlier is extreme; the vertical banding is not always this obvious.

9.6   Further Reading and Extensions

A good expository discussion of imbalance can be found in Herr (1986); more advanced treatments can be found in texts on linear models, such as Hocking (1985).

The computational woes of imbalance are less for *proportional balance.* In a two-factor design, we have proportional balance if $n_{ij}/N = n_{i\bullet}/N \times n_{\bullet j}/N$. For example, treatments at level 1 of factor A might have replication 4, and all other treatments have replication 2. Under proportional balance, contrasts in one main effect or interaction are orthogonal to contrasts in any other main effect or interaction. Thus the order in which terms enter a model does not matter, and ordinary, Type II, and Type III sums of squares all agree. Balanced data are obviously a special case of proportional balance. For more than two factors, the rule for proportional balance is that the fraction of the data in one cell should be the product of the fractions in the different margins.

When we have specific hypotheses that we would like to test, but they do not correspond to standard factorial terms, then we must address them with special-purpose contrasts. This is reasonably easy for a single degree of freedom. For hypotheses with several degrees of freedom, we can form multidegree of freedom sums of squares for a set of contrasts using methods described in Hocking (1985) and implemented in many software packages. Alternatively, we may use Bonferroni to combine the tests of individual degrees of freedom.

It is somewhat instructive to see the hypotheses tested by approaches other than Type III. Form row and column averages of treatment means using weights proportional to cell counts:

$$\mu_{i\star} = \sum_{j=1}^{b} n_{ij}\mu_{ij}/n_{i\bullet}$$

$$\mu_{\star j} = \sum_{i=1}^{a} n_{ij}\mu_{ij}/n_{\bullet j} \ ;$$

and form averages for each row of the column weighted averages, and

weighted averages for each column of the row weighted averages:

$$(\mu_{\star j})_{i\star} = \sum_{j=1}^{b} n_{ij}\mu_{\star j}/n_{i\bullet}$$

$$(\mu_{i\star})_{\star j} = \sum_{i=1}^{a} n_{ij}\mu_{i\star}/n_{\bullet j} .$$

Thus there is a $(\mu_{\star j})_{i\star}$ value for each row i, formed by taking a weighted average of the column weighted averages $\mu_{\star j}$. The values may differ between rows because the counts n_{ij} may differ between rows, leading to different weighted averages.

Consider two methods for computing a sum of squares for factor A. We can calculate the sum of squares for factor A ignoring all other factors; this is SAS Type I for factor A first in the model, and is also called "weighted means." This sum of squares is the change in error sum of squares in going from a model with just a grand mean to a model with row effects and is appropriate for testing the null hypothesis

$$\mu_{1\star} = \mu_{2\star} = \cdots = \mu_{a\star} .$$

Alternatively, calculate the sum of squares for factor A adjusted for factor B; this is a Type II sum of squares for a two-way model and is appropriate for testing the null hypothesis

$$\mu_{1\star} = (\mu_{\star j})_{1\star}; \quad \mu_{2\star} = (\mu_{\star j})_{2\star}; \quad \ldots; \quad \mu_{a\star} = (\mu_{\star j})_{a\star} .$$

That is, the Type II null hypothesis for factor A allows the row weighted means to differ, but only because they are different weighted averages of the column weighted means.

Daniel (1976) is an excellent source for the analysis of two-series designs, including unreplicated two-series designs. Much data-analytic wisdom can be found there.

One way of understanding Tukey models is to suppose that we have a simple structure for values $\mu_{ij} = \mu + \alpha_i + \beta_j$. Let's divide through by μ and assume that row and column effects are relatively small compared to the mean. We now have $\mu_{ij} = \mu(1 + \alpha_i/\mu + \beta_j/\mu)$. But instead of working with data on this scale, suppose that we have these data raised to the $1/\lambda$ power.

Then the observed mean structure looks like

$$
\begin{aligned}
(1 + \frac{\alpha_i}{\mu} + \frac{\beta_j}{\mu})^{1/\lambda} &\approx 1 + \frac{\alpha_i}{\mu} + \frac{\beta_j}{\mu} + \frac{1-\lambda}{2\mu^2\lambda^2}(\alpha_i^2 + 2\alpha_i\beta_j + \beta_j^2) \\
&= 1 + \frac{\alpha_i}{\mu} + \frac{1-\lambda}{2\mu^2\lambda^2}\alpha_i^2 + \frac{\beta_j}{\mu} + \frac{1-\lambda}{2\mu^2\lambda^2}\beta_j^2 + \frac{1-\lambda}{\mu^2\lambda^2}\alpha_i\beta_j \\
&\approx 1 + \frac{\alpha_i}{\mu} + \frac{1-\lambda}{2\mu^2\lambda^2}\alpha_i^2 + \frac{\beta_j}{\mu} + \frac{1-\lambda}{2\mu^2\lambda^2}\beta_j^2 + \\
&\quad (1-\lambda)(\frac{\alpha_i}{\mu} + \frac{1-\lambda}{2\mu^2\lambda^2}\alpha_i^2)(\frac{\beta_j}{\mu} + \frac{1-\lambda}{2\mu^2\lambda^2}\beta_j^2) \\
&= (\mu + r_i + c_j + (1-\lambda)\frac{r_i c_j}{\mu})\frac{1}{\mu} \quad,
\end{aligned}
$$

where the first approximation is via a Taylor series and

$$
\begin{aligned}
r_i &= \frac{\alpha_i}{\mu} + \frac{1-\lambda}{2\mu^2\lambda^2}\alpha_i^2 \\
c_j &= \frac{\beta_j}{\mu} + \frac{1-\lambda}{2\mu^2\lambda^2}\beta_j^2 \quad.
\end{aligned}
$$

Thus when we see mean structure of the form $\mu + r_i + c_j + (1-\lambda)r_i c_j/\mu$, we should be able to recover an additive structure by taking the data to the power λ. That is, the transformation power is one minus the coefficient of the cross product term. The cross products $r_i c_j/\mu$ are called the comparison values, because we can compare the residuals from the additive model to these comparison values to see if Tukey style interaction is present.

Here is why our algorithm works for assessing Tukey interaction. We are computing the improvement sum of squares for adding a single degree of freedom term X to a model M. In any ANOVA or regression, the improvement sum of squares obtained by adding the X to M is the same as the sum of squares for the single degree of freedom model consisting of the residuals of X fit to M. For the Tukey interaction procedure in a two-way factorial, the predicted values have the form $\widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j$, so the rescaled squared predicted values equal

$$
\frac{\widehat{\mu}}{2} + (\widehat{\alpha}_i + \frac{\widehat{\alpha}_i^2}{2\widehat{\mu}}) + (\widehat{\beta}_j + \frac{\widehat{\beta}_j^2}{2\widehat{\mu}}) + \frac{\widehat{\alpha}_i\widehat{\beta}_j}{\widehat{\mu}} \quad.
$$

If we fit the additive model to these rescaled squared predicted values, the residuals will be $\widehat{\alpha}_i\widehat{\beta}_j/\widehat{\mu}$. These residuals are exactly the comparison values, so the sum of squares for the squared predicted values entered last will be equal to the sum of squares for the comparison values.

What do we do for comparison values in more complicated models; for example, three factors instead of two? For two factors, the comparison values are the product of the row and column effects divided by the mean. The comparison values for other models are the sums of the cross products of all the terms in the simple model divided by the mean. For example:

Simple Model	Tukey Interaction
$\mu + \alpha_i + \beta_i + \gamma_k$	$\eta(\dfrac{\alpha_i\beta_j}{\mu} + \dfrac{\alpha_i\gamma_k}{\mu} + \dfrac{\beta_i\gamma_k}{\mu})$
$\mu + \alpha_i + \beta_i + \gamma_k + \delta_l$	$\eta(\dfrac{\alpha_i\beta_i}{\mu} + \dfrac{\alpha_i\gamma_k}{\mu} + \dfrac{\alpha_i\delta_i}{\mu} + \dfrac{\beta_i\gamma_k}{\mu} +$ $\dfrac{\beta_i\delta_l}{\mu} + \dfrac{\gamma_k\delta_l}{\mu})$
$\mu + \alpha_i + \beta_i + \alpha\beta_{ij} + \gamma_k$	$\eta(\dfrac{\alpha_i\beta_i}{\mu} + \dfrac{\alpha_i\gamma_k}{\mu} + \dfrac{\alpha_i\alpha\beta_{ij}}{\mu} + \dfrac{\beta_i\gamma_k}{\mu} +$ $\dfrac{\beta_i\alpha\beta_{ij}}{\mu} + \dfrac{\gamma_k\alpha\beta_{ij}}{\mu})$

Once we have the comparison values, we can get their coefficient and the Tukey sum of squares by adding the comparison values to our ANOVA model. In all cases, using the rescaled squared predicted values from the base model accomplishes the same task.

There are several further models of interaction that can be useful, particularly for designs with only one data value per treatment. (See Cook and Weisberg 1982, section 2.5, for a fuller discussion.) Mandel (1961) introduced the *row-model, column-model,* and *slopes-model.* These are generalizations of the Tukey model of interaction, and take the forms

Row-model: $\quad \mu_{ij} = \mu + \alpha_i + \beta_j + \zeta_j\alpha_i$

Column-model: $\quad \mu_{ij} = \mu + \alpha_i + \beta_j + \xi_i\beta_j$

Slopes-model: $\quad \mu_{ij} = \mu + \alpha_i + \beta_j + \zeta_j\alpha_i + \xi_i\beta_j \quad .$

Clearly, the slopes-model is just the union of the row- and column models. These models have the restrictions that

$$\sum_j \zeta_j = \sum_i \xi_i = 0 \quad ,$$

so they represent $b - 1$, $a - 1$, and $a + b - 2$ degrees of freedom respectively in the $(a - 1)(b - 1)$ degree of freedom interaction. The Tukey model is the special case where $\zeta_j = \eta\beta_j$ or $\xi_i = \eta\alpha_i$. It is not difficult to verify that the row- and column models of interaction are orthogonal to the main effects and each other (though not to the Tukey model, which they include, or the slopes-model, which includes both of them).

The interpretation of these models is not too hard. The row model states that mean value of each treatment is a linear function of the row effects, but the slope $(1 + \zeta_j)$ and intercept $(\mu + \beta_j)$ differ from column to column. Similarly, the column model states that the mean value of each treatment is a linear function of the column effects, but the slope $(1 + \xi_i)$ and intercept $(\mu + \alpha_i)$ differ from row to row.

Johnson and Graybill (1972) proposed a model of interaction that does not depend on the main effects:

$$\alpha\beta_{ij} = \delta v_i u_j \ ,$$

with the restrictions that $\sum_i v_i = \sum_j u_j = 0$, and $\sum_i v_i^2 = \sum_j u_j^2 = 1$. This more general structure can model several forms of nonadditivity, including one cell interactions and breakdown of the table into separate additive parts. The components δ, v_i, and u_j are computed from the singular value decomposition of the residuals from the additive model. See Cook and Weisberg for a detailed discussion of this procedure.

9.7   Problems

Three ANOVA tables are given for the results of a single experiment. **Exercise 9.1**
These tables give sequential (Type I) sums of squares. Construct a Type II ANOVA table. What would you conclude about which effects and interactions are needed?

	DF	SS	MS
a	1	1.9242	1.9242
b	2	1584.2	792.1
a.b	2	19.519	9.7595
c	1	1476.7	1476.7
a.c	1	17.527	17.527
b.c	2	191.84	95.92
a.b.c	2	28.567	14.284
Error	11	166.71	15.155

	DF	SS	MS
b	2	1573	786.49
c	1	1428.7	1428.7
b.c	2	153.62	76.809
a	1	39.777	39.777
b.a	2	69.132	34.566
c.a	1	27.51	27.51
b.c.a	2	28.567	14.284
Error	11	166.71	15.155

	DF	SS	MS
c	1	1259.3	1259.3
a	1	9.0198	9.0198
c.a	1	0.93504	0.93504
b	2	1776.1	888.04
c.b	2	169.92	84.961
a.b	2	76.449	38.224

```
c.a.b                2          28.567         14.284
Error               11          166.71         15.155
```

A single replicate of a 2^4 factorial is run. The results in standard order are 1.106, 2.295, 7.074, 6.931, 4.132, 2.148, 10.2, 10.12, 3.337, 1.827, 8.698, 6.255, 3.755, 2.789, 10.99, and 11.85 (data set `TwoSeriesDataA`). Analyze the data to determine the important factors and find which factor-level combination should be used to maximize the response.

Exercise 9.2

Here are two sequential (Type I) ANOVA tables for the same data. Complete the second table. What do you conclude about the significance of row effects, column effects, and interactions?

Exercise 9.3

	DF	SS	MS
r	3	3.3255	1.1085
c	3	112.95	37.65
r.c	9	0.48787	0.054207
ERROR	14	0.8223	0.058736

	DF	SS	MS
c	3	116.25	38.749
r	3		
c.r	9		
ERROR	14		

We have an unbalanced three-way factorial design with factors A, B, and C. I compute both Type II and Type III ANOVAs. Which of the following mean squares will be the same in the two tables (be sure to explain why)?

Exercise 9.4

(a) MS_{AB}.

(b) MS_{ABC}.

(c) MS_E.

An experiment investigated the release of the hormone ACTH from rat pituitary glands under eight treatments: the factorial combinations of CRF (0 or 100 nM; CRF is believed to increase ACTH release), calcium (0 or 2 mM of $CaCl_2$), and Verapamil (0 or 50 μM; Verapamil is thought to block the effect of calcium). Thirty-six rat pituitary cell cultures are assigned at random to the factor-level combinations, with control (all treatments 0) getting 8 units, and other combinations getting 4. The data follow (Giguere, Lefevre, and Labrie 1982, data set `Verapamil`). Analyze these data and report your conclusions.

Problem 9.1

Treatment	ACTH			
Control	1.73	1.57	1.53	2.1
	1.31	1.45	1.55	1.75
V (Verapamil)	2.14	2.24	2.15	1.87
CRF	4.72	2.82	2.76	4.44
CRF + V	4.36	4.05	6.08	4.58
Ca (Calcium)	3.53	3.13	3.47	2.99
Ca + V	3.22	2.89	3.32	3.56
CRF + Ca	13.18	14.26	15.24	11.18
CRF + Ca + V	19.53	16.46	17.89	14.69

Problem 9.2

Consumers who do not regularly eat yogurt are polled and asked to rate on a 1 to 9 scale the likelihood that they would buy a certain yogurt product at least once a month; 1 means very unlikely, 9 means very likely. The product is hypothetical and described by three factors: cost ("C"—low, medium, and high), sensory quality ("S"—low, medium, and high), and nutritional value ("N"—low and high). The plan was to poll three consumers for each product type, but it became clear early in the experiment that people were unlikely to buy a high-cost, low-nutrition, low-quality product, so only one consumer was polled for that combination. Each consumer received one of the eighteen product descriptions chosen at random. The data follow (data set `Yogurt`):

CSN	Scores			CSN	Scores		
HHH	2.6	2.5	2.9	HHL	1.5	1.6	1.5
HMH	2.3	2.1	2.3	HML	1.4	1.5	1.4
HLH	1.05	1.06	1.05	HLL	1.01		
MHH	3.3	3.5	3.3	MHL	2.2	2.0	2.1
MMH	2.6	2.6	2.3	MML	1.8	1.7	1.8
MLH	1.2	1.1	1.2	MLL	1.07	1.08	1.07
LHH	7.9	7.8	7.5	LHL	5.5	5.7	5.7
LMH	4.5	4.6	4.0	LML	3.8	3.3	3.1
LLH	1.7	1.8	1.8	LLL	1.5	1.6	1.5

Analyze these data for the effects of cost, quality, and nutrition on likelihood of purchase.

Problem 9.3

Modern ice creams are not simple recipes. Many use some type of gum to enhance texture, and a non-cream protein source (for example, whey protein solids). A food scientist is trying to determine how types of gum and protein added change a sensory rating of the ice cream. She runs a five by five factorial with two replications using five gum types and five protein sources. Unfortunately, six of the units did not freeze properly, and these units were not rated. Ice creams are rated by a trained panel, with higher ratings being better (data set `IceCream`).

Gum	Protein 1	2	3	4	5
1	3.5	3.6	2.1	4.0	3.1
	3.0	2.9	4.5		
2	7.2	6.8	6.7	7.5	6.8
			4.8	6.9	9.3
3	4.1	5.8	4.5	5.3	4.1
	5.6	4.8	4.6	7.3	5.3
4	5.3	4.8	5.0	6.7	5.2
		3.2	7.2	6.7	4.2
5	4.5	5.1	5.0	4.9	4.5
	2.7	3.7	4.5	4.7	

Analyze these data to determine if protein and/or gum have any effect on the sensory rating. Determine which, if any, proteins and/or gums differ in their sensory ratings.

Here is the output of three different ANOVAs on the same set of (unbalanced) data. **Problem 9.4**

```
WARNING: summaries are sequential
              DF             SS            MS             F        P-value
CONSTANT       1         480.19        480.19      67.98994    8.5625e-13
a              3         194.58         64.86       9.18355    2.1171e-05
b              3         40.143        13.381       1.89460       0.13565
a.b            9         48.572        5.3969       0.76414       0.64955
ERROR1        96         678.01        7.0626

WARNING: summaries are sequential
              DF             SS            MS             F        P-value
CONSTANT       1         480.19        480.19      67.98994    8.5625e-13
b              3         217.67        72.555      10.27307    6.2893e-06
a              3         17.058         5.686       0.80508       0.49406
b.a            9         48.572        5.3969       0.76414       0.64955
ERROR1        96         678.01        7.0626

WARNING: SS are Type III sums of squares
              DF             SS            MS             F        P-value
CONSTANT       1         22.288        22.288       3.15575      0.078828
a              3         19.816        6.6055       0.93527       0.42682
b              3         21.207        7.0691       1.00091       0.39596
a.b            9         48.572        5.3969       0.76414       0.64955
ERROR1        96         678.01        7.0626
```

What do you conclude about the significance of the effects? (You may assume that all assumptions about normality, constant variance, etc are met.)

Gums are used to alter the texture and other properties of foods, in part by binding water. An experiment studied the water-binding of various carrageenan gums in gel systems under various conditions. The experiment had factorial treatment structure with four factors. Factor 1 was the type of gum (kappa, mostly kappa with some lambda, and iota). Factor 2 was the concentration of the gum in the gel in g/100g H_2O (level 1 is .1; level 2 is .5; and level 3 is 2 for gums 1 and 2, and 1 for gum 3). The third factor was type of solute (NaCl, Na_2SO_4, sucrose). The fourth factor was solute concentration (ku/kg H_2O). For sucrose, the three levels were .05, .1, and .25; for NaCl and Na_2SO_4, the levels were .1, .25, and 1. The response is the water-binding for the gel in mOsm (data from Rey 1981, data set `WaterBinding`). This experiment was completely randomized. There were two units at each factor-level combination except solute concentration 3, where all but one combination had four units.

Analyze these data to determine the effects and interactions of the factors. Summarize your analysis and conclusions in a report.

Problem 9.5

S.	S. conc.	G. conc. 1			G. conc. 2			G. conc. 3		
		kappa	k&l	iota	kappa	k&l	iota	kappa	k&l	iota
NaCl	1	99.7	97.6	99.0	100.0	104.7	107.3	123.0	125.7	117.3
		98.3	103.7	98.0	104.3	105.7	106.7	116.3	121.7	117.3
	2	239.0	239.7	237.0	249.7	244.7	243.7	277.0	266.3	268.0
		236.0	246.7	237.7	255.7	245.7	247.7	262.3	276.3	266.7
	3	928.7	940.0	899.3	937.0	942.7	953.3	968.0	992.7	1183.7
		930.0	961.3	941.0	938.7	988.0	991.0	975.7	1019.0	1242.0
		929.0	939.7	944.3	939.7	945.7	988.7	972.7	1018.7	1133.0
		930.0	931.3	919.0	924.3	933.0	965.7	968.0	1021.0	1157.0
Na_2SO_4	1	87.3	80.0	88.0	92.3	94.5	86.7	104.3	115.7	101.0
		89.0	89.3	89.0	97.7	94.3	95.3	104.0	118.0	104.3
	2	203.7	204.0	203.0	209.0	210.7	203.7	218.0	241.0	214.7
		204.0	206.3	201.7	209.3	210.0	209.0	221.5	232.7	222.7
	3	695.0	653.0	668.7	688.7	697.7	726.7	726.0	731.0	747.7
		679.7	642.7	686.7	701.3	701.7	744.7	747.7	790.3	897.0
		692.7	686.0	665.0	698.0	698.0	741.0	736.7	799.7	812.7
		688.0	646.0	688.3	711.7	698.7	708.7	743.7	806.0	885.0
Sucrose	1	55.0	56.7	54.7	61.7	62.7	63.7	90.7	99.0	72.7
		55.3	56.0	56.3	62.0	64.0	65.0	99.3	102.3	75.0
	2	123.7	109.7	105.0	113.3	115.0	114.3	229.3	213.4	123.7
		106.0	111.0	105.7	115.0	115.7	116.7	193.7	196.3	132.7
	3	283.3	271.7	258.3	277.3	279.3	282.0	426.5	399.7	291.7
		276.0	275.3	268.0	277.0	283.0	279.3	389.3	410.3	308.0
		266.0	267.3	273.3	281.3		282.7	420.0	360.0	310.0
		263.0	268.7	272.7	279.0		281.0	421.7	409.3	303.3

Expanded/extruded wheat flours (think breakfast cereals or cheese puffs) **Problem 9.6**
have air cells that vary in size, and the size may depend on the variety of
wheat used to make the flour, the location where the wheat was grown, and
the temperature at which the flour was extruded. An experiment has been
conducted to assess these factors. The first factor is the variety of wheat
used (Butte 86, 2371, or Grandin). The second factor is the growth location
(MN or ND). The third factor is the temperature of the extrusion (120°C or
180°C). The response is the area in mm^2 of the air cells (data from Sutheer-
awattananonda 1994, data set `AirCells`).

Analyze these data and report your conclusions; variety and temperature
effects are of particular interest.

Temp.	Loc.	Var.	Area		
1	1	1	4.63	10.37	7.53
1	1	2	6.83	7.43	2.99
1	1	3	11.02	13.87	2.47
1	2	1	3.44	5.88	
1	2	2	2.60	4.48	
1	2	3	4.29	2.67	
2	1	1	2.80	3.32	
2	1	2	3.01	4.51	
2	1	3	5.30	3.58	
2	2	1	3.12	2.58	2.97
2	2	2	2.15	2.62	3.00
2	2	3	2.24	2.80	3.18

Problem 9.7

Anticonvulsant drugs may be effective because they encourage the effect of the neurotransmitter GABA (γ-aminobutyric acid). Calcium transport may also be involved. The present experiment randomly assigned 48 rats to eight experimental conditions. These eight conditions are the factor-level combinations of three factors, each at two levels. The factors are the anticonvulsant Trifluoperazine (brand name Stelazine) present or absent, the anticonvulsant Diazepam (brand name Valium) present or absent, and the calcium-binding protein calmodulin present or absent. The response is the amount of GABA released when brain tissues are treated with 33 mM K$^+$ (data based on Table I of de Belleroche, Dick, and Wyrley-Birch 1982, data set GABA).

Tri	Dia	Cal	GABA							
A	A	A	1.19	1.33	1.34	1.23	1.24	1.23	1.28	1.32
		P	1.07	1.44	1.14	.87	1.35	1.19	1.17	.89
	P	A	.58	.54	.63	.81				
		P	.61	.60	.51	.88				
P	A	A	.89	.40	.89	.80	.65	.85	.45	.37
		P	1.21	1.20	1.40	.70	1.10	1.09	.90	1.28
	P	A	.19	.34	.61	.30				
		P	.34	.41	.29	.52				

Analyze these data and report your findings. We are interested in whether the drugs affect the GABA release, by how much, and if the calmodulin changes the drug effects.

Problem 9.8

In a study of patient confidentiality, a large number of pediatricians was surveyed. Each pediatrician was given a "fable" about a female patient less than 18 years old. There were sixteen different fables, the combinations of the factors complaint (drug problem or sexually transmitted disease), age (14 years or 17 years), the length of time the pediatrician had known the family (less than 1 year or more than 5 years), and the maturity of patient (immature for age or mature for age). The response at each combination of factor levels is the fraction of doctors who would keep confidentiality and not inform the patient's parents (data modeled on Moses 1987, data set CALM). Analyze these data to determine which factors influence the pediatrician's decision.

C	A	L	M	Response	C	A	L	M	Response
1	1	1	1	.445	2	1	1	1	.578
1	1	1	2	.624	2	1	1	2	.786
1	1	2	1	.360	2	1	2	1	.622
1	1	2	2	.493	2	1	2	2	.755
1	2	1	1	.513	2	2	1	1	.814
1	2	1	2	.693	2	2	1	2	.902
1	2	2	1	.534	2	2	2	1	.869
1	2	2	2	.675	2	2	2	2	.902

Problem 9.9

A consulting client comes to me with an unbalanced, two-factor design (factor A has 4 levels, factor B has 3 levels). He has done type II sums of squares and type III sums of squares; nothing is significant. He also did a one-way anova between the 12 treatments and found significant differences. He is very puzzled and asks what he did wrong in using his statistics software to get these bizarre results. What should I tell him?

Problem 9.10

An animal nutrition experiment was conducted to study the effects of protein in the diet on the level of leucine in the plasma of pigs. Pigs were randomly assigned to one of twelve treatments. These treatments are the combinations of protein source (fish meal, soybean meal, and dried skim milk) and protein concentration in the diet (9, 12, 15, or 18 percent). The response is the free plasma leucine level in mcg/ml (data from Windels 1964, data set `PlasmaLeucine`).

Meal	9%	12%	15%	18%
Fish	27.8	31.5	34.0	30.6
	23.7	28.5	28.7	32.7
	32.8	28.3	33.7	
Soy	39.3	39.8	38.5	42.9
	34.8	40.0	39.2	49.0
	29.8	39.1	40.0	44.4
Milk	40.6	42.9	59.5	72.1
	31.0	50.1	48.9	59.8
	34.6	37.4	41.4	67.6

Analyze these data to determine the effects of the factors on leucine level.

Problem 9.11

Fat acidity is a measure of flour quality that depends on the kind of flour, how the flour has been treated, and how long the flour is stored. In this experiment there are two types of flour (Patent or First Clear); the flour treatment factor (extraction) has eleven levels, and the flour has been stored for one of six periods (0, 3, 6, 9, 15, or 21 weeks). We observe only one unit for each factor-level combination. The response is fat acidity in mg KOH/100 g flour (data from Nelson 1961, data set `FatAcidity`). Analyze these data. Of particular interest are the effect of storage time and how that might depend on the other factors.

T	W	Extraction										
		1	2	3	4	5	6	7	8	9	10	11
P	0	12.7	12.3	15.4	13.3	13.9	30.3	123.9	53.4	29.4	11.4	19.0
	3	11.3	16.4	18.1	14.6	10.5	27.5	112.3	48.9	31.4	11.6	29.1
	6	16.5	24.3	27.2	10.9	11.6	34.1	117.5	52.9	38.3	15.8	17.1
	9	10.9	30.8	24.5	13.5	13.2	33.2	107.4	49.6	42.9	17.8	15.9
	15	12.5	30.6	26.5	15.8	13.3	36.2	109.5	51.0	15.2	18.2	13.5
	21	15.2	36.3	36.8	14.4	13.1	43.2	98.6	48.2	58.6	22.2	17.6
FC	0	36.5	38.5	38.4	27.1	35.0	38.3	274.6	241.4	21.8	34.2	34.2
	3	35.4	68.5	63.6	41.4	34.5	76.8	282.8	231.8	47.9	33.9	33.2
	6	35.7	93.2	76.7	50.2	34.0	96.4	270.8	223.2	65.2	38.9	35.2
	9	33.8	95.0	113.0	44.9	36.1	94.5	271.6	200.1	75.0	39.0	34.7
	15	43.0	156.7	160.0	30.2	33.0	75.8	269.5	213.6	88.9	37.9	33.0
	21	53.0	189.3	199.3	41.0	45.5	143.9	136.1	198.9	104.0	39.2	37.1

Problem 9.12 Artificial insemination is an important tool in agriculture, but freezing semen for later use can reduce its potency (ability to produce offspring). Here we are trying to understand the effect of freezing on the potency of chicken semen. Four semen mixtures are prepared, consisting of equal parts of either fresh or frozen Rhode Island Red semen, and either fresh or frozen White Leghorn semen. Sixteen batches of Rhode Island Red hens are assigned at random, four to each of the four treatments. Each batch of hens is inseminated with the appropriate mixture, and the response measured is the fraction of the hatching eggs that have white feathers and thus White Leghorn fathers (data from Tajima 1987, data set `ChickenSemen`). Analyze these data to determine how freezing affects potency of chicken semen.

RIR	WL	Fraction			
Fresh	Fresh	.435	.625	.643	.615
Frozen	Frozen	.500	.600	.750	.750
Fresh	Frozen	.250	.267	.188	.200
Frozen	Fresh	.867	.850	.846	.950

Problem 9.13 Explore the interaction in the pacemaker delamination data introduced in Problem 8.3.

Problem 9.14 Explore the interaction in the tropical grass production data introduced in Problem 8.5.

Problem 9.15 Explore the interaction in the dye adsorption data introduced in Problem 8.12.

Problem 9.16 Explore the interaction in the dye removal data introduced in Problem 8.14.

Problem 9.17 Explore the interaction in the reaction time data introduced in Problem 8.13.

Problem 9.18 One measure of the effectiveness of cancer drugs is their ability to reduce the number of viable cancer cells in laboratory settings. In this experiment, the A549 line of malignant cells is plated onto petri dishes with various concentrations of the drug cisplatin. After 7 days of incubation, half the petri

dishes at each dose are treated with a dye, and the number of viable cell colonies per 500 mm^2 is determined as a response for all petri dishes (after Figure 1 of Alley, Uhl, and Lieber 1982, data set `Cisplatin`). The dye is supposed to make the counting machinery more specific to the cancer cells.

	Cisplatin (ng/ml)					
	0	.15	1.5	15	150	1500
Conventional	200	178	158	132	63	40
Dye added	56	50	45	63	18	14

Analyze these data for the effects of concentration and dye. What can you say about interaction?

We have a 2^4 factorial with a single replication. Looking at the Daniel plot we see A and B are significant; C and D are probably significant; and maybe AB (it's the next one) is significant.

Problem 9.19

Half Normal Plot for y, alpha=0.05
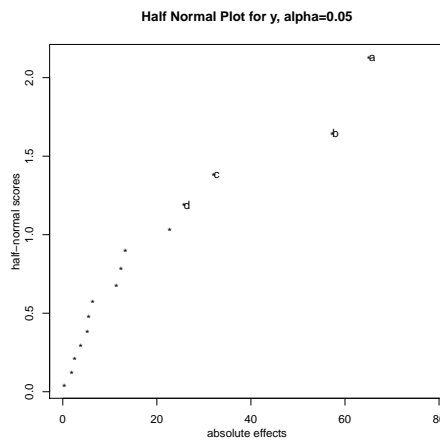
If we fit a model with just A, B, C, and D we get a residual plot that looks like this. (It's not much different if you add AB.) What do we do now?

Residuals vs Fitted
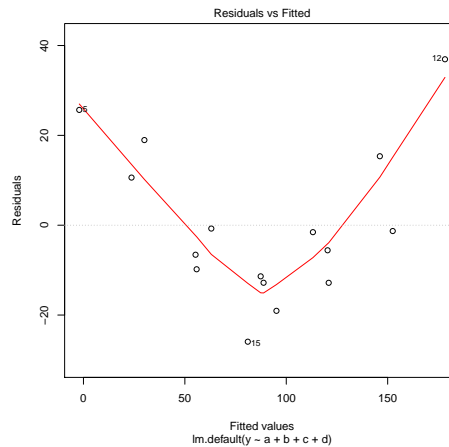
An experiment studied the effects of starch source, starch concentration, and temperature on the strength of gels. This experiment was completely randomized with sixteen units. There are four starch sources (adzuki bean, corn, wheat, and potato), two starch percentages (5% and 7%), and two temperatures (22°C and 4°C). The response is gel strength in grams (data from Tjahjadi 1983, data set `GelStrength`).

Problem 9.20

Temperature	Percent	Bean	Corn	Wheat	Potato
22	5	62.9	44.0	43.8	34.4
	7	110.3	115.6	123.4	53.6
4	5	60.1	57.9	58.2	63.0
	7	147.6	180.7	163.8	92.0

Analyze these data to determine the effects of the factors on gel strength.

We have a 2^4 experiment on the mass of a nanoscale polymer. We produce 16 polymer layers by varying the factors A — concentration of solution; B — number of coats; C — speed of coating machine; D — flow rate of solution. There are four plots below. The plots in the first row are for data on the original scale, while the plots in the second row are plots on the log scale. For each scale we show a Daniel Plot (with α set to .3 so lots of terms are labelled) and a residuals versus predicted plot for the main effects only model.

Problem 9.21
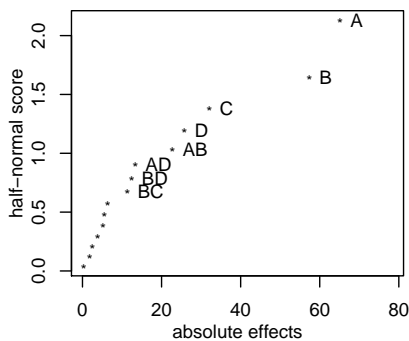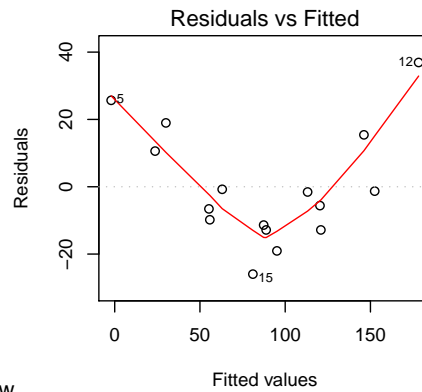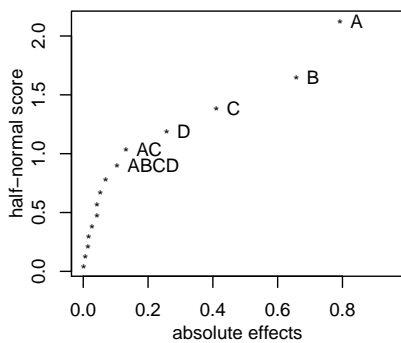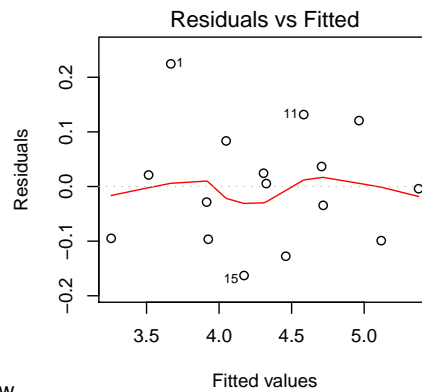
Explain which scale you would choose and why.

Heavy metals should be removed from the waste water stream before it

Problem 9.22

is released into the environment. This experiment examines how five factors affect the removal of zinc and copper from water via liquid-liquid extraction. There are five factors under consideration: pH of initial solution (4.5 or 6.5), initial concentration of the metal (25 or 75 mg/l), concentration of the extractant (5 or 10 % by volume), medium of the solution (sulfate or chloride), and stirring rate (400 or 500 rpm). The 32 factor/level combinations were assigned to 32 independent runs, and two responses were measured: percent removal of zinc and percent removal of copper.

The data for the experiment are in the following two tables (data from Berrama, Benaouag, Kaouah, and Bendjama 2013 via Lye 2019, data set `CopperZinc`).

Zn Pct. removed		Extractant/Initial Conc./pH							
		5				10			
		25		75		25		75	
Stirring	Medium	4.5	6.6	4.5	6.6	4.5	6.6	4.5	6.6
400	Sulfate	90.1	93.4	83.7	85.9	96.7	97.0	93.7	98.3
	Chloride	95.9	97.2	95.0	96.6	99.3	99.0	91.9	97.3
500	Sulfate	98.0	94.8	88.0	94.5	97.0	97.0	89.5	93.4
	Chloride	94.4	95.4	89.3	88.1	98.0	96.8	85.8	90.1

Cu Pct. removed		Extractant/Initial Conc./pH							
		5				10			
		25		75		25		75	
Stirring	Medium	4.5	6.6	4.5	6.6	4.5	6.6	4.5	6.6
400	Sulphate	96.5	96.7	96.1	96.8	98.4	98.8	98.5	98.8
	Chloride	95.8	95.9	95.3	99.5	98.4	98.6	98.1	98.2
500	Sulphate	96.1	96.7	96.0	97.4	94.3	98.2	98.5	98.9
	Chloride	96.2	96.7	95.7	98.4	98.5	98.5	98.0	98.2

Analyze these data to determine the effects of the factors on removal of metals from solution.

Problem 9.23 Burning coal containing sulfur produces acid rain, so this experiment studies a process to remove sulfur from coal. There are five factors, each at two levels: pH (1.5 or 2.5), particle size (180 or 500 um), Fe^{2+} (0 or 60 mmol), pulp density (2 or 10%), and leaching time (6 or 14 days). Thirty-two units are assigned at random to the factor/level combinations. The response of interest is percentage of sulfur removed, as shown in the following table (data from Golshani, Jorjani, Chelgani S. Chehreh, and Heidari 2013 via Lye 2019, data set `SulfurRemoval`).

		Iron/Size/pH							
		0				60			
		180		500		180		500	
Time	Density	1.5	2.5	1.5	2.5	1.5	2.5	1.5	2.5
6	2	29.76	25.56	28.87	24.65	29.75	24.35	26.24	22.47
	10	34.52	29.85	35.88	25.36	34.26	34.62	31.35	24.08
14	2	42.31	38.03	40.76	35.31	42.88	39.27	39.92	37.53
	10	48.47	38.50	43.78	37.58	53.12	51.97	48.82	46.82

Analyze these data to determine the effects of the factors on removal of metals from solution.

Problem 9.24

Wax from crude oil may settle out and literally gum up the works. This experiment considers how four factors affect the deposition of wax from Malaysian crude oil. The factors are speed of rotation (0 or 600 rpm), cold finger temperature (5 or 15 C), duration (2 or 24 h), and inhibitor concentration (200 or 5000 ppm). The 16 factor/level combinations are run in random order, and the response is the amount of wax deposited (in g). Data from Ridzuan, Adam, and Yaacob (2016) via Lye (2019), data set `WaxDeposit`.

		Duration/Temp./Rotation							
		2				24			
		5		15		5		15	
Inhibitor	0	600	0	600	0	600	0	600
200	1.9	1.8	0.80	1.0	2.65	2.8	1.40	1.6
5000	1.5	2.1	0.75	0.9	2.50	3.0	1.05	1.2

Analyze these data to determine the effects of the factors.

Problem 9.25

Ginger contains an essential oil that we would like to extract without using a solvent, in this case by using a microwave. This experiment considers eight different treatments for extracting this oil. These treatments are the factor/level combinations of duration (10 or 30 minutes), wattage (288 or 640), and preparation (crushed or sliced). The response is the oil yield (in %). Sixteen samples of ginger are randomly assigned to the eight factor/level combinations, two per combination. The data are in the table below (data from Shah and Garg 2014 via Lye 2019, data set `GingerOil`).

| Type | Sliced | | | | Crushed | | | |
| Wattage | 288 | | 640 | | 288 | | 640 | |
Duration	10	30	10	30	10	30	10	30
	0.10	0.26	0.14	0.35	0.22	0.20	0.28	0.44
	0.12	0.25	0.14	0.32	0.24	0.17	0.31	0.46

Analyze these data paying particular attention to the interactions that are present.

Question 9.1

A study uses computational fluid dynamics (CFD) to calculate the diffusion time for moisture. There are four factors, each at two levels (length of the opening, radius of the opening, temperature, and relative humidity). The

physics are calculated for the 16 different situations and the diffusion time (response) computed. The responses are wildly different, but only length and radius appear to be significant when analyzing the log diffusion times.

This is a "computer experiment," which means that if we redo one of the sixteen factor/level runs we will get *exactly* the same response. Put another way, σ^2 is apparently zero.

What are the inferential consequences of an experiment with apparently no error?

Show how to construct simultaneous confidence intervals for all pairwise differences of interaction effects $\widehat{\alpha\beta}_{ij}$ using Bonferroni. Hint: first find the variances of the differences.

Question 9.2

Determine the condition for orthogonality of two main-effects contrasts for the same factor when the data are unbalanced.

Question 9.3

Show that an interaction contrast w_{ij} in the means $\overline{y}_{ij\bullet\bullet}$ equals the corresponding contrast in the interaction effects $\widehat{\alpha\beta}_{ij}$.

Question 9.4

Chapter 10

Random and Mixed Effects Models

Random effects are another approach to designing experiments and modeling data. Random effects are appropriate when the treatments are random samples from a population of potential treatments. They are also useful for random subsampling from populations, even if we did not apply a treatment *per se*. Random-effects models make the same kinds of decompositions into overall mean, treatment effects, and random error that we have been using, but random-effects models assume that the treatment effects are random variables. Also, the focus of inference is often on the variance of the population of potential treatment effects, not the individual treatment effects themselves. This chapter introduces random-effects models along with mixed effects, nesting, and the Hasse diagram to visualize the model.

10.1   Models for Random Effects

A company has 50 machines that make cardboard cartons for canned goods, and they want to understand the variation in strength of the cartons. They choose ten machines at random from the 50 and make 40 cartons on each machine, assigning 400 lots of feedstock cardboard at random to the ten chosen machines. The resulting cartons are tested for strength. This is a completely randomized design, with ten treatments and 400 units; we will refer to this as carton experiment one.

We have been using models for data that take the form

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij} .$$

The parameters of the mean structure (μ_i, μ, and α_i) have been treated as fixed, unknown numbers with the treatment effects summing to zero, and the primary thrust of our inference has been learning about these mean parameters. These sorts of models are called *fixed-effects* models, because the

treatment effects are fixed numbers.

These fixed-effects models are not appropriate for our carton strength data. It still makes sense to decompose the data into an overall mean, treatment effects, and random error, but the fixed-effects assumptions don't make much sense here for a couple of reasons. First, we are trying to learn about and make inferences about the whole population of machines, not just these ten machines that we tested in the experiment, so we need to be able to make statements for the whole population, not just the random sample that we used in the experiment. Second, we can learn all we want about these ten machines, but a replication of the experiment will give us an entirely different set of machines. Learning about α_1 in the first experiment tells us nothing about α_1 in the second experiment—they are probably different machines. We need a new kind of model.

Random-effects designs study populations of treatments

The basic random effects model begins with the usual decomposition:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \ .$$

We assume that the errors ϵ_{ij} are independent and normally distributed with mean 0 and variance σ^2, as we did in fixed effects. For random effects, we also assume that the treatment effects α_i are independent and normally distributed with mean 0 and variance σ_α^2, and that the α_i's and the ϵ_{ij}'s are independent of each other. Random effects models do not require that the sum of the α_i's be zero.

Treatment effects are random in random-effects models

The variance of y_{ij} is $\sigma_\alpha^2 + \sigma^2$. The terms σ_α^2 and σ^2 are called *components of variance* or *variance components*. Thus the random-effects model is sometimes called a components of variance model. The correlation between y_{ij} and y_{kl} is

Variance components

$$\mathrm{Cor}(y_{ij}, y_{kl}) = \left\{ \begin{array}{lll} 0 & & i \neq k \\ \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma^2) & \text{for} & i = k \text{ and } j \neq l \\ 1 & & i = k \text{ and } j = l \end{array} \right. \ .$$

The correlation is nonzero when $i = k$ because the two responses share a common value of the random variable α_i. The correlation between two responses in the same treatment group is called the *intraclass* correlation. Another way of thinking about responses in a random-effects model is that they all have mean μ, variance $\sigma_\alpha^2 + \sigma^2$, and a correlation structure determined by the variance components. The additive random-effects model and the correlation structure approach are nearly equivalent (the additive random-effects model can only induce positive correlations, but the general correlation structure model allows negative correlations).

Intraclass correlation

Random effects can be specified by correlation structure

The parameters of the random effects model are the overall mean μ, the error variance σ^2, and the variance of the treatment effects σ_α^2; the treatment effects α_i are random variables, not parameters. We want to make inferences about these parameters; we are often not as interested in making inferences about the α_i's and ϵ_{ij}'s, which will be different in the next experiment anyway. Typical inferences would be point estimates or confidence intervals for

Tests and confidence intervals for parameters

the variance components, or a test of the null hypothesis that the treatment variance σ_α^2 is 0.

Now extend carton experiment one. Suppose that machine operators may also influence the strength of the cartons. In addition to the ten machines chosen at random, the manufacturer also chooses ten operators at random. Each operator will produce four cartons on each machine, with the cardboard feedstock assigned at random to the machine-operator combinations. We now have a two-way factorial treatment structure with both factors random effects and completely randomized assignment of treatments to units. This is carton experiment two.

<div style="float:right; text-align:right">Carton experiment two, two random factors</div>

The model for two-way random effects is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \,,$$

where α_i is a main effect for factor A, β_j is a main effect for factor B, $\alpha\beta_{ij}$ is an AB interaction, and ϵ_{ijk} is random error. The model assumptions are that all the random effects α_i, β_j, $\alpha\beta_{ij}$, and ϵ_{ijk} are independent, normally distributed, with mean 0. Each effect has its own variance: $\text{Var}(\alpha_i) = \sigma_\alpha^2$, $\text{Var}(\beta_j) = \sigma_\beta^2$, $\text{Var}(\alpha\beta_{ij}) = \sigma_{\alpha\beta}^2$, and $\text{Var}(\epsilon_{ijk}) = \sigma^2$. The variance of y_{ijk} is $\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2$, and the correlation of two responses is the sum of the variances of the random components that they share, divided by their common variance $\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma^2$.

<div style="float:right; text-align:right">Two-factor model</div>

This brings us to another way that random effects differ from fixed effects. In fixed effects, we have a table of means onto which we impose a structure of equally weighted main effects and interactions. There are other plausible structures based on unequal weightings that can have different main effects and interactions, so testing main effects when interactions are present in fixed effects makes sense only when we are truly interested in the specific, equally-weighted null hypothesis corresponding to the main effect. Random effects set up a correlation structure among the responses, with autonomous contributions from the different variance components. It is reasonable to ask if a main-effect contribution to correlation is absent even if interaction contribution to correlation is present. Similarly, equal weighting is about the only weighting that makes sense in random effects; after all, the row effects and column effects are chosen randomly and exchangeably. Why weight one row or column more than any other? So for random effects, we more or less automatically test for main effects, even if interactions are present.

<div style="float:right; text-align:right">Hierarchy less important in random-effects models</div>

We can, of course, have random effects models with more than two factors. Suppose that there are many batches of glue, and we choose two of them at random. Now each operator makes two cartons on each machine with each batch of glue. We now have 200 factor-level combinations assigned at random to the 400 units. This is carton experiment three.

<div style="float:right; text-align:right">Carton experiment three, three random factors</div>

The model for three-way random effects is

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon_{ijkl} \,,$$

where α_i, β_j, and γ_k are main effects; $\alpha\beta_{ij}$, $\alpha\gamma_{ik}$, $\beta\gamma_{ik}$, and $\alpha\beta\gamma_{ijk}$ are

<div style="float:right; text-align:right">Three-factor model</div>

interactions; and ϵ_{ijkl} is random error. The model assumptions remain that all the random effects are independent and normally distributed with mean 0. Each effect has its own variance: $\text{Var}(\alpha_i) = \sigma_\alpha^2$, $\text{Var}(\beta_j) = \sigma_\beta^2$, $\text{Var}(\gamma_k) = \sigma_\gamma^2$, $\text{Var}(\alpha\beta_{ij}) = \sigma_{\alpha\beta}^2$, $\text{Var}(\alpha\gamma_{ik}) = \sigma_{\alpha\gamma}^2$, $\text{Var}(\beta\gamma_{jk}) = \sigma_{\beta\gamma}^2$, $\text{Var}(\alpha\beta\gamma_{ijk}) = \sigma_{\alpha\beta\gamma}^2$, and $\text{Var}(\epsilon_{ijkl}) = \sigma^2$. Generalization to more factors is straightforward.

10.2 Why Use Random Effects?

The carton experiments described above are all completely randomized designs: the units are assigned at random to the treatments. The difference from what we have seen before is that the treatments have been randomly sampled from a population. Why should anyone design an experiment that uses randomly chosen treatments?

The answer is that we are trying to draw inferences about the population from which the treatments were sampled. Specifically, we are trying to learn about variation in the treatment effects. Thus we want to design an experiment that looks at variation in a population by looking at the variability that arises when we sample from the population. When you want to study variances and variability, think random effects. *Random effects study variances in populations*

Random-effects models are also used in subsampling situations. Revise carton experiment one. The manufacturer still chooses ten machines at random, but instead of making new cartons, she simply goes to the warehouse and collects 40 cartons at random from those made by each machine. It still makes sense to model the carton strengths with a random effect for the randomly chosen machine and a random error for the randomly chosen cartons from each machine's stock; that is precisely the random effects model. *Use random effects when subsampling*

In the subsampling version of the carton example, we have done no experimentation in the sense of applying randomly assigned treatments to units. Instead, the stochastic nature of the data arises because we have sampled from a population. The items we have sampled are not exactly alike, so the responses differ. Furthermore, the sampling was done in a structured way (in the example, first choose machines, then cartons for each machine) that produces some correlation between the responses. For example, we expect cartons from the same machine to be a bit similar, but cartons from different machines should be unrelated. The pattern of correlation for subsampling is the same as the pattern of correlation for randomly chosen treatments applied to units, so we can use the same models for both. *Subsampling induces random variation*

10.3 Nesting Versus Crossing

The vitamin A content of baby food carrots may not be consistent. To evaluate this possibility, we go to the grocery store and select four jars of carrots at random from each of the three brands of baby food that are sold in our

region. We then take two samples from each jar and measure the vitamin A in every sample for a total of 24 responses.

It makes sense to consider decomposing the variation in the 24 responses into various sources. There is variation between the brands, variation between individual jars for each brand, and variation between samples for every jar.

<div style="text-align: right">*Multiple sources of variation*</div>

It does *not* make sense to consider jar main effects and brand by jar interaction. Jar one for brand A has absolutely nothing to do with jar one for brand B. They might both have lots of vitamin A by chance, but it would just be chance. They are not linked, so there should be no jar main effect across the brands. If the main effect of jar doesn't make sense, then neither does a jar by brand interaction, because that two-factor interaction can be interpreted as how the main effect of jar must be altered at each level of brand to obtain treatment means.

<div style="text-align: right">*No jar effect across brands*</div>

Main effects and interaction are appropriate when the treatment factors are *crossed*. Two factors are crossed when treatments are formed as the combinations of levels of the two factors, and we use the same levels of the first factor for every level of the second factor, and vice versa. All factors we have considered until the baby carrots have been crossed factors. The jar and brand factors are not crossed, because we have different jars (levels of the jar factor) for every brand.

<div style="text-align: right">*Crossed factors form treatments with their combinations*</div>

The alternative to crossed factors is *nested* factors. Factor B is nested in factor A if there is a completely different set of levels of B for every level of A. Thus the jars are nested in the brands and not crossed with the brands, because we have a completely new set of jars for every brand. We write nested models using parentheses in the subscripts to indicate the nesting. If brand is factor A and jar (nested in brand) is factor B, then the model is written

<div style="text-align: right">*Factor B nested in A has different levels for every level of A*</div>

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{k(ij)} \ .$$

The $j(i)$ indicates that the factor corresponding to j (factor B) is nested in the factor corresponding to i (factor A). Thus there is a different β_j for each level i of A. In terms of term labels, this is sometimes written as B(A) to indicate that factor B is nested in factor A.

Note that we wrote $\epsilon_{k(ij)}$, nesting the random errors in the brand-jar combinations. This means that we get a different, unrelated set of random errors for each brand-jar combination. In the crossed factorials we have used until now, the random error is nested in the all-way interaction, so that for a three-way factorial the error ϵ_{ijkl} could more properly have been written $\epsilon_{l(ijk)}$. Random errors are always nested in some model term; we've just not needed to deal with it before now.

<div style="text-align: right">*Errors are nested*</div>

Nested factors can be random or fixed, though they are usually random and often arise from some kind of subsampling. As an example of a factor that is fixed and nested, consider a company with work crews, each crew consisting of four members. Members are nested in crews, and we get the same four crew members whenever we look at a given crew, making member a fixed effect.

<div style="text-align: right">*Nested factors are usually random*</div>

When we have a chain of factors, each nested in its predecessor, we say that the design is fully nested. The baby carrots example is fully nested, with jars nested in brand, and sample nested in jar. Another example comes from genetics. There are three subspecies. We randomly choose five males from each subspecies (a total of fifteen males); each male is mated with four females (of the same subspecies, a total of 60 females); we observe three offspring per mating (a total of 180 offspring); and we make two measurements on each offspring (a total of 360 measurements). Offspring are nested in females, which are nested in males, which are nested in subspecies.

<div align="right">Fully nested design</div>

10.4   Why Nesting?

We may design an experiment with nested treatment structure for several reasons. Subsampling produces small units by one or more layers of selection from larger bundles of units. For the baby carrots we went from brands to jars to samples, with each layer being a group of units from the layer beneath it. Subsampling can be used to select treatments as well as units. In some experiments crossing is theoretically possible, but logistically impractical. There may be two or three clinics scattered around the country that can perform a new diagnostic technique. We could in principle send our patients to all three clinics to cross clinics with patients, but it is more realistic to send each patient to just one clinic. In other experiments, crossing simply cannot be done. For example, consider a genetics experiment with females nested in males. We need to be able to identify the father of the offspring, so we can only breed each female to one male at a time. However, if females of the species under study only live through one breeding, we must have different females for every male.

<div align="right">Unit generation, logistics, and constraints may lead to nesting</div>

We do not simply choose to use a nested model for an experiment. We use a nested model because the treatment structure of the experiment was nested, and we must build our models to match our treatment structure.

<div align="right">Models must match designs</div>

10.5   Crossed and Nested Factors

Designs can have both crossed and nested factors. One common source of this situation is that "units" are produced in some sense through a nesting structure. In addition to the nesting structure, there are treatment factors, the combinations of which are assigned at random to the units in such a way that all the combinations of nesting factors and treatment factors get an equal number of units.

<div align="right">Units with nesting crossed with treatments</div>

Example 10.1 Gum arabic

Gum arabic is used to lengthen the shelf life of emulsions, including soft drinks, and we wish to see how different gums and gum preparations affect emulsion shelf life. Raw gums are ground, dissolved, treated (possible

treatments include pasteurization, demineralization, and acidification), and then dried; the resulting dry powder is used as an emulsifier in food products.

Gum arabic comes from acacia trees; we obtain four raw gum samples from each of two varieties of acacia tree (a total of eight samples). Each sample is split into two subsamples. One of the subsamples (chosen at random) will be demineralized during treatment, the other will not. The sixteen subsamples are now dried, and we make five emulsions from each subsample and measure as the response the time until the ingredients in the emulsion begin to separate.

This design includes both crossed and nested factors. The samples of raw gum are nested in variety of acacia tree; we have completely different samples for each variety. The subsamples are nested in the samples. Subsample is now a unit to which we apply one of the two levels of the demineralization factor. Because one subsample from each sample will be demineralized and the other won't be, each sample occurs with both levels of the demineralization treatment factor. Thus sample and treatment factor are crossed. Similarly, each variety of acacia occurs with both levels of demineralization so that variety and treatment factor are crossed. The five individual emulsions from a single subsample are nested in that subsample, or equivalently, in the variety-sample-treatment combinations. They are measurement units.

If we let variety, sample, and demineralization be factors A, B, and C, then an appropriate model for the responses is

$$y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk(i)} + \epsilon_{l(ijk)} \ .$$

Not all designs with crossed and nested factors have such a clear idea of unit. For some designs, we can identify the sources of variation among responses as factors crossed or nested, but identifying "treatments" randomly assigned to "units" takes some mental gymnastics.

Treatments and units not always clear

Example 10.2 Cheese tasting

Food scientists wish to study how urban and rural consumers rate cheddar cheeses for bitterness. Four 50-pound blocks of cheddar cheese of different types are obtained. Each block of cheese represents one of the segments of the market (for example, a sharp New York style cheese). The raters are students from a large introductory food science class. Ten students from rural backgrounds and ten students from urban backgrounds are selected at random from the pool of possible raters. Each rater will taste eight bites of cheese presented in random order. The eight bites are two each from the four different cheeses, but the raters don't know that. Each rater rates each bite for bitterness.

The factors in this experiment are background, rater, and type of cheese. The raters are nested in the backgrounds, but both background and rater are crossed with cheese type, because all background-cheese type combinations and all rater/cheese type combinations occur. This is an experiment with both

crossed and nested factors. Perhaps the most sensible formulation of this as treatments and units is to say that bites of cheese are units (nested in type of cheese) and that raters nested in background are treatments applied to bites of cheese.

If we let background, rater, and type be factors A, B, and C, then an appropriate model for the responses is

$$y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk(i)} + \epsilon_{l(ijk)} \ \ .$$

This is the same model as Example 10.1, even though the structure of units and treatments is very different!

These two examples illustrate some of the issues of working with designs having both crossed and nested factors. You need to

1. Determine the sources of variation,

2. Decide which cross and which nest,

3. Decide which factors are fixed and which are random, and

4. Decide which interactions should be in the model.

Identifying the appropriate model with fixed-random-crossed-nested designs can be a challenge; it takes a lot of practice.

10.6   Mixed Effects

In addition to having both crossed and nested factors, Example 10.1 has both fixed (variety and demineralization) and random (sample) factors; Example 10.2 also has fixed (background and cheese type) and random (rater) factors. An experiment with both fixed and random effects is said to have *mixed* effects. The interaction of a fixed effect and a random effect must be random, because a new random sample of factor levels will also lead to a new sample of interactions.

Analysis of mixed-effects models reminds me of the joke in the computer business about standards: "The wonderful thing about standards is that there are so many to choose from." For mixed effects, there are two sets of assumptions that have a reasonable claim to being standard. Unfortunately, the two sets of assumptions lead to different analyses, and potentially different answers.

Before stating the mathematical assumptions, let's visualize two mechanisms for producing the data in a mixed-effects model; each mechanism leads to a different set of assumptions. By thinking about the mechanisms behind the assumptions, we should be able to choose the appropriate assumptions in any particular experiment. Let's consider a two-factor model, with factor A fixed and factor B random, and a very small error variance so that the data are really just the sums of the row, column, and interaction effects.

Here is one way to get the data. Imagine a table with a rows and a very large number of columns. Our random factor B corresponds to selecting b of

the columns from the table at random, and the data we observe are the items in the table for the columns that we select.

This construction implies that if we repeated the experiment and we happened to get the same column twice, then the column totals of the data for the repeated column would be the same in the two experiments. Put another way, once we know the column we choose, we know the total for that column; we don't need to wait and see what particular interaction effects are chosen before we see the column total. Thus column differences are determined by the main effects of column; we can assume that the interaction effects in a given column add to zero. This approach leads to the *restricted* model, since it restricts the interaction effects to add to zero when summed across a fixed effect.

The second approach treats the main effects and interactions independently. Now we have two populations of effects; one population contains random column main effects β_j, and the other population contains random interaction effects $\alpha\beta_{ij}$. In this second approach, we have fixed row effects, we choose column effects randomly and independently from the column main effects population, and we choose interaction effects randomly and independently from the interaction effects population; the column and interaction effects are also independent.

When we look at column totals in these data, the column total of the interaction effects can change the column total of the data. Another sample with the same column will have a different column total, because we will have a different set of interaction effects. This second approach leads to the *unrestricted* model, because it has no zero-sum restrictions.

Choose between these models by answering the following question: if you reran the experiment and got a column twice, would you have the same interaction effects or an independent set of interaction effects for that repeated column? If you have the same set of interaction effects, use the restricted model. If you have new interaction effects, use the unrestricted model. I tend to use the restricted model by default and switch to the unrestricted model when appropriate.

We will see that the unrestricted model is more conservative in the sense that it implies that random variability for estimated effects is at least as large, and possibly larger, than the corresponding model using restricted assumptions.

Example 10.3 Cheese tasting, continued

In the cheese tasting example, one of our raters is Mary; Mary likes sharp cheddar cheese and dislikes mild cheese. Any time we happen to get Mary in our sample, she will rate the sharp cheese higher and the mild cheese lower. John, on the other hand, likes milder cheeses. We get the same rater by cheese interaction effects every time we choose Mary or John, so the restricted model is appropriate.

Example 10.4 Particle sampling

To monitor air pollution, a fixed volume of air is drawn through disk-shaped filters, and particulates deposit on the filters. Unfortunately, the particulate deposition is not uniform across the filter. Cadmium particulates on a filter are measured by X-ray fluorescence. The filter is placed in an instrument that chooses a random location on the filter, irradiates that location twice, measures the resulting fluorescence spectra, and converts them to cadmium concentrations. We compare three instruments by choosing ten filters at random and running each filter through all three instruments, for a total of 60 cadmium measurements.

In this experiment we believe that the primary interaction between filter and instrument arises because of the randomly chosen locations on that filter that are scanned and the nonuniformity of the particulate on the filter. Each time the filter is run through an instrument, we get a different location and thus a different "interaction" effect, so the unrestricted model is appropriate.

Unfortunately, the choice between restricted and unrestricted models is not always clear.

Example 10.5 Gum arabic, continued

Gum sample is random (nested in variety) and crosses with the fixed demineralization factor. Should we use the restricted or unrestricted model? If a gum sample is fairly heterogeneous, then at least some of any interaction that we observe is probably due to the random split of the sample into two subsamples. The next time we do the experiment, we will get different subsamples and probably different responses. In this case, the demineralization by sample interaction should be treated as unrestricted, because we would get a new set of effects every time we redid a sample.

On the other hand, how a sample reacts to demineralization may be a shared property of the complete sample. In this case, we would get the same interaction effects each time we redid a sample, so the restricted model would be appropriate.

We need to know more about the gum samples before we can make a reasoned decision on the appropriate model.

Here are the technical assumptions for mixed effects. For the unrestricted model, all random effects are independent and have normal distributions with mean 0. Random effects corresponding to the same term have the same variance: σ_β^2, $\sigma_{\alpha\beta}^2$, and so on. Any purely fixed effect or interaction must add to zero across any subscript.

Unrestricted model assumptions

The assumptions for the restricted model are the same, except for interactions that include both fixed and random factors. Random effects in a mixed-interaction term have the same variance, which is written as a factor times the usual variance component: for example, $r_{ab}\,\sigma_{\alpha\beta}^2$. These effects must sum to zero across any subscript corresponding to a fixed factor, but are independent if the random subscripts are not the same. The zero sum

Restricted model assumptions

requirement induces negative correlation among the random effects with the same random subscripts.

The scaling factors like r_{ab} are found as follows. Get the number of levels for all fixed factors involved in the interaction. Let r_1 be the product of these levels, and let r_2 be the product of the levels each reduced by 1. Then the multiplier is r_2/r_1. For an AB interaction with A fixed and B random, this is $(a-1)/a$; for an ABC interaction with A and B fixed and C random, the multiplier is $(a-1)(b-1)/(ab)$.

Scale factors in restricted model variances

10.6.1   A matrix formulation

Here we briefly write a general matrix formulation for mixed effects models. We will not be using this formulation directly, but those who are comfortable with matrix algebra may find this a unifying explanation. To be concrete, we will use a two-factor design as an example. Factor A is fixed with three levels; factor B is random with two levels and crosses with A; and $n = 2$ for a total of $N = 12$.

Let $\mathbf{y}$ be the N-vector of our data; that is, we establish an order and put our y_{ijkl} responses into that order. In our example, we will put our responses into the order $y_{111}, y_{211}, y_{311}, y_{121}, y_{221}, y_{321}, y_{112}, y_{212}, y_{312}, y_{122}, y_{222}, y_{322}$.

Let $\mathbf{X}$ be a known $N \times p$ matrix, where p is the number of degrees of freedom in the fixed effects; $p = 3$ in our example. Let $\boldsymbol{\beta}$ be a vector of length p containing the fixed effect coefficients. In our example, $\boldsymbol{\beta}$ contains the elements μ, α_1, and α_2 (recall that $\sum_i \alpha_i = 0$, so we can represent α_3 in terms of the other fixed effects as $\alpha_3 = -\alpha_1 - \alpha_2$). Do not confuse this bold beta with other (non-bold) beta coefficients in the model.

The fixed effects contribution to the model is $\mathbf{X}\boldsymbol{\beta}$. In our example, the fixed effects contributions are $\mu+\alpha_1$, $\mu+\alpha_2$, or $\mu+\alpha_3 = \mu-\alpha_1-\alpha_2$ depending on the level of factor A for that unit. In our example,

$$
\mathbf{X}\boldsymbol{\beta} =
\begin{bmatrix}
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & -1 & -1 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & -1 & -1 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & -1 & -1 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & -1 & -1
\end{bmatrix}
\begin{pmatrix}
\mu \\
\alpha_1 \\
\alpha_2
\end{pmatrix}
$$

Let $\mathbf{Z}$ be a known $N \times q$ matrix, where q is the total number of random effects in the model. Sometimes it is convenient to break $\mathbf{Z}$ into separate matrices corresponding to each random term in the model. In our example, we could have $\mathbf{Z} = [\mathbf{Z}^{[1]}\mathbf{Z}^{[2]}]$, where $\mathbf{Z}^{[1]}$ is an $N \times 2$ matrix for the B effect and $\mathbf{Z}^{[2]}$ is an $N \times 6$ matrix for the AB effect.

Let $\boldsymbol{\zeta}$ be a vector of length q containing the random coefficients. In our example, $\boldsymbol{\zeta}$ contains the elements β_1, β_2, $\alpha\beta_{11}$, $\alpha\beta_{21}$, $\alpha\beta_{31}$, $\alpha\beta_{12}$, $\alpha\beta_{22}$, $\alpha\beta_{32}$. If we split $\mathbf{Z}$ into submatrices corresponding to different random terms, we can also split $\boldsymbol{\zeta}$ in an analogous way. In our example, $\boldsymbol{\zeta}^{[1]}$ has elements β_1 and β_2, while $\boldsymbol{\zeta}^{[2]}$ has elements $\alpha\beta_{11}$, $\alpha\beta_{21}$, $\alpha\beta_{31}$, $\alpha\beta_{12}$, $\alpha\beta_{22}$, $\alpha\beta_{32}$.

The random effects contribution to our model is $\mathbf{Z}\boldsymbol{\zeta}$ (or broken out into subcomponents by terms, like $\mathbf{Z}\boldsymbol{\zeta} = \mathbf{Z}^{[1]}\boldsymbol{\zeta}^{[1]} + \mathbf{Z}^{[2]}\boldsymbol{\zeta}^{[2]}$ in our example). Each row of $\mathbf{Z}$ (or $\mathbf{Z}^{[j]}$) corresponds to a single unit in the experiment, and that row of $\mathbf{Z}$ creates a linear combination of the random effect coefficients that then get added to the fixed effect contribution. In many cases, rows of $\mathbf{Z}$ will consist of zeroes and ones, which are thus simply picking up appropriate random effect coefficients and adding them to the fixed effect contribution. However, $\mathbf{Z}$ can be more complicated, giving this formulation much of its power. In our example, we would have

$$\mathbf{Z}^{[1]}\boldsymbol{\zeta}^{[1]} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\mathbf{Z}^{[2]}\boldsymbol{\zeta}^{[2]} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \alpha\beta_{11} \\ \alpha\beta_{21} \\ \alpha\beta_{31} \\ \alpha\beta_{12} \\ \alpha\beta_{22} \\ \alpha\beta_{32} \end{pmatrix}$$

The elements of $\boldsymbol{\zeta}$ are random. We assume that $\boldsymbol{\zeta}^{[i]}$ is independent of $\boldsymbol{\zeta}^{[j]}$ for all i and j. Each vector $\boldsymbol{\zeta}^{[i]}$ is multivariate normal with mean 0 and variance/covariance matrix $\Sigma^{[i]}$. In many cases, we will assume that the elements of $\boldsymbol{\zeta}^{[i]}$ are independent with constant variance. In that case, $\Sigma^{[i]} = \sigma_i^2 \mathbf{I}$, where $\mathbf{I}$ is the identity matrix with the appropriate dimension.

Finally, we assume there is a random error on each observation. Usually we assume these errors are independent with constant variance σ_0^2, but we can reformulate the model to include different variances, autocorrelation, and so on.

Our complete model for the data is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\zeta} + \boldsymbol{\epsilon} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^{[1]}\boldsymbol{\zeta}^{[1]} + \mathbf{Z}^{[2]}\boldsymbol{\zeta}^{[2]} + \cdots + \boldsymbol{\epsilon} \end{aligned}$$

Statistically, $\mathbf{y}$ is distributed multivariate normal with mean $\mathbf{X}\boldsymbol{\beta}$ and variance/covariance $\mathbf{Z}^{[1]}\Sigma^{[1]}\mathbf{Z}^{[1]T} + \mathbf{Z}^{[2]}\Sigma^{[2]}\mathbf{Z}^{[2]T} + \cdots + \sigma_0^2\mathbf{I}$. (The T superscript indicates matrix transposition.) In our example, if we are using unrestricted model assumptions, this variance reduces to $\sigma_A^2\mathbf{Z}^{[1]}\mathbf{Z}^{[1]T} + \sigma_{AB}^2\mathbf{Z}^{[2]}\mathbf{Z}^{[2]T} + \sigma_0^2\mathbf{I}$. If we want the restricted model assumptions, then $\Sigma^{[2]} = \sigma_{AB}^2\mathbf{C} \neq \sigma_{AB}^2\mathbf{I}$, where $\mathbf{C}$ is a known matrix that ensures zero variance when adding across the levels of any fixed factor for any combination of levels for random factors. In that case, the variance is $\sigma_A^2\mathbf{Z}^{[1]}\mathbf{Z}^{[1]T} + \sigma_{AB}^2\mathbf{Z}^{[2]}\mathbf{C}\mathbf{Z}^{[2]T} + \sigma_0^2\mathbf{I}$.

Modern statistical software uses this $\mathbf{X}, \mathbf{Z}$ formulation internally to represent a wide range of potential models, including many that would be very difficult to handle using the old $\overline{y}_{ij\bullet\bullet}$ sorts of formulae.

10.7   Developing a Model

A table of data alone does not tell us the correct model. Before we can analyze data, we have to have a model on which to build the analysis. This model reflects the structure of the experiment (nesting and/or crossing of effects); how broadly we are trying to make inference (just these treatments or a whole population of treatments); and whether mixed effects should be restricted or unrestricted. Once we have answered these questions, we can build a model. Parameters are only defined within a model, so we need the model to make tests, compute confidence intervals, and so on.

Analysis depends on model

We must decide whether each factor is fixed or random. This decision is usually straightforward but can actually vary depending upon the goals of an experiment. Suppose that we have an animal breeding experiment with four sires. Now we know that the four sires we used are the four sires that were available; we did no random sampling from a population. If we are trying to make inferences about just these four sires, we treat sire as a fixed effect. On

Fixed or random factors?

the other hand, if we are trying to make inferences about the population of potential sires, we would treat sires as a random effect. This is reasonable, provided that we can consider the four sires at hand to be a random sample from the population, even though we did no actual sampling. If these four sires are systematically different from the population, trying to use them to make inferences about the population will not work well.

We must decide whether each factor is nested in some other factor or interaction. The answer is determined by examining the construction of an experiment. Do all the levels of the factor appear with all the levels of another effect (crossing), or do some levels of the factor appear with some levels of the effect and other levels of the factor appear with other levels of the effect (nesting)? For the cheese raters example, we see a different set of raters for rural and urban backgrounds, so rater must be nested in background. Conversely, all the raters taste all the different kinds of cheese, so rater is crossed with cheese type.

My model generally includes interactions for all effects that could interact, but we will see in some designs later on (for example, split plots) that not all possible interactions are always included in models. To some degree the decision as to which interactions to include is based on knowledge of the treatments and experimental materials in use, but there is also a degree of tradition in the choice of certain models.

Finally, we must decide between restricted and unrestricted model assumptions. I generally use the restricted model as a default, but we must think carefully in any given situation about whether the zero-sum restrictions are appropriate.

10.8   Hasse Diagrams

A Hasse diagram (Lohr 1995) is a graphical representation of a model showing the nesting/crossing and random/fixed structure. We can go back and forth between models and Hasse diagrams. I find Hasse diagrams to be useful when I am trying to build my model, as I find the graphic easier to work with and comprehend than a cryptic set of parameters and subscripts.

Figure 10.1 shows three Hasse diagrams that we will use for illustration. First, every term in a model has a *node* on the Hasse diagram. A node consists of a label to identify the term (for example, AB), a subscript giving the degrees of freedom for the term, and a superscript giving the number of different effects in a given term (for example, ab for $\beta_{j(i)}$). Some nodes are joined by line segments. Term U is above term V (or term V is below term U) if you can go from U to V by moving *down* line segments. For example, in Figure 10.1(b), AC is below A, but BC is not. The label for a random factor or any term below a random factor is enclosed in parentheses to indicate that it is random.

Nesting or crossing?

Which interactions?

Restricted or unrestricted?

Nodes for terms, joined by lines for above/below

Random terms in parentheses

Intercept$_1^1$ Intercept$_1^1$ Intercept$_1^1$

A$_4^5$ (B)$_3^4$ (A)$_4^5$ (B)$_3^4$ C$_1^2$ (A)$_4^5$

(A:B)$_{12}^{20}$ (A:B)$_{12}^{20}$ (A:C)$_4^{10}$ (B:C)$_3^8$ (B)$_{15}^{20}$

(Error)$_{20}^{40}$ (A:B:C)$_{12}^{40}$ (C)$_{20}^{40}$

(Error)$_{40}^{80}$ (Error)$_{40}^{80}$

(a) (b) (c)
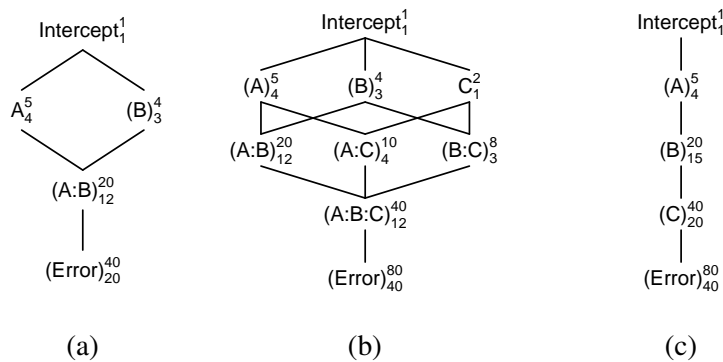
Figure 10.1: Hasse diagrams: (a) two-way factorial with A fixed and B random, A and B crossed; (b) three-way factorial with A and B random, C fixed, all factors crossed; (c) fully nested, with B fixed, A and C random. In all cases, A has 5 levels, B has 4 levels, and C has 2 levels.

10.8.1 Constructing a Hasse diagram

A Hasse diagram always has a node at the top for the grand mean (it could be called M or Intercept), a node at the bottom for random error (it could be called (E) or (Error)), and nodes for each factorial term in between. I build Hasse diagrams from the top down, but to do that I need to know which terms go above other terms. Hasse diagrams have the same above/below relationships as ANOVA tables.

Build from top down

A term U is above a term V in an ANOVA table if all of the factors in term U are in term V. Sometimes these factors are explicit; for example, factors A, B, and C are in the ABC interaction. When nesting is present, some of the factors may be implicit or implied in a term. For example, factors A, B, and C are all in the term C nested in the AB interaction. When we write the term as C, A and B are there implicitly. We will say that term U is above term V if all of the factors in term U are present or implied in term V.

Nested factors include implicit factors

Before we start the Hasse diagram, we must determine the factors in the model, which are random and which are fixed, and which nest and which cross. Once these have been determined, we can construct the diagram using the steps in Display 10.1.

Example 10.6 Cheese tasting Hasse diagram

The cheese tasting experiment of Example 10.2 had three factors: the fixed factor for background (two levels, labeled B), the fixed factor cheese type (four levels, labeled C), and the random factor for rater (ten levels, random, nested in background, labeled R). Cheese type crosses with both background and rater.

1. Start row 0 with node M for the grand mean.

2. Put a node on row 1 for each factor that is not nested in any term. Add lines from the node M to each of the nodes on row 1. Put parentheses around random factors.

3. On row 2, add a node for any factor nested in a row 1 node, and draw a line between the two. Add nodes for terms with two explicit or implied factors and draw lines to the terms above them. Put parentheses around nodes that are below random nodes.

4. On each successive row, say row i, add a node for any factor nested into a row $i-1$ node, and draw a line between the two. Add nodes for terms with i explicit or implied factors and draw lines to the terms above them. Put parentheses around nodes that are below random nodes.

5. When all interactions have been exhausted, add a node for error on the bottom line, and draw a line from error to the dangling node(s) above it.

6. For each node, add a superscript that indicates the number of effects in the term.

7. For each node, add a subscript that indicates the degrees of freedom for the term. Degrees of freedom for a term U are found by starting with the superscript for U and subtracting out the degrees of freedom for all terms above U.

Display 10.1: Steps for constructing a Hasse diagram.

Figure 10.2(a) shows the first stage of the diagram, with the Intercept node for the mean and nodes for each factor that is not nested.

Figure 10.2(b) shows the next step. We have added rater nested in background. It is in parentheses to denote that it is random, and we have a line up to background to show the nesting. Also in this row is the BC two-factor interaction, with lines up to B and C.

Figure 10.2(c) shows the third stage, with the rater by cheese RC interaction. This is random (in parentheses) because it is below rater. It is also below BC; B is present implicitly in any term containing R, because R nests in B.

Figure 10.2(d) adds the node for random error. I call this stage a skeleton, or frame-only, Hasse diagram.

Figure 10.2(e) adds the superscripts for each term. The superscript is the number of different effects in the term and equals the product of the number of levels of all the implied or explicit factors in a term.

Finally, Figure 10.2(f) adds the subscripts, which give the degrees of free-

(a)   (b)   (c)

Intercept   Intercept   Intercept

B   C   B   C   B   C

                       (R)   B:C   (R)   B:C

                                                   (C:R)

(d)   (e)   (f)

Intercept   Intercept^1 Intercept$^1_1$

B   C   B^2 C^4 B$^2_1$ C$^4_3$

(R)   B:C   (R)^{20} B:C^8 (R)$^{20}_{18}$ B:C$^8_3$

(C:R)   (C:R)^{80} (C:R)$^{80}_{54}$

(Error)   (Error)^{160} (Error)$^{160}_{80}$
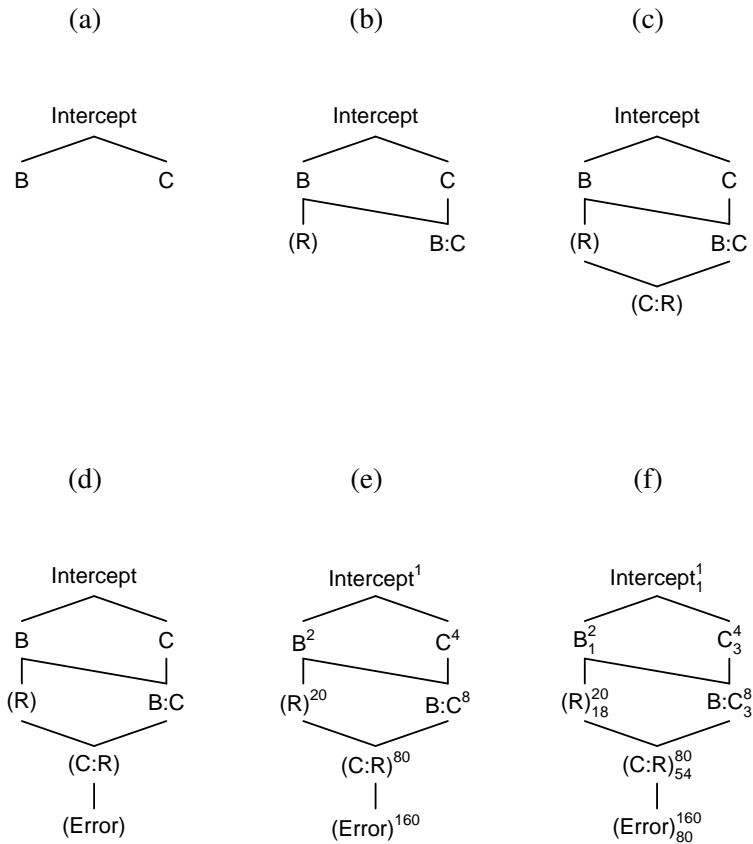
Figure 10.2: Stages in the construction of Hasse diagram for the cheese rating example.

dom. Compute the degrees of freedom by starting with the superscript and subtracting out the degrees of freedom for all terms above the given term. It is easiest to get degrees of freedom by starting with terms at the top and working down.

Perhaps the most difficult part of this text is the processing of a verbal description of an experiment into a Hasse diagram (or model). It takes time, and it takes practice, and a few more examples might help.

Example 10.7 Honey Bee Hasse

Many insects respond to odors (of food, for example), but their responses may depend on many factors. In this study, scientists wish to study the effects of genetic background, type of food, and concentration of odor on a physiological response in bees. The experiment involves 312 bees, 52 from

each of six colonies (hives). The 52 bees from each colony are divided at random into two groups of 26. Group 1 gets control food and group 2 gets experimental food. Each bee is then exposed to three different concentrations of the odor in random order, with a wash-out odor exposure between each experimental condition. Each bee is measured for its response after exposure to each experimental condition.

The first step is to identify the factors. Factors are the "reasons" why the responses are not all the same. In our case, the factors are the source of the bees (the hives, H), the type of food (F), the concentration of the odor (C), and the individual bee (B). Now we think about what is crossed and what is nested. Each individual bee is from one hive and only gets one kind of food. Thus bee must be nested in the hive by food interaction. Each bee sees every concentration, so bee crosses with concentration. Similarly, everything else crosses as well. What is fixed and what is random? Bee is probably random, as there are many more than 52 bees in each hive. Concentration and food are almost certainly fixed. Hive is less clear in this example. If we are talking about just these six hives, hive is a fixed factor. If we are trying to draw inference to the population of all potential hives and consider our six to be a random sample, then hive would be a random factor.

If we assume that hive is fixed, we get the Hasse diagram in Figure 10.3 (a). On the other hand, with hive random we get the Hasse diagram in Figure 10.3 (b).

Note that even though we have 936 observations, we have 0 degrees of freedom for estimating error. We would need to observe the response of each bee to each concentration more than once to be able to estimate error.

Example 10.8 Plant growth Hasse

We are conducting an experiment on plant growth. The experiment will be conducted at two sites, and our interest is in how the plants grow at these particular two sites. We are interested in four specific populations of plants, and from within each population we have randomly chosen 10 sub-populations called lines. At each site there are 8 parcels of land called quadrats. At each site, each population is randomly assigned to two quadrats. At each quadrat, all 10 lines from the chosen population are grown. Finally, we take two biomass measurements from each line. This gives us a total of 320 biomass measurements.

The factors in this experiment are population, line, site, and quadrat. The first three are fairly obvious, but quadrat will appear like a source of error. Some quadrats may be highly productive, and others less so. We can identify this source through the randomization, so we include it in the model. Line is clearly nested in population, but every population sees every site, and every line sees every site, so those pairs cross. There is a separate set of quadrats for every population-site combination, so quadrat is nested in the population-site interaction. The description of the experiment indicates that interest is in these particular populations and sites; thus the population and site factors are fixed (we are not making inference to some encompassing universe of sites
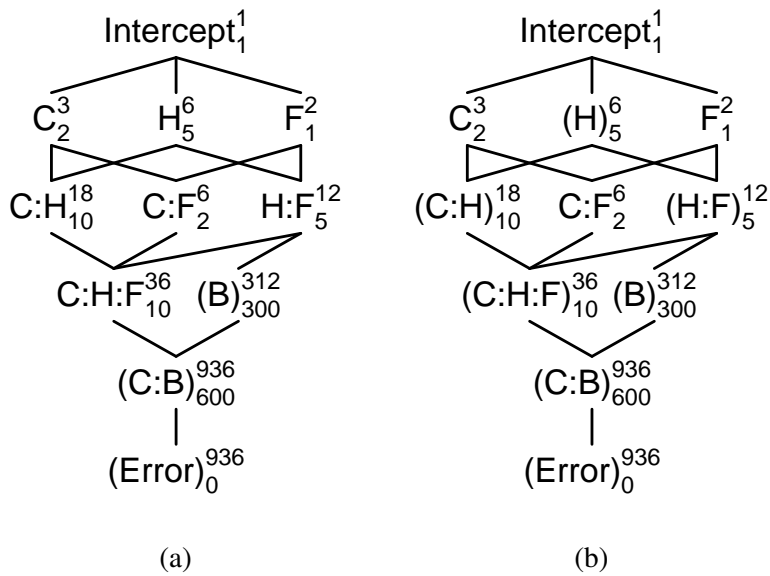
Figure 10.3: Hasse diagrams for the honey bee food and odor concentration experiment in Example 10.7: (a) hive assumed to be fixed; (b) hive assumed to be random. H is hive, C is concentration, F is food type, and B is individual bee.

or populations). Line, however, is random and quadrat will be random.

The Hasse diagram for this example is in Figure 10.4.

Example 10.9 Personality Types

A psychologist wishes to explore a theory about interactions between young men and women. She has identified four male personality types and five female personality types. Five young men of each type are selected at random (total 20 men), and 5 young women of each type are selected at random (total 25 women). Each of the 20 men will meet each of the 25 women for a five minute conversation. The conversation is taped, and the psychologist rates each conversation for its degree of interaction.

The factors here are the female and male personality types as well as the individual men and women in the experiment. The personality types are fixed factors; there are four for men and five for women and no additional types that these are supposed to represent. Individual men and women are random factors. The male personality types cross with the female personality types, and the individual males cross with the individual females. However, males are nested in male personality type, and females are nested in female personality type.

The Hasse diagram for this example is in Figure 10.5. Note that, again in
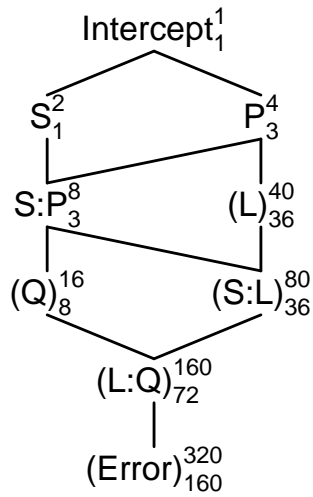
Figure 10.4: Hasse diagram for the plant growth experiment in Example 10.8. S is site; P is population; L is line; and Q is quadrat.
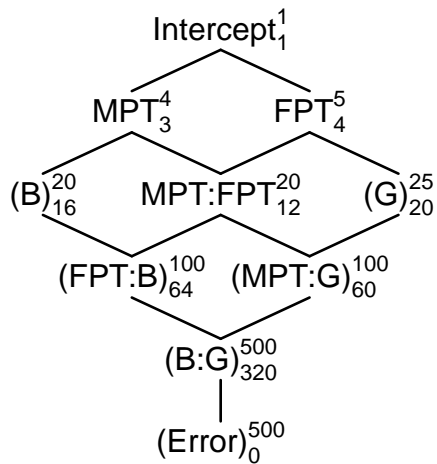
Figure 10.5: Hasse diagram for the conversation experiment in Example 10.9. MPT is male personality type; FPT is female personality type; B is man (the boys); and G is woman (the girls).

this example, there are no degrees of freedom to estimate error, because we only get one observation for each man:woman combination.

$$\text{Intercept}_1^1$$

$$O_3^4$$

$$(S)_8^{12}$$

$$(P)_{36}^{48}$$

$$(\text{Error})_0^{48}$$
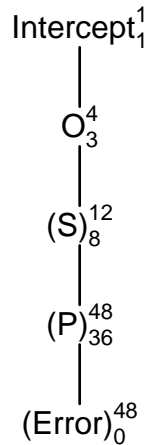
Figure 10.6: Hasse diagram for the pacemaker pins experiment in Example 10.10. O is operator; S is substrate; and P is pin.

Example 10.10 Pacemaker Pins

Consider the situation of Problem 3.1. Cardiac pacemakers contain electrical connections that are platinum pins soldered onto a substrate. The question of interest is whether different operators produce solder joints with the same strength. Twelve substrates are randomly assigned to our four operators. Each operator solders four pins on each substrate, and then these solder joints are assessed by measuring the shear strength of the pins.

This question was originally about experimental units (substrates) versus measurement units (pins). Back in that chapter, we needed to average the measurement units to get a response for the experimental unit. However, with mixed effects and nesting, we can create a model that encompasses all the data.

The factors in this experiment are the operators and the substrates; we will also include the pins as a factor, but we will see that we cannot distinguish pin to pin variability from measurement error. Pins are nested in substrates (each pin can only belong to one substrate), and substrates are nested in operators (each substrate was only assembled by one operator). Substrate is a random factor, being randomly assigned to operators. If there are only four pins per substrate, pins is a fixed factor (nested inside a random factor). If we chose four of many pins, then pins is a random factor. The problem description seems to imply that operator is a fixed factor ("our four operators" indicating that there are no others). The fixed/random status of pin (and potentially operator) should be verified with the subject matter expert.

The Hasse diagram is shown in Figure 10.6. Note that the term "pin nested in substrate" will be random whether or not we think that the factor

pin itself is random, because pin is below substrate, and substrate is random. Thus the Hasse diagram looks the same whether pin is fixed or random.

10.9   Random Coefficient Models

Random coefficient models are the convergence of two types of models we have seen before: polynomial (dose response) models and random effects. In the random effects models we have considered so far, each random effect is simply added to an appropriate subset of units; effectively, each random effect is multiplied by 0 or 1 depending on whether the effect should be included in that unit. In random coefficient models, a random effect can be multiplied by something other than 0 or 1 before being added to the unit.

Random coefficients for polynomial terms

Consider the honey bee experiment in Example 10.7. One of the factors in that model was concentration. Bee was a random effect in that experiment, and we included a random interaction term between concentration and bee. In the example, we just considered concentration to be a factor with three levels, but we could have done polynomial modeling with linear and quadratic terms. Similarly, we could have included a linear in concentration by bee random term, and a quadratic in concentration by bee random term. These terms allow the overall linear and quadratic coefficients to be modified on a bee by bee basis, with these modifications considered to be random with mean zero and a variance to be estimated.

In the more mathematical framing of Section 10.6.1, the powers of the quantitative predictor appear in the columns of $\mathbf{Z}$.

10.10   Staggered Nested Designs

One feature of standard fully-nested designs is that we have few degrees of freedom for the top-level terms and many for the low-level terms. For example, in Figure 10.1(c), we have a fully-nested design with 4, 15, 20, and 40 degrees of freedom for A, B, C, and error. This difference in degrees of freedom implies that our estimates for the top-level variance components will be more variable than those for the lower-level components. If we are equally interested in all the variance components, then some other experimental design might be preferred.

Ordinary nesting has more degrees of freedom for nested terms

Staggered nested designs can be used to distribute the degrees of freedom more evenly (Smith and Beverly 1981). There are several variants of these designs; we will only discuss the simplest. Factor A has a levels, where we'd like a as large as feasible. A has $(a-1)$ degrees of freedom. Factor B has two levels and is nested in factor A; B appears at two levels for every level of A. B has $a(2-1) = a$ degrees of freedom. Factor C has two levels and is nested in B, but in an unbalanced way. Only level 2 of factor B will have two levels of factor C; level 1 of factor B will have just one level of factor C. Factor D is nested in factor C, but in the same unbalanced way. Only level 2 of factor C will have two levels of factor D; level 1 of factor C will have just one level of

Staggered nested designs nest in an unbalanced way
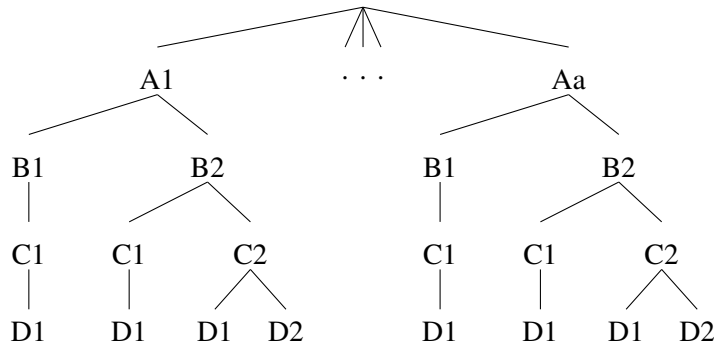
Figure 10.7: Example of staggered nested design.

factor D. Any subsequent factors are nested in the same unbalanced fashion. Figure 10.7 illustrates the idea for a four-factor model.

For a staggered nested design with h factors (counting error), there are ha units. There is 1 degree of freedom for the overall mean, $a - 1$ degrees of freedom for A, and a degrees of freedom for each nested factor below A.

10.11 Problems

Consider a four-factor model with A and D fixed, each with three levels. Factors B and C are random with two levels each. There is a total of 72 observations. All factors are crossed. Draw the Hasse diagram. **Exercise 10.1**

Consider a four-factor model with A and D fixed, each with three levels. Factors B and C are random with two levels each. B nests in A, C nests in B, and D crosses with the others. There is a total of 72 observations. Draw the Hasse diagram. **Exercise 10.2**

Consider a four-factor model with A and D fixed, each with three levels. Factors B and C are random with two levels each. B nests in A, C nests in D, and all other combinations cross. There is a total of 72 observations. Draw the Hasse diagram. **Exercise 10.3**

Draw the Hasse diagram for each of the following experiments. **Problem 10.1**

(a) We are interested in the relationship between atmospheric sulfate aerosol concentration and visibility. As a preliminary to this study, we examine how we will measure sulfate aerosol. Sulfate aerosol is measured by drawing a fixed volume of air through a filter and then chemically analyzing the filter for sulfate. There are four brands of filter available and two methods to analyze the filters chemically. We randomly select eight filters for each brand-method combination. These 64 filters are then used (by drawing a volume of air with a known concentration of sulfate through the filter), split in half, and both halves are

chemically analyzed with whatever method was assigned to the filter, for a total of 128 responses.

(b) A research group often uses six contract analytical laboratories to determine total nitrogen in plant tissues. However, there is a possibility that some labs are biased with respect to the others. Forty-two tissue samples are taken at random from the freezer and split at random into six groups of seven, one group for each lab. Each lab then makes two measurements on each of the seven samples they receive, for a total of 84 measurements.

(c) A research group often uses six contract analytical laboratories to determine total nitrogen in plant tissues. However, there is a possibility that some labs are biased with respect to the others. Seven tissue samples are taken at random from the freezer and each is split into six parts, one part for each lab. We expect some variation among the subsamples of a given sample. Each lab then makes two measurements on each of the seven samples they receive, for a total of 84 measurements.

Problem 10.2

An investigative group at a television station wishes to determine if doctors treat patients on public assistance differently from those with private insurance. They measure this by how long the doctor spends with the patient. There are four large clinics in the city, and the station chooses three pediatricians at random from each of the four clinics. Ninety-six families on public assistance are located and divided into four groups of 24 at random. All 96 families have a one-year-old child and a child just entering school. Half the families will request a one-year checkup, and the others will request a preschool checkup. Half the families will be given temporary private insurance for the study, and the others will use public assistance. The four groupings of families are the factorial combinations of checkup type and insurance type. Each group of 24 is now divided at random into twelve sets of two, with each set of two assigned to one of the twelve selected doctors. Thus each doctor will see eight patients from the investigation. Recap: 96 units (families); the response is how long the doctor spends with each family; and treatments are clinic, doctor, checkup type, and insurance type. Draw the Hasse diagram.

Problem 10.3

Eurasian water milfoil is an exotic water plant that is infesting North American waters. Some weevils will eat milfoil, so we conduct an experiment to see what may influence weevils' preferences for Eurasian milfoil over the native northern milfoil. We may obtain weevils that were raised on Eurasian milfoil or northern milfoil. From each source, we take ten randomly chosen males (a total of twenty males). Each male is mated with three randomly chosen females raised on the same kind of milfoil (a total of 60 females). Each female produces many eggs. Eight eggs are chosen at random from the eggs of each female (a total of 480 eggs). The eight eggs for each female are split at random into four groups of two, with each set of two assigned to one of the factor-level combinations of hatching species and growth species (an egg may be hatched on either northern or Eurasian milfoil, and after hatching grows to maturity on either northern or Eurasian

milfoil). After the hatched weevils have grown to maturity, they are given ten opportunities to swim to a plant. The response is the number of times they swim to Eurasian. Draw the Hasse diagram.

City hall wishes to learn about the rate of parking meter use. They choose eight downtown blocks at random (these are *city* blocks, not *statistical* blocks!), and on each block they choose five meters at random. Six weeks are chosen randomly from the year, and the usage (money collected) on each meter is measured every day (Monday through Sunday) for all the meters on those weeks. Draw the Hasse diagram.

Problem 10.4

We want to investigate the color richness of wood stain applied to various woods. We work with boards that are made of either oak, maple, or pine. The boards can be either "new" (that is, freshly cut) or "old" (cut one year ago). All boards are .75 inches thick, 3.5 inches wide, and 6 inches long. We have 18 boards, three from each of the species by age combinations. To each board we will apply two stain colors (walnut or mahogony), with the colors randomly applied to the two sides of the board. After applying the stain, all boards are finished with a standard varnish. Each side of each board is then evaluated for color richness.

Problem 10.5

Draw a Hasse diagram for this experiment (just the basic diagram, no subscripts or superscripts).

Carpal tunnel syndrome is a repetitive motion disorder that produces pain in the wrists when sufferers write, type, or do other repetitive uses of their hands. We wish to study the effects of two medications and two rest strategies on the relief of pain. Twenty subjects are recruited, 10 men and 10 women. Each subject will participate in the study for four weeks. During each week of the study, a subject will use one of the four combinations of medication and rest strategy. The order of the four combinations is randomized separately for each subject. At the end of each week, subjects report their level of pain as the response. We do not expect that all the medications and rest strategies will be equally effective for all subjects.

Problem 10.6

Draw a Hasse diagram for this experiment.

MDMP (2-methoxy-3,5-dimethylpyrazine) is a chemical compound that is responsible for some off odors in wines, including the finest wines. It is said to smell like "an old damp dishcloth that has gone moldy with slightly coffee, slightly nutty overtones," so you can imagine that wine makers are eager to keep this stuff out of their wine. There is speculation that this compound comes from a bacterium, and the source of the bacterium could be oak chips or natural corks or maybe both. If it turns out to be oak chips, some winemakers believe that toasting the oak chips before use will kill the bacteria.

Problem 10.7

The following experiment was conducted. There are five suppliers of oak chips. We obtain three batches of chips from each supplier. From each batch we take two samples. One sample is randomly selected; this sample is toasted and the other sample is left untoasted. Two barrels of wine are produced from each sample, and four bottles of wine are sampled from each

barrel. That is 240 bottles of wine, and from each of these bottles we measure MDMP concentration.

Draw a Hasse diagram for this experiment.

A new instrument is being evaluated for measuring concentration of DNA in solution. It is not expected to be more accurate (in fact, it may be less accurate), but it is quick and cheap. An additional quirk is that the result may depend on the volume of sample that is used in the measurement. There is a loaner instrument that we can try before we buy, and we will assess how accurate the new instrument is on our particular population of samples.

Problem 10.8

We have a lab full of technicians and a storage freezer full of lots of different sample solutions that have had their DNA concentration measured by the current "gold standard" instrument. Three lab technicians are chosen at random, and each technician is directed to go into the freezer and choose five random sample solutions. Each technician then withdraws 12 aliquots (subsamples) from each of the sample solutions. However, these aliquots have six different volumes, so each technician has two aliquots of each of the six volumes for each of their five sample solutions. The technicians then use the loaner instrument to measure DNA concentrations of all of their aliquots (this done in random order), and the response is the percentage variation in DNA measured on the new instrument relative to the measurement on the gold standard instrument.

Construct a Hasse diagram for this experiment.

The following experiment was conducted (seriously, something very similar to this was run ..., you just can't make this stuff up). Interest focusses on whether breast size on a female hitchhiker affects the likelihood of drivers to stop for the hitchhiker. The test hitchhiker is an "average looking confederate with A cup breasts." The three treatments are the hitchhiker as is, or augmented with either B cup or C cup latex appliances. During three consecutive time slots, the woman goes to the side of a highway and sticks out her thumb to hitchhike. The breast size is randomized to the three time slots. In each time slot, a hidden confederate monitors the numbers of male and female drivers that pass, and the time slot is done when 200 men and 200 women drivers have been observed. The response is whether or not a driver stops to offer a ride. The woman does not accept any rides, and there is hidden security available as well.

Problem 10.9

(a) Draw a Hasse diagram for this study.

(b) What *statistical* criticisms would you make of this study?

An opened bottle of wine will deteriorate and become undrinkable. Various products purport to help preserve opened bottles. In this experiment, we use three different varieties (Merlot, Cabernet, and Shiraz) from each of five randomly chosen vinyards (just labeled A through E). We obtain six bottles of each variety from each vinyard. These six are randomly assigned to three different wine preservation products, two bottles to each of the three products. The bottles are opened, one half of the wine is poured out, and then the bottles are closed with their assigned product. Four days later, a professional

Problem 10.10

wine judge tastes wine from each of the bottles (fully blinded) and gives a rating on a 1–100 scale for the wine.

Draw a Hasse diagram for this experiment.

People's tendency to go along with the crowd, also called the bandwagon effect or herd instinct, may explain part of consumer behavior. Psychologists would like to learn more about the role of social cues in advertising. This study examines three factors in the success of an advertisement. The first is visual: whether the advertisement shows a single person or a group of four enjoying the product (Champagne). The second is textual: whether the text makes a quality statement (taste is everything) or a social statement (everyone enjoys it). Finally, the third factor is personal. Some people are "high self-monitoring"; high self-monitoring people use social cues to adjust their own behavior to agree with social norms. Low self-monitoring people are more likely to go their own way.

Problem 10.11

In this experiment, student volunteers from a large psychology class took a questionnaire to determine their self-monitoring status. After scoring the questionnaires, 100 high self-monitors were selected at random, and 100 low self-monitors were selected at random. The 100 high self-monitors were randomly assigned to the four combinations of visual and verbal cues, 25 per combination. The same was done for the low self-monitors. Volunteers then viewed their advertisement and gave as their response their rating of the product.

Draw the Hasse diagram for this experiment.

Poly-3-hydroxybutyrate (PHB) is a biologically produced polymer that is becoming popular because it is biodegradable. This experiment studies the lab method used for measuring the concentration of PHB in a sample. The overall method is to digest samples of known concentration and then measure the concentration via gas chromotography. The procedure involves internal standards, recalibration of instruments, various independent dilutions, and so on. In particular, at this stage we have to deal with the possibility that everything could interact with everything else.

Problem 10.12

We will do this on three randomly chosen days. On each day we will make up eight samples of PHB. Each sample is randomly assigned to one of the combinations of four concentrations and two digestion methods. Once we have made the sample, we will measure the concentration twice on the GC.

Draw the Hasse diagram for this experiment.

My daughters have supplied their Christmas wish lists (single spaced, double column, multipage—enough to bankrupt Bill Gates). These lists include many CDs and DVDs. You can buy these on-line or at "brick and mortar" stores. Being an impoverished academic, I'm always looking for good prices, so I collect some data. I randomly choose four each of CDs and DVDs from their combined wish list. From a list of retail and online stores, I randomly choose three brick and mortar stores and three on-line stores that sell digital media. I then price the eight selected items at the six selected stores.

Problem 10.13

Construct a Hasse diagram for analyzing the collected prices.

Arthritis can cause pain that limits shoulder range of motion. This experiment is conducted to compare how three drugs affect arthritis pain in the evening. In addition to studying the drugs, the experiment also looks at how time of medication (morning or noon) and type of arm motion (up/down, front/back, or circular) affect perceived pain.

Problem 10.14

Thirty-six women aged 55 to 65 are randomly assigned to the six factor/level combinations of drug and medication time, six women per group. All the women stay on the medication for two weeks and then go to the clinic for the pain study. Each woman does all three motions in randomly assigned order and gives a pain rating for each motion. We thus have a total of 108 pain observations.

Draw a Hasse diagram for this experiment.

Plant breeders are trying to produce good hybrid barley. As part of the study, nine "parental" lines are crossed with three inbred "tester" lines producing 27 hybrids. The experiment is to determine the factors affecting the yields of the 27 hybrids. The experiment is conducted in four locations (Crookston, Waseca, Morris and St. Paul). At each location, four fields are chosen at random. Each field is split into 27 strips, and the 27 hybrids are randomly assigned to the 27 strips in each field. At the end of the season, the seed yield is determined for each strip (g/strip).

Problem 10.15

For our purposes, we may consider the four locations to be a random sample of locations. We anticipate that differences in meteorology and soils will cause yield differences by location. Similarly, some hybrids to do better in some locations, and so on.

Draw a Hasse diagram for this experiment.

The following experiment was conducted to study whether and how plants adapt to environment. Three populations of the same plant species are available (MN, KS, and OK). Five plants from each population are chosen as males (total of 15 males). Four plants from each population are chosen for each male; these plants will be females (total of 60 females). Pollen from each male plant is used to fertilize the flowers on its assigned female plants. Three seeds are collected from each pollinated female (total 180 seeds). These three seeds are randomly assigned to one of three growth environments. The response is the height of each plant after 10 weeks of growth.

Problem 10.16

Construct a Hasse diagram for this design.

An experiment is performed to determine the effects of different pasteurization methods on bacterial survival. We work with whole milk, 2% milk, and skim milk. We obtain four gallons of each kind of milk from a grocery store. These gallons are assumed to be a random sample from all potential gallons. Each gallon is then dosed with an equal number of bacteria. (We assume that this dosing is really equal so that dosing is not a factor of interest in the model.) Each gallon is then subdivided into two parts, with the two pasteurization methods assigned at random to the two parts. Our observations are 24 bacterial concentrations after pasteurization. Draw the Hasse

Problem 10.17

diagram.

Consider the following study: adolescent subjects (N=119) are classified as to whether they have major depressive disorder (n=30), bipolar disorder (n=45), or are healthy controls (n=44); all bipolar and depressive subjects are in remission from their disorder (generally due to drug therapy). Each subject takes three mathematics tests: WRAT-R2, PIAT, and BAFPETOA (be glad I used the acronyms) in that order; we observe their scores and analyze for differences.

Problem 10.18

(a) Is this a randomized, controlled experiment as we have defined one in this course; why or why not?

(b) If this is an experiment, describe the kinds of inferences that can be made. If this is not an experiment, describe the kinds of problems we might have with inference.

Start with a four by three table of independent normals with mean 0 and variance 1. Compute the row means and then subtract out these row means. Find the distribution of the resulting differences and relate this to the restricted model for mixed effects.

Question 10.1

Chapter 11

Inference for Random and Mixed-Effects Models

Inference in random and mixed-effects models is primarily about (interval) estimates and/or tests for the fixed effects and (interval) estimates and/or tests for the variances of the random terms. To a lesser degree, we may be interested in predicting (estimating) the random effects themselves, but whether this is needed is specific to the problem. This chapter discusses three approaches to inference: a modern REML approach, the classical expected mean squares approach, and the Bayesian approach. All three approaches have advantages and disadvantages, and it pays to be aware of all three.

REML, classical, and Bayes

The classical approach is just dead easy in simple cases, becomes a bit challenging in even modestly more complicated cases, and is very difficult to use with unbalanced data. In its simplest form, it carries the potential risk of negative estimates of variances. However, the classical approach provides useful insights into a problem, and, in particular, it is a good way to consider the issue of power (sample sizes might not be what you think they are with random effects).

The REML approach is never really dead easy to implement or understand, but it does not get significantly more difficult as the problem becomes more complicated, or even unbalanced. In simple, balanced situations, REML results will be the same as classical results.

Bayesian analysis of mixed effects is practically identical to Bayesian analysis of fixed effects. In addition, some of the more challenging things to do well in the non-Bayesian approaches (interval estimates of variance components in particular) are just sitting there for us in the summary of Bayesian model fits.

Before going into the details, we need to state up front that making inferences about variances is inherently more difficult than making inference about means. This shows up in a couple of ways. First, it takes a lot of data to estimate a variance precisely. For example, assume that you can get data from a normal distribution with mean 0 and variance σ^2 and you want to have

Inference about a variance is difficult

a 95% confidence interval for σ^2 that is no longer than $.2\sigma^2$ (roughly plus or minus 10%). That does not seem like much to ask, but you will need a sample size of about 800 to achieve that goal. Second, the quality of our inference for variances (interval estimates in particular) is much more sensitive to normality than what we have experienced for means. This can substantially narrow the range of applicability for our methods.

11.1 Restricted Maximum Likelihood

The standard frequentist methods we have discussed are based on the assumption that the data have a mean structure, but the variability from unit to unit is independent, normally distributed, with constant variance. These assumptions lead to ordinary least squares (OLS), where we choose the mean parameters to minimize the sum of squared errors. Random and mixed-effects models generally lead to dependent data, so the OLS assumptions are not met. If we know the covariance structure of the data V, or we know it up to a multiple of a known matrix (that is, $V = \sigma^2 V_1$ with V_1 known), then we can use generalized least squares (GLS) to account for the correlations. GLS models linear combinations of the response by the same linear combinations of the predictors (and errors), using linear combinations that make the (new) errors independent with constant variance. GLS estimates might, or might not, be the same as OLS estimates; when they differ, GLS estimates have less variance overall.

Ordinary versus generalized least squares

In our random and mixed effects context, we have correlations, so OLS is not optimal, but we do not know V_1, so we need to estimate the variances as well as the mean structure. The REML approach does this in two stages. First, we estimate the variances using REML. Then, conditional on the estimated variances, we estimate the mean structure using GLS. We now have variance estimates from step 1, and mean structure estimates (along with standard errors) from step 2.

REML: estimate variance, then conditional GLS

One important thing to note is that the standard errors for the mean structure estimates in step 2 are computed assuming that we know the correlation in the data. However, we only have an estimate of this correlation. Using this estimate as if it were truth hides some of the variability in the estimates. The upshot is that the variability values we get for our mean structure estimates in REML will in general be too small. Often this difference is ignorable, but not always.

REML works by first getting regression residuals for the observations modeled by the fixed effects portion of the model, ignoring at this point any correlations or random effects. We then figure out the statistical model for these residuals, which will be normal with mean 0 (because we have regressed out any contribution to the mean from the fixed effects) and a complicated covariance. This covariance is a matrix formula involving the structure of the random effects and the matrix of predictors from the fixed effects and parameterized by the variances of the random effects. The residuals are now linear combinations of the random effects part of the model and error part of

REML does MLE on regression residuals

Table 11.1: Frequency change in a quartz crystal as a function of type of crystal (gold or titanium dioxide), individual crystal, environment, and intensity. Data from K. Greden; data set `QCM`.

Type	Crs.	Env.	Intensity 1	2	3	4	5	6	7	8
Au	1	Air	1.79	3.80	6.98	11.75	10.27	34.99	62.41	101.76
		Water	4.54	6.81	11.14	16.48	24.85	38.55	65.89	101.46
	2	Air	0.97	3.31	7.31	13.36	27.17	44.09	106.72	166.07
		Water	2.57	4.71	8.29	15.42	29.45	50.85	107.31	150.15
	3	Air	1.31	3.46	8.13	16.09	34.37	67.80	132.47	185.99
		Water	2.61	4.61	10.13	20.01	31.70	63.66	111.23	160.27
TiO2	1	Air	104.24	109.36	164.77	116.13	161.74	249.43	316.59	415.39
		Water	53.77	96.67	103.45	119.67	162.26	184.79	253.73	299.62
	2	Air	70.91	90.80	136.62	124.29	168.63	234.40	284.28	384.63
		Water	12.76	22.19	31.29	45.73	71.17	104.42	165.86	244.32
	3	Air	134.52	121.22	182.30	112.39	177.58	269.68	322.57	429.10
		Water	19.86	44.69	48.90	59.97	90.70	117.40	265.10	238.60

the model. REML estimates the variances by doing maximum likelihood on the likelihood function for the residuals.

The main advantage of this approach is that the MLE applied to the residuals adjusts the variance estimates for the degrees of freedom used in the fixed effects. For example, the error variance is estimated as if using a denominator of $n - p$, where p is the number of fixed effects parameters. In the simplest case of just a common mean, the REML estimate of variance would be the usual MS_E with an $n - 1$ denominator. This unbiasing of the error variance is REML's main claim to fame.

REML variances less biased

REML makes use of a different likelihood function (for the residuals) than ordinary likelihood (for the data). In particular, the REML likelihood doesn't even depend on the fixed effects coefficients, so its achieved likelihood is also independent of the fixed effects. One consequence of this is that you cannot use REML likelihood to compare models with different fixed effects, but you can use REML likelihood to compare models with the same fixed effects but different random effects. You can use ordinary likelihood to compare models that have different fixed effects, but this will also yield biased variance estimates, and we will recommend a different approach for fixed effects.

No fixed effect inference via REML likelihood

Example 11.1 Quartz Crystal Microbalance

A quartz crystal will oscillate when an alternating current is applied. However, the frequency of the oscillation will change if some material is adsorbed to the surface of the crystal. This is the basis of the quartz crystal microbalance (QCM), which can measure in nanogram quantities. In a QCM, the crystal will typically be coated with some material and placed in

$$\text{Intercept}_1^1$$

$$I_7^8 \qquad E_1^2 \qquad T_1^2$$

$$I{:}E_7^{16} \qquad I{:}T_7^{16} \qquad E{:}T_1^4 \qquad (C)_4^6$$

$$I{:}E{:}T_7^{32} \qquad (I{:}C)_{28}^{48} \qquad (E{:}C)_4^{12}$$

$$(I{:}E{:}C)_{28}^{96}$$

$$(\text{Error})_0^{96}$$
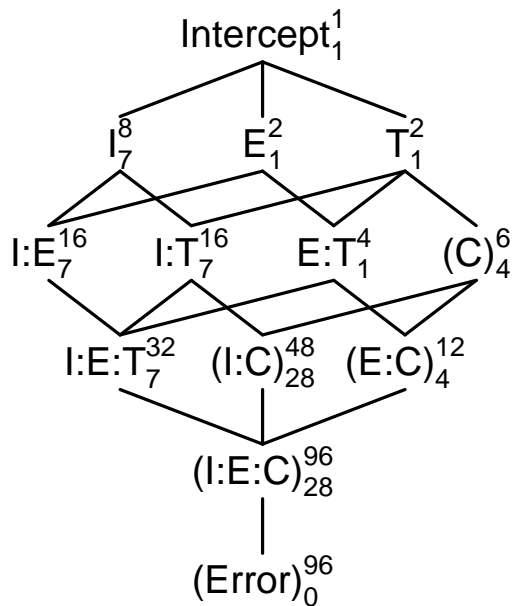
Figure 11.1: Hasse diagram for the quartz crystal microbalance experiment in Example 11.1. T is type; E is environment; I is intensity; C is crystal.

some gaseous or liquid environment. The substance to be measured is introduced to the environment, and the concentration is inferred from the change in frequency of the crystal.

In this experiment, we have two types of crystals: those coated with gold and those coated with titanium dioxide. We have three random crystals from each type of coating. The environments we will use are air and water. The main purpose of this experiment is to investigate the change in frequency when a third factor called intensity is varied across eight levels. For each individual crystal, we make 16 observations on frequency change, one for each of the combinations of environment and intensity. These 96 observations are done in random order. Note, there is not actually anything being "weighed."

There are four factors in this design: type, environment, intensity, and individual crystal. Individual crystal is random and nested in type. Otherwise, everything crosses. Here we have not applied type to crystal, we have subsampled crystals from types. Type and crystal explain variability in the data even if they are not applied experimentally in the same way as environment and intensity. Table 11.1 shows the data for this experiment, and Figure 11.1 gives the Hasse diagram.

nlme::lme and lme4::lmer are two standard functions in **R** for fitting linear mixed-effects models. lme can only estimate models that have chains of nested effects; that restriction gives it computational efficiency so that it can fit quite large data sets, and it makes some forms of inference more straightforward. lmer can fit models with crossed random effects. Both have "generalized" variants in the sense that they can fit Poisson, binomial, and other responses with analogous mixed effects predictors. We will typically use lmer by default, but, in this example, there are non-nested random effects (I:C and E:C), so we must use lmer.

> Be aware that lme and lmer *always* fit using the *unrestricted* model assumptions.

That is a slight exaggeration, because I have been able to coerce lme into fitting the restricted assumptions in one narrow class of models, but it took many lines of **R** code to do so. The classical analysis approach and the Bayesian approach with bglmm can both do both sets of model assumptions.

In an lmer formula, a random effect is indicated by something like (1|A) or (1|A:B). The term(s) to the left of the bar indicates the form of the effect being added, and the term(s) to the right of the bar indicate the different groupings of the data over which the effect should be applied. In these two examples above, the form of the model is just "1", indicating that we simply add a "constant". In this way, (1|A) is a random α_i term, and (1|A:B) is a random $\alpha\beta_{ij}$ term. There are various short cut forms as well. For example, (1|A/B) expands to (1|A)+(1|A:B).

You may put one or more non-factors (that is, continuous, regression-like predictors) to the left of the bar, but you may not put factors to the left of the bar. With a continuous predictor on the left you get a random coefficients model with grouping determined by the terms to the right of the bar. For example, (0+z|A) computes a random coefficient for the linear predictor z, separately and independently for each level of A. If there is more than one degree of freedom to the left of the bar, the corresponding coefficients are fit assuming they can be correlated. Thus (z|A), which is the same as (1+z|A), would fit random slopes for z and intercepts separately for each level of A, but allowing the slopes and intercepts to be correlated within levels of A. If you want them to be independent, you need to use (1|A)+(0+z|A).

Line 1 attempts to fit the full model. In the QCM data set, crystal is indicated as 1,2,3 for Au and 1,2,3 for TiO_2 instead of 1,2,3,4,5,6. This means that we need to use the type by crystal interaction to indicate all six crystals individually. If we had crystal indicated as 1–6, we could have used crystal by itself (although using the type by crystal interaction would not have harmed anything).
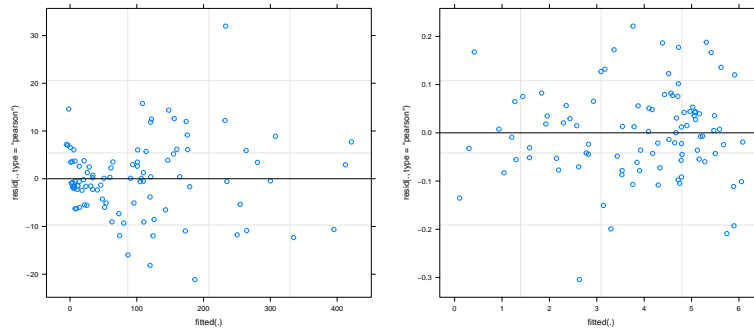
Figure 11.2: Residual plots for original (pane 1) and log-transformed (panel 2) data in the quartz crystal microbalance experiment in Example 11.1.

```
1 > fit <- lmer(y~type*env*intensity+(1|type:crystal)+(1|intensity:type:crystal)+
    (1|env:type:crystal)+(1|type:crystal:intensity:env),data=QCM)
  Error: number of levels of each grouping factor must be < number of observations
2 > fit <- lmer(y~type*env*intensity+(1|type:crystal)+(1|intensity:type:crystal)+
    (1|env:type:crystal),data=QCM)
3 > plot(fit)
4 > fit2 <- lmer(log(y)~type*env*intensity+(1|type:crystal)+(1|intensity:type:crystal)+
    (1|env:type:crystal),data=QCM)
5 > plot(fit2)
```

Our first attempt did not work. Internally, **R** represents the variance components as σ^2 and ratios of the other variances to σ^2 (for example, $\sigma^2_{crystal}/\sigma^2$). There are sound theoretical and numerical reasons to do it this way, but it does mean that we need to be able to estimate σ^2. We saw in the Hasse diagram that there were 0 degrees of freedom for error, meaning that we cannot estimate error in the full model. What we will do is drop the all-way interaction from the model. This will become a "surrogate" error.

Line 2 reduces the model and refits; line 3 plots the residuals as seen in panel 1 of Figure 11.2. These residuals are smaller on the left than on the right, indicating that the variance is non-constant. Line 4 refits with a log transformation for the response, and line 5 plots the residuals, which look much better (panel 2 of Figure 11.2). Line 6 shows (abbreviated) summary information for the log model.

```
6 > summary(fit2)
 Linear mixed model fit by REML ['lmerMod']
 Formula: log(y) ~ type * env * intensity + (1 | type:crystal) +
      (1 | intensity:type:crystal) +  (1 | env:type:crystal)
    Data: QCM

 REML criterion at convergence: 96.8

 Random effects:
  Groups                Name         Variance Std.Dev.
  intensity:type:crystal (Intercept) 0.03050  0.1746
  env:type:crystal       (Intercept) 0.04563  0.2136
  type:crystal           (Intercept) 0.01172  0.1083
  Residual                            0.02273  0.1508
 Number of obs: 96, groups:  intensity:type:crystal, 48; env:type:crystal, 12;
    type:crystal, 6

 Fixed effects:
                     Estimate Std. Error t value
 (Intercept)          3.88839    0.08141   47.76
 type1               -0.93647    0.08141  -11.50
 env1                 0.11923    0.06355    1.88
 intensity1          -1.59109    0.07814  -20.36
 intensity2          -1.03496    0.07814  -13.25
 ...
 type1:env1          -0.24470    0.06355   -3.85
 type1:intensity1    -0.65449    0.07814   -8.38
 type1:intensity2    -0.45537    0.07814   -5.83
 ...
 env1:intensity1      0.02208    0.04071    0.54
 env1:intensity2     -0.01030    0.04071   -0.25
 ...
 type1:env1:intensity1 -0.32900    0.04071   -8.08
 type1:env1:intensity2 -0.06807    0.04071   -1.67
 ...
7 > with(QCM,interactplot(intensity,type:env,log(y)))
```

The summary begins by recapitulating the model definition. The bulk of the summary information is divided into two parts: estimates of random effects and estimates of fixed effects. There is one row for each random effect giving the grouping term (the term to the right of the bar), the form of the effect (the term to the left of the bar), and the estimated variance and standard deviation. Obviously, you only need one of the last two, but sometimes one scale is needed and sometimes the other; showing both saves you a step. (Intercept) as the form of the term corresponds to 1 to the left of the bar. In this data set, the random effect variances are similar in size.

The fixed effects information looks similar to other summary information we have seen before, but with two big differences. First, the summary contains no p-values. This was a decision of the designers of lmer, and we will later see some ways to get p-values/do testing.

The second difference is more subtle. In previous models we considered, the standard errors of the estimated effects were larger and larger as we moved to higher and higher order interactions. This is because the sample size for an interaction coefficient is smaller for higher order interactions.

However, the standard errors shown after line 6 are getting smaller as we go to higher order interactions! This is because the variability affecting a term comes from random terms below the term of interest in the Hasse diagram. As you go farther down the diagram (to higher order interactions), fewer random terms can be below and thus the effective variance is smaller.

There is another way in which mixed-effects model results differ from results in the same data with all factors treated as fixed, and this is completely hidden in the usual model summary. Our predictions for random effects ("estimates" of the random quantities are often referred to as predictions) α_i in the random factor A will not be the same as those we would get by treating factor A as if it were fixed. In general, random effects are predicted to be closer to zero than the corresponding fixed effect; they are *shrunk* toward their known mean of 0. The amount of shrinking depends on the relative sizes of the different variance components, with proportionally more shrinking in a term when more of the variance in the data appears "below" the term.

One consequence is that residual plots (where the residual is the data minus our estimates of the fixed and random effects) can often show trend, even when nothing is wrong with the model. The more shrinking that was done in the prediction of random effects, the more trend will be visible.

Even lacking p-values, some of the t-statistics for interaction terms look pretty large, so examining the interaction is prudent. Line 7 produces the interaction plot in Figure 11.3. The type effect is dramatic, and the effect of environment is much stronger in TiO_2 coated crystals. The effect of intensity is stronger for Au coated crystals. Finally, the effect of intensity is stronger in air for Au coated crystals and stronger in water for TiO_2 coated crystals.

Information about the intensity factor (including its levels) was conspicuous by its absence in the source of these data, but if we assume for the sake of argument that the levels are equally spaced, Figure 11.3 suggests that a model consisting of linear in intensity rather than the full 7 degree of freedom effect might be useful. The column `int.z` in the QCM data set ranges from –3.5 to 3.5 in steps of 1. Centering this linear term at 0 makes it orthogonal to other terms and improves interpretability of the resulting models.

There are several models we could investigate while considering intensity to be quantitative. Most obviously, we can replace `intensity` with polynomials of `int.z` in the fixed effects. In random effects, we can consider the term `(0+int.z|type:crystal)`, which fits a random slope for each crystal. We could add this along with `(1|intensity:type:crystal)` or use it to replace `(1|intensity:type:crystal)` entirely. If these were fixed effects, the linear term would be redundant with the individual level effects. However, as random effects, these are not redundant. The `(0+int.z|type:crystal` term sets up a correlation where high values at one end of the intensity scale tend to go with low values at the other end (or vice versa). This kind of correlation is totally different from that of the `(1|intensity:type:crystal)` term, and the two can be used together.

While we cannot fit the full `(1|intensity:env:type:crystal)` term, we can fit a `(0+int.z|env:type:crystal)` term and still have

degrees of freedom left to estimate σ^2.

Lines 8–13 fit a variety of different models using these linear-in-intensity variables.

```
8 > fit3 <- lmer(log(y)~type*env*intensity+(1|type:crystal)+(0+int.z|type:crystal)+
      (1|intensity:type:crystal)+(1|env:type:crystal),data=QCM)
9 > fit3b <- lmer(log(y)~type*env*intensity+(int.z|type:crystal)+
      (1|intensity:type:crystal)+(1|env:type:crystal),data=QCM)
0 > fit3c <- lmer(log(y)~type*env*intensity+(int.z||type:crystal)+
      (1|intensity:type:crystal)+(1|env:type:crystal),data=QCM)
1 > fit4 <- lmer(log(y)~type*env*intensity+(1|type:crystal)+(0+int.z|type:crystal)+
      (1|intensity:type:crystal)+(1|env:type:crystal)+(0+int.z|env:type:crystal),
      data=QCM)
2 > fit5 <- lmer(log(y)~type*env*int.z+(1|type:crystal)+(1|intensity:type:crystal)+
      (1|env:type:crystal),data=QCM)
3 > fit6 <- lmer(log(y)~type*env*int.z+(1|type:crystal)+(0+int.z|type:crystal)+
      (1|intensity:type:crystal)+(1|env:type:crystal)+(0+int.z|env:type:crystal),
      data=QCM)
```

First compare the models in lines 8–10. The models on lines 8 and 10 are actually the same model, because the double bar short cut for the random by `type:crystal` term in line 10 expands to the two random by `type:crystal` terms in line 8. On the other hand, the model in line 9 differs from the model in line 8 because line 9 allows the intercept and slope (random by `type:crystal`) to be correlated, and line 8 does not.

Line 11 adds the linear-in-intensity random by `type:crystal:env` term. Lines 12 and 13 produce models analogous to lines 4 and 11 except that the 7 degree of freedom intensity fixed effect term is replaced by the quantitative intensity term.

Lines 14 and 15 show partial summary information for models `fit3` and `fit3b` (lines 8–9).

```
4 > summary(fit3)
  ...
  Random effects:
   Groups               Name        Variance Std.Dev.
   intensity.type.crystal (Intercept) 0.000000 0.00000
   env.type.crystal       (Intercept) 0.045798 0.21400
   type.crystal           int.z       0.005309 0.07286
   type.crystal.1         (Intercept) 0.015534 0.12464
   Residual                           0.021375 0.14620
  ...
5 > summary(fit3b)
  ...
  Random effects:
   Groups               Name        Variance Std.Dev. Corr
   intensity:type:crystal (Intercept) 0.000000 0.00000
   env:type:crystal       (Intercept) 0.045798 0.21400
   type:crystal           (Intercept) 0.015534 0.12464
                          int.z       0.005309 0.07286  -0.24
   Residual                           0.021375 0.14620
```

We see that the intercept and linear in intensity by crystal variances are the same in the two models. In line 14 we two lines for `type:crystal`, indi-
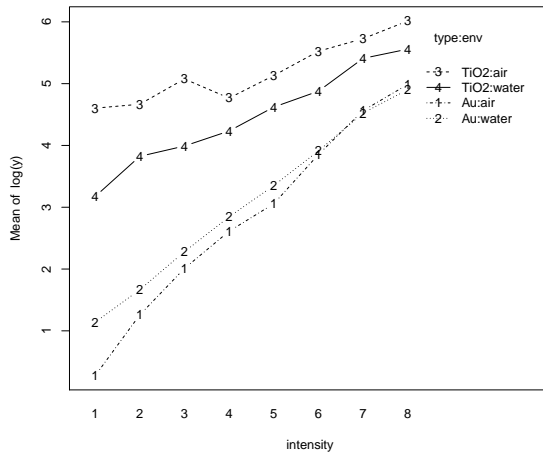
Figure 11.3: Interaction plot of intensity by type and environment for the logged response in the quartz crystal microbalance experiment in Example 11.1.

cating two independent terms, but in line 15 we only a single `type:crystal` term, and we have an estimate for the correlation between the intercept and slope random effects.

11.1.1 Inference for random terms

In principle, we can test the null hypothesis that a random effect has variance of 0 using a likelihood ratio test. We do this by fitting a second simpler model that does not include the term we want to test. We then take twice the difference of the (REML) log likelihoods of the two models as our test statistic. According to the theory of likelihood ratio tests, the LRT should be treated as a chi-square random variable with degrees of freedom equal to the difference in the number of parameters between the two model. The p-value is then the probability that a chi-square with the appropriate degrees of freedom is larger than the value of the LRT.

Notice the "in principle" above, because it turns out that the standard chi-square approximation to the distribution of the LRT does not work well when testing a null that a variance is zero. (The reason is that the null value is on the edge of the set of possible parameters instead of in the middle.) In particular, it tends to produce p-values that are too big. One rough guideline is to divide the nominal p-value by 2.

REML LRT for variance needs adjustment

Crainiceanu and Ruppert (2004) derived the exact null distribution for the LRT in linear mixed effects and provided an algorithm to simulate it. This result is too complicated to explain in detail here, but it works well as long as

Simulate exact LRT distribution

the number of fixed parameters plus the number of levels of random effects is less than the number of data (Greven, Crainiceanu, and Kuechenhoff 2008). The function `RLRSim::exactRLRT` uses this approach to approximate the p-value for tests of random effects.

Confidence intervals for variance components derived from REML results are all approximate. The built-in `confint` function in **R** provides three methods for computing these. One method is "Wald," which is simply the estimate plus or minus a multiple of the standard error of the variance estimate. This is very fast, but it is also usually a very poor estimate unless the variance component has been estimated from many levels (for example, in a single factor model the number of levels a must be large; it is not sufficient for n to be large). The "profile" method is a confidence interval consisting of all values for the parameter that would not be rejected as null hypothesis values in a (restricted) likelihood ratio test. The profile interval is an improvement over the Wald interval, because the profile interval can accommodate the inherent asymmetry when estimating variance components (the Wald interval is always symmetric). However, this interval is based on the asymptotic chi-square approximation for likelihood ratio tests, and our sample size might not be large enough for that to work well.

The best, but also the slowest, method we have available is a parametric bootstrap. In the parametric bootstrap, we simulate repeated data sets assuming that our current estimates of the variance components are correct. We then fit the model to each of the simulated data sets and observe the distribution of the estimated variance components. For example, if we simulate under the assumption that a variance component is equal to 2, and 95% of the variance components range from .75 to 5 (.375 to 2.5 times the true value), then our bootstrap confidence interval will be from 2/2.5 = .8 to 2/.375 = 5.33.

Bootstrap confidence intervals

Example 11.2 Quartz Crystal Microbalance, continued.

Continuing with the crystal frequency change data from Example 11.1, we begin on lines 1–4 by fitting the obvious model suggested by the Hasse diagram, and then refitting three times with one of the three random terms excluded in each of the new fits.

```
 1 > fit2 <- lmer(log(y)~type*env*intensity+(1|type:crystal)+
   (1|intensity:type:crystal)+(1|env:type:crystal),data=QCM)
 2 > fit2noetc <- lmer(log(y)~type*env*intensity+(1|type:crystal)+
   (1|intensity:type:crystal),data=QCM)
 3 > fit2noitc <- lmer(log(y)~type*env*intensity+(1|type:crystal)+
   (1|env:type:crystal),data=QCM)
 4 > fit2notc <- lmer(log(y)~type*env*intensity+
   (1|intensity:type:crystal)+(1|env:type:crystal),data=QCM)
 5 > logLik(fit2);logLik(fit2noetc);logLik(fit2noitc);logLik(fit2notc)
   'log Lik.' -48.39012 (df=36)
   'log Lik.' -60.33435 (df=35)
   'log Lik.' -53.96201 (df=35)
   'log Lik.' -48.4595 (df=35)
 6 > 2*(-48.3901- -48.4595)
[1] 0.1388
 7 > pchisq(.139,1,lower.tail=FALSE)
[1] 0.7092772
```

Line 5 computes the (REML) log likelihoods for these models (recall that these are comparable because all of the models have the same fixed effects). Removing the (1|type:crystal) effect barely changes the log likelihood, but removing either of the others makes a substantial change to the log likelihood. Line 6 computes the LRT, and line 7 computes the nominal p-value.

In order to use RLRsim::exactRLRT to get good p-values, we need to have three model fits: the full model, the null model with the random term of interest removed, and the reduced model containing the effect of interest as the only random effect.

```
 8 > fit2etc <- lmer(log(y)~type*env*intensity+(1|env:type:crystal),data=QCM)
 9 > fit2itc <- lmer(log(y)~type*env*intensity+(1|intensity:type:crystal),data=QCM)
10 > fit2tc <- lmer(log(y)~type*env*intensity+(1|type:crystal),data=QCM)
11 > RLRsim::exactRLRT(fit2etc,fit2,fit2noetc)
   ...
   RLRT = 23.888, p-value < 2.2e-16
12 > RLRsim::exactRLRT(fit2itc,fit2,fit2noitc)
   ...
   RLRT = 11.144, p-value = 7e-04
13 > RLRsim::exactRLRT(fit2tc,fit2,fit2notc)
   ...
   RLRT = 0.13876, p-value = 0.2753
```

Lines 8–10 compute the reduced models, and lines 11–13 illustrate the exact test. The p-value for (1|type:crystal) shrinks from the nominal .71 down to the corrected .28. This is a substantial reduction (a bit more than the rough and ready divide the p-value by 2), but the p-value is still large enough to indicate the random effect is unneeded.

Let us repeat this testing program on fit4, the model that contains the random effects from the Hasse diagram plus the linear in intensity by type and crystal and by environment, type, and crystal. Lines 14–19 compute the full and null models,

```
14 > fit4 <- lmer(log(y)~type*env*intensity+(1|type:crystal)+(0+int.z|type:crystal)+
   (1|intensity:type:crystal)+(1|env:type:crystal)+(0+int.z|env:type:crystal),
   data=QCM)
15 > fit4nozetc <- lmer(log(y)~type*env*intensity+(1|type:crystal)+
   (0+int.z|type:crystal)+(1|intensity:type:crystal)+(1|env:type:crystal),
   data=QCM)
16 > fit4noetc <- lmer(log(y)~type*env*intensity+(1|type:crystal)+
   (0+int.z|type:crystal)+(1|intensity:type:crystal)+(0+int.z|env:type:crystal),
   data=QCM)
17 > fit4noitc <- lmer(log(y)~type*env*intensity+(1|type:crystal)+
   (0+int.z|type:crystal)+(1|env:type:crystal)+(0+int.z|env:type:crystal),
   data=QCM)
18 > fit4noztc <- lmer(log(y)~type*env*intensity+(1|type:crystal)+
   (1|intensity:type:crystal)+(1|env:type:crystal)+(0+int.z|env:type:crystal),
   data=QCM)
19 > fit4notc <- lmer(log(y)~type*env*intensity+(0+int.z|type:crystal)+
   (1|intensity:type:crystal)+(1|env:type:crystal)+(0+int.z|env:type:crystal),
   data=QCM)
```

and lines 20–25 compute the reduced models:

```
20 > fit4zetc <- lmer(log(y)~type*env*intensity+(0+int.z|env:type:crystal),data=QCM)
21 > fit4etc <- lmer(log(y)~type*env*intensity+(1|env:type:crystal),data=QCM)
22 > fit4itc <- lmer(log(y)~type*env*intensity+(1|intensity:type:crystal),data=QCM)
23 > fit4ztc <- lmer(log(y)~type*env*intensity+(0+int.z|type:crystal),data=QCM)
24 > fit4tc <- lmer(log(y)~type*env*intensity+(1|type:crystal),data=QCM)
25 > RLRsim::exactRLRT(fit4zetc,fit4,fit4nozetc)
...
RLRT = 8.1588, p-value = 0.0012
26 > RLRsim::exactRLRT(fit4etc,fit4,fit4noetc)
...
RLRT = 40.613, p-value < 2.2e-16
27 > RLRsim::exactRLRT(fit4itc,fit4,fit4noitc)
...
RLRT = 0.9839, p-value = 0.1611
28 > RLRsim::exactRLRT(fit4ztc,fit4,fit4noztc)
...
RLRT = 2.7091, p-value = 0.0341
29 > RLRsim::exactRLRT(fit4tc,fit4,fit4notc)
...
RLRT = 0.2304, p-value = 0.2417
```

Then lines 26–29 compute the p-values for testing these random effects. As before, we find there is no evidence that the (1|type:crystal) term is needed. We also see that the (1|intensity:type:crystal) is not needed when we have the (0+int.z|type:crystal) term. In general, random effects involving both crystal and environment tend to be largest.

Finally, we refit using only the needed random terms on line 30,

```
30 fit7 <- lmer(log(y)~type*env*intensity+(0+int.z|type:crystal)+(1|env:type:crystal)+
   (0+int.z|env:type:crystal),data=QCM)
31 > confint(fit7,method="boot",oldNames=FALSE)
   Computing bootstrap confidence intervals ...
                                         2.5 %        97.5 %
   sd_int.z|env:type:crystal        1.031776e-03  0.0844182033
   sd_(Intercept)|env:type:crystal  1.438325e-01  0.4458674781
   sd_int.z|type:crystal            1.579681e-08  0.1469088319
   sigma                            9.939023e-02  0.1482196628
   ...
```

and compute parametric bootstrap confidence intervals for the random effects
on line 31. Note the `method="boot"` to request the bootstrap intervals,
and the `oldNames=FALSE` to request human readable names for the vari-
ance components. Because these are bootstrap intervals, the results will be
slightly different each time. In line 28, the `(0+int.z|type:crystal)`
term had a p-value of .034, and in line 31 we see that the lower bound for
the confidence interval is nearly 0. Zero can be in the confidence interval, as
we see for the `(1|type:crystal)` term in the `fit2` model as shown on
line 32.

```
32 > confint(fit2,method="boot",oldNames=FALSE)
   Computing bootstrap confidence intervals ...
                                                2.5 %        97.5 %
   sd_(Intercept)|intensity:type:crystal  0.1596627402  0.303183324
   sd_(Intercept)|env:type:crystal        0.0733674456  0.404806555
   sd_(Intercept)|type:crystal            0.0000000000  0.379323783
   sigma                                  0.0918880577  0.158990062
     ...
```

We need to close this section with a warning:

> Confidence intervals for variance components are *very* sensitive to non-
> normality.

The coverage for your confidence interval can be far from nominal, even
when the random effects are only slightly non-normal. More data does not
solve this problem (unlike a confidence interval for a mean, where more data
does solve the problem).

11.1.2   Inference for fixed terms

You cannot use restricted likelihood to test fixed effects; it simply doesn't
work, and there is no reason that it should work. Remember, the first thing
that REML does is remove fixed effects from the model. Because it would
be removing different fixed effects in the two cases, the log likelihoods for
the larger and smaller models are not commensurate. You could use ordinary
likelihood instead of restricted likelihood and get an LRT; that will work, but
we have other alternatives.

It is relatively straightforward to compute an "F" statistic for a null hypothesis about fixed effects using the (estimated) variance/covariance matrix. The numerator degrees of freedom for the F are also straightforward. If the random effects are all nested in a chain, there are good heuristics for computing a denominator degrees of freedom. The `lme` function in **R** only handles nested random effects, so an `anova` of an `lme` object produces F-tests complete with numerator and denominator degrees of freedom and p-values.

<div style="text-align: right">Denominator df for F can be difficult</div>

However, if random effects are crossed, as they can be in `lmer`, none of the denominator degree of freedom heuristics works well across the range of possible models, and a call to `anova` after `lmer` will not produce denominator degrees of freedom or p-values.

Kenward and Roger (1997) created a method for approximating the denominator degrees of freedom for F-tests of fixed effects in a mixed-effects model, and this method is implemented in `car::Anova` (and in other packages as well). This will produce Type II tests of the factorial effects using the Kenward and Roger approximation. `pbkrtest::KRmodcomp` lets you compare full and nested models that differ by more than a single term.

<div style="text-align: right">Kenward and Roger df</div>

Example 11.3 Quartz Crystal Microbalance, continued.

The crystal experiment in Example 11.1 contains three fixed factors: type, environment, and intensity. In the context of purely fixed effects, we would probably begin with an analysis of variance and continue with various contrasts and pairwise comparisons. We would examine interactions and perhaps consider polynomial modeling. We can do all of those things in the mixed-effects context.

Lines 1 and 2 use `car::Anova` to get Kenward and Roger type II tests of the fixed effects.

```
1 > car::Anova(fit2,test="F")
  Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)

  Response: log(y)
                         F Df Df.res    Pr(>F)
  type             132.3113  1      4 0.0003261 ***
  env                3.5194  1      4 0.1338965
  intensity        163.2944  7     28 < 2.2e-16 ***
  type:env          14.8249  1      4 0.0182978 *
  type:intensity    25.3333  7     28 1.617e-10 ***
  env:intensity      1.1943  7     28 0.3380488
  type:env:intensity 13.9297  7     28 1.248e-07 ***
2 > car::Anova(fit7,test="F")
  Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)

  Response: log(y)
                         F Df Df.res    Pr(>F)
  type             164.4238  1  8.000 1.291e-06 ***
  env                2.6653  1  8.000 0.1412052
  intensity         31.0984  7 32.576 1.130e-12 ***
  type:env          11.2268  1  8.000 0.0100707 *
  type:intensity     7.1626  7 32.576 3.344e-05 ***
  env:intensity      1.5142  7 32.576 0.1972937
  type:env:intensity  6.3135  7 32.576 0.0001003 ***
```

The difference between lines 1 and 2 is the different random effects structures in the two models; the different random effects structures produce different estimates of error that go into the denominator of the F-test, thus producing somewhat different results. The conclusions are the same (as we would hope): the three-factor interaction is highly significant, so we retain all terms. The argument `test="F"` requests KR tests, and results are type II by default.

Intensity is quantitative, and the interaction plot seemed to suggest linear in intensity might be an adequate explanation of the intensity effect (at least when results are on the log scale).

```
3 > fit7 <- lmer(log(y)~type*env*intensity+(0+int.z|type:crystal)+
    (1|env:type:crystal)+(0+int.z|env:type:crystal),data=QCM)
4 > fit7lin <- lmer(log(y)~type*env*int.z+(0+int.z|type:crystal)+
    (1|env:type:crystal)+(0+int.z|env:type:crystal),data=QCM)
5 > pbkrtest::KRmodcomp(fit7,fit7lin)
  F-test with Kenward-Roger approximation; computing time: 0.13 sec.
  large : log(y) ~ type * env * intensity + (0 + int.z | type:crystal) +
      (1 | env:type:crystal) + (0 + int.z | env:type:crystal)
  small : log(y) ~ type * env * int.z + (0 + int.z | type:crystal) + (1 |
      env:type:crystal) + (0 + int.z | env:type:crystal)
          stat    ndf    ddf F.scaling   p.value
  Ftest  3.3873 24.0000 48.0000        1 0.0001586 ***
```

Line 4 refits replacing intensity with linear in intensity in the fixed effects. This model uses 24 fewer degrees of freedom, and it is reasonable to ask whether this reduced model adequately describes the mean structure. The function `pbkrtest::KRmodcomp` lets you compare two different models provided that the random effects are the same and one of the fixed effects models is a submodel of the other. Line 5 uses that test, and we get a very

significant p-value. In fact, while the bulk of the variability is linear, there are significant higher order effects of intensity that need to be modeled.

Finally, line 6 illustrates that we can compute linear contrasts in the mixed-effects setting.

```
6 > linear.contrast(fit7,env,c(-1,1))
    estimates       se   t-value   p-value   lower-ci   upper-ci
1 -0.2384581 0.146064 -1.632559 0.1412052 -0.5752822 0.09836601
7 > car::Anova(fit7,test="F",type=3)
Analysis of Deviance Table (Type III Wald F tests with Kenward-Roger df)

Response: log(y)
                        F Df Df.res     Pr(>F)
(Intercept)      2834.7391  1  8.000 1.717e-11 ***
type              164.4238  1  8.000 1.291e-06 ***
env                 2.6653  1  8.000 0.1412052
intensity          31.0984  7 32.576 1.130e-12 ***
type:env           11.2268  1  8.000 0.0100707 *
type:intensity      7.1626  7 32.576 3.344e-05 ***
env:intensity       1.5142  7 32.576 0.1972937
type:env:intensity  6.3135  7 32.576 0.0001003 ***
```

Line 7 is there to remind us that contrasts are type III effects (although in this example the fixed effects are orthogonal so types II and III are the same).

11.2 Classical Analysis for Mixed Effects

We will only consider the classical approach in the situation where the data are balanced and the factors are treated as non-quantitative. In this situation, testing in the classical approach can be succinctly described as "Ignore the fact that you have mixed effects and do an ANOVA; then go back and modify the tests in the ANOVA to take into account the mixed effects." Thus the classical analysis is sometimes called the ANOVA approach. This works quite well in simple models, but rapidly becomes unwieldy. That said, results of the REML and classical approaches are often identical in simple, balanced situations.

Classical works well for balanced data

The primary reason for continuing to think about the classical approach is the insight it gives us into power and sample size selection for mixed effects designs. Unfortunately, we will need to cover quite a bit of the classical approach before we can talk sensibly about power. We will also see that the Hasse diagram is more than just a pretty picture.

Power

11.2.1 ANOVA and Expected Mean Squares

The analysis of variance for mixed effects is computed *exactly* the same as for fixed effects. The ANOVA table is mostly the same. It has rows for every term in the model and columns for source, sums of squares, degrees of freedom, mean squares, and F-statistics; the sources, sums of squares, degrees of freedom, and mean squares are just like for fixed effects. The

ANOVA table includes column for EMS

F-statistics often differ, and a mixed-effects ANOVA table often includes an additional column for expected mean squares (EMS). The EMS for a term is literally the expected value of its mean square.

The EMS for error is σ^2, exactly the same as in fixed effects. For balanced single-factor data with a random treatment factor, the EMS for treatments is $\sigma^2 + n\sigma_\alpha^2$.

To test the null hypothesis that $\sigma_\alpha^2 = 0$, we use the F-ratio $\text{MS}_{\text{Trt}}/\text{MS}_{\text{E}}$ and compare it to an F-distribution with $g-1$ and $N-g$ degrees of freedom to get a p-value. Let's start looking for the pattern now. To test the null hypothesis that $\sigma_\alpha^2 = 0$, we try to find two expected mean squares that would be the same if the null hypothesis were true and would differ otherwise. Put the mean square with the larger EMS in the numerator. If the null hypothesis is true, then the ratio of these mean squares should be about 1 (give or take some random variation). If the null hypothesis is false, then the ratio tends to be larger than 1, and we reject the null for large values of the ratio. In a one-factor ANOVA there are only two mean squares to choose from, and we use $\text{MS}_{\text{Trt}}/\text{MS}_{\text{E}}$ to test the null hypothesis of no treatment variation.

Construct tests by examining EMS

It's a bit puzzling at first that fixed- and random-effects models, which have such different assumptions about parameters, should have the same test for the standard null hypothesis. However, think about the effects when the null hypotheses are true. For fixed effects, the α_i are fixed and all zero; for random effects, the α_i are random and all zero. Either way, they're all zero. It is this commonality under the null hypothesis that makes the two tests the same.

Now consider a two-factor experiment with both factors random. The sources in a two-factor ANOVA are A, B, the AB interaction, and error; the following table gives the general two-factor skeleton ANOVA.

Two-factor EMS

Source	DF	EMS
A	$a-1$	$\sigma^2 + n\sigma_{\alpha\beta}^2 + nb\sigma_\alpha^2$
B	$b-1$	$\sigma^2 + n\sigma_{\alpha\beta}^2 + na\sigma_\beta^2$
AB	$(a-1)(b-1)$	$\sigma^2 + n\sigma_{\alpha\beta}^2$
Error	$N - ab = ab(n-1)$	σ^2

Suppose that we want to test the null hypothesis that $\sigma_{\alpha\beta}^2 = 0$. The EMS for the AB interaction is $\sigma^2 + n\sigma_{\alpha\beta}^2$, and the EMS for error is σ^2. These differ only by the variance component of interest, so we can test this null hypothesis using the ratio $\text{MS}_{\text{AB}}/\text{MS}_{\text{E}}$, with $(a-1)(b-1)$ and $ab(n-1)$ degrees of freedom.

That was pretty familiar; how about testing the null hypothesis that $\sigma_\alpha^2 = 0$? The only two lines that have EMS's that differ by a multiple of σ_α^2 are A and the AB interaction. Thus we use the F-ratio $\text{MS}_{\text{A}}/\text{MS}_{\text{AB}}$ with $a-1$ and $(a-1)(b-1)$ degrees of freedom to test $\sigma_\alpha^2 = 0$. Similarly, the test for $\sigma_\beta^2 = 0$ is $MS_B/\text{MS}_{\text{AB}}$ with $b-1$ and $(a-1)(b-1)$ degrees of freedom. Not having MS_{E} in the denominator is a major difference from fixed effects.

> The denominator mean square for F-tests in classical analysis for mixed-effects models will not always be MS_E!

Let's press on to three random factors. The sources in a three-factor ANOVA are A, B, and C; the AB, AC, BC, and ABC interactions; and error. The following table gives the generic expected mean squares:

Three-factor model

Source	EMS
A	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nc\sigma_{\alpha\beta}^2 + nb\sigma_{\alpha\gamma}^2 + nbc\sigma_\alpha^2$
B	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nc\sigma_{\alpha\beta}^2 + na\sigma_{\beta\gamma}^2 + nac\sigma_\beta^2$
C	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nb\sigma_{\alpha\gamma}^2 + na\sigma_{\beta\gamma}^2 + nab\sigma_\gamma^2$
AB	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nc\sigma_{\alpha\beta}^2$
AC	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nb\sigma_{\alpha\gamma}^2$
BC	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + na\sigma_{\beta\gamma}^2$
ABC	$\sigma^2 + n\sigma_{\alpha\beta\gamma}^2$
Error	σ^2

Testing for interactions is straightforward using our rule for finding two terms with EMS's that differ only by the variance component of interest. Thus error is the denominator for ABC, and ABC is the denominator for AB, AC, and BC. What do we do about main effects? Suppose we want to test the main effect of A, that is, test whether $\sigma_\alpha^2 = 0$. If we set σ_α^2 to 0 in the EMS for A, then we get $\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nc\sigma_{\alpha\beta}^2 + nb\sigma_{\alpha\gamma}^2$. A quick scan of the table of EMS's shows that *no* term has $\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nc\sigma_{\alpha\beta}^2 + nb\sigma_{\alpha\gamma}^2$ for its EMS. What we see is that there is no exact F-test for the null hypothesis that a main effect is zero in a three-way random-effects model. The lack of an exact F-test turns out to be relatively common in models with many random effects.

No exact F-tests for some hypotheses

In the absence of an exact F-test, we must form an approximate F-test. But before creating approximate tests, let's review a bit about exact F-tests.

An exact F-test is the ratio of two positive, independently distributed random quantities (mean squares). The denominator is distributed as a multiple τ_d of a chi-square random variable divided by its degrees of freedom (the denominator degrees of freedom), and the numerator is distributed as a multiple τ_n of a chi-square random variable divided by its degrees of freedom (the numerator degrees of freedom). The multipliers τ_d and τ_n are the expected mean squares; $\tau_n = \tau_d$ when the null hypothesis is true, and $\tau_n > \tau_d$ when the null hypothesis is false. Putting these together gives us a test statistic that has an F-distribution when the null hypothesis is true and tends to be bigger when the null is false.

Mean squares are multiples of chi-squares divided by their degrees of freedom

We want the approximate test to mimic the exact test as much as possible. The approximate F-test should be the ratio of two positive, independently distributed random quantities. When the null hypothesis is true, both quantities should have the same expected value. For exact tests, the numerator and denominator are each a single mean square. For approximate tests, the

Approximate tests mimic exact tests

numerator and denominator are sums of mean squares. Because the numerator and denominator should be independent, we need to use different mean squares for the two sums.

The key to the approximate test is to find sums for the numerator and denominator that have the same expectation when the null hypothesis is true. We can do this by inspection of the table of EMS, but we will shortly describe a simpler way using the Hasse diagram. One helpful comment: you always have the same number of mean squares in the numerator and denominator.

Example 11.4 Finding mean squares for an approximate test

Consider testing for no factor A effect (H_0: $\sigma_\alpha^2 = 0$) in a three-way model with all random factors. Refer to the table of expected mean squares given above.

1. The only mean square with an EMS that involves σ_α^2 is MS_A, so it must be in the numerator.

2. The EMS for A under the null hypothesis $\sigma_\alpha^2 = 0$ is $\sigma^2 + n\sigma_{\alpha\beta\gamma}^2 + nc\sigma_{\alpha\beta}^2 + nb\sigma_{\alpha\gamma}^2$.

3. We need to find a term or terms that will include $nc\sigma_{\alpha\beta}^2$ and $nb\sigma_{\alpha\gamma}^2$ without extraneous variance components that do not appear in the numerator. We can get $nc\sigma_{\alpha\beta}^2$ from MS_{AB}, and we can get $nb\sigma_{\alpha\gamma}^2$ from MS_{AC}. Our provisional denominator is now $\text{MS}_{AB} + \text{MS}_{AC}$; its expected value is $2\sigma^2 + 2n\sigma_{\alpha\beta\gamma}^2 + nc\sigma_{\alpha\beta}^2 + nb\sigma_{\alpha\gamma}^2$.

4. The denominator now has an expected value that is $\sigma^2 + n\sigma_{\alpha\beta\gamma}^2$ larger than that of the numerator. We can make them equal in expectation by adding MS_{ABC} to the numerator.

5. The numerator $\text{MS}_A + \text{MS}_{ABC}$ and denominator $\text{MS}_{AB} + \text{MS}_{AC}$ have the same expectations under the null hypothesis, and the numerator has a larger expectation when $\sigma_\alpha^2 > 0$. Their ratio forms our approximate F-test.

Now that we have the numerator and denominator, the test statistic is their ratio. To compute a p-value, we have to know the distribution of the ratio, and this is where the approximation comes in. We don't know the distribution of the ratio exactly; we approximate it. Exact F-tests follow the F-distribution, and we are going to compute p-values assuming that our approximate F-test also follows an F-distribution, even though it doesn't really. The degrees of freedom for our approximating F-distribution come from Satterthwaite formula (Satterthwaite 1946) shown below. These degrees of freedom will almost never be integers, but your software won't mind. If you only have a table, rounding the degrees of freedom down gives a conservative result.

Get approximate p-value using F-distribution

The simplest situation is when we have the sum of several mean squares, say MS_1, MS_2, and MS_3, with degrees of freedom ν_1, ν_2, and ν_3. The ap-

proximate degrees of freedom are calculated as

$$\nu^\star = \frac{(\text{MS}_1 + \text{MS}_2 + MS_3)^2}{\text{MS}_1^2/\nu_1 + \text{MS}_2^2/\nu_2 + MS_3^2/\nu_3} \ .$$

In more complicated situations, we may have a general linear combination of mean squares $\sum_k g_k \text{MS}_k$. This linear combination has approximate degrees of freedom

$$\nu^\star = \frac{(\sum_k g_k \text{MS}_k)^2}{\sum_k g_k^2 \text{MS}_k^2/\nu_k} \ .$$

Satterthwaite approximate degrees of freedom

Unbalanced data will lead to these more complicated forms. The approximation tends to work better when all the coefficients g_k are positive.

Example 11.5 Approximate degrees of freedom

Suppose that we obtain the following ANOVA table for an experiment with machine, operator, and glue as three crossed random factors (we called this carton experiment three in the last chapter, data not shown):

	DF	SS	MS	EMS
m	9	2706	300.7	$\sigma^2 + 2\sigma_{\alpha\beta\gamma}^2 + 4\sigma_{\alpha\beta}^2 + 20\sigma_{\alpha\gamma}^2 + 40\sigma_\alpha^2$
o	9	8887	987.5	$\sigma^2 + 2\sigma_{\alpha\beta\gamma}^2 + 4\sigma_{\alpha\beta}^2 + 20\sigma_{\beta\gamma}^2 + 40\sigma_\beta^2$
g	1	2376	2376	$\sigma^2 + 2\sigma_{\alpha\beta\gamma}^2 + 20\sigma_{\alpha\gamma}^2 + 20\sigma_{\beta\gamma}^2 + 200\sigma_\gamma^2$
m:o	81	1683	20.78	$\sigma^2 + 2\sigma_{\alpha\beta\gamma}^2 + 4\sigma_{\alpha\beta}^2$
m:g	9	420.4	46.71	$\sigma^2 + 2\sigma_{\alpha\beta\gamma}^2 + 20\sigma_{\alpha\gamma}^2$
o:g	9	145.3	16.14	$\sigma^2 + 2\sigma_{\alpha\beta\gamma}^2 + 20\sigma_{\beta\gamma}^2$
m:o:g	81	1650	20.37	$\sigma^2 + 2\sigma_{\alpha\beta\gamma}^2$
error	200	4646	23.23	σ^2

We illustrate approximate tests with a test for machine. We have already discovered that the numerator should be the sum of the mean squares for machine and the three-way interaction; these are 300.7 and 20.37 with 9 and 81 degrees of freedom. Our numerator is 321.07, and the approximate degrees of freedom are:

$$\nu_n^\star = \frac{321.07^2}{300.7^2/9 + 20.37^2/81} \approx 10.3 \ .$$

The denominator is the sum of the mean squares for the machine by operator and the machine by glue interactions; these are 20.78 and 46.71 with 81 and 9 degrees of freedom. The denominator is 67.49, and the approximate degrees of freedom are

$$\nu_d^\star = \frac{67.49^2}{20.78^2/81 + 46.71^2/9} \approx 18.4 \ .$$

> 1. The denominator for testing a term U is the leading eligible random term below U in the Hasse diagram.
>
> 2. An eligible random term V below U is leading if there is no eligible random term that is above V and below U.
>
> 3. If there are two or more leading eligible random terms, then we must use an approximate test.
>
> 4. In the unrestricted model, all random terms below U are eligible.
>
> 5. In the restricted model, all random terms below U are eligible except those that contain a fixed factor not found in U.
>
> **Display 11.1:** Rules for finding test denominators in balanced factorials using the Hasse diagram.

The F-test is $321.07/67.49 = 4.76$ with 10.3 and 18.4 approximate degrees of freedom and an approximate p-value of .0018; this is strong evidence against the null hypothesis of no machine to machine variation.

11.2.2 Hasse Diagrams, Test Denominators, and Expected Mean Squares

Using the Hasse diagram, we can determine the appropriate test denominator visually, without ever calculating the EMS for the design. We will briefly discuss denominators, and then move on to the main topic of EMS, which we need for power.

Recall that we considered unrestricted model assumptions, where all random effects were independent, and restricted model assumptions, where certain sums of the random effects were constrained to be 0. When we were discussing the REML method of analysis, we only considered the unrestricted assumptions. That was due to a limitation of `lme` and `lmer`; restricted assumptions can still appropriate in some circumstances, and the classical approach can handle the restricted assumptions.

Test denominators

Hasse diagrams look the same whether you use the restricted model or the unrestricted model, but the models are different and we must therefore use the Hasse diagram slightly differently for restricted and unrestricted models. Display 11.1 gives the steps for finding test denominators using the Hasse diagram. In general, you find the leading random term below the term to be tested, but only random terms without additional fixed factors are eligible in the restricted model. If there is more than one leading random term, we have an approximate test.

Finding test denominators
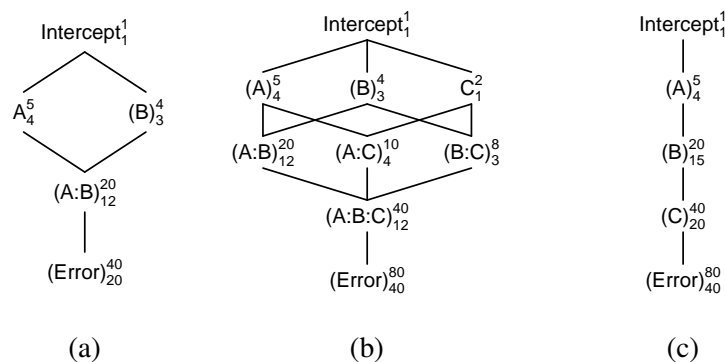
Figure 11.4: Hasse diagrams: (a) two-way factorial with A fixed and B random, A and B crossed; (b) three-way factorial with A and B random, C fixed, all factors crossed; (c) fully nested, with B fixed, A and C random. In all cases, A has 5 levels, B has 4 levels, and C has 2 levels.

Example 11.6 Test denominators in the restricted model

Consider the Hasse diagram in Figure 11.4(a). The next random term below A is the AB interaction. The only fixed factor in AB is A, so AB is the denominator for A. The next random term below B is also the AB interaction. However, AB contains A, an additional fixed factor not found in B, so AB is ineligible to be the denominator for B. Proceeding down, we get to error, which is random and does not contain any additional fixed factors. Therefore, error is the denominator for B. Similarly, error is the denominator for AB.

Figure 11.4(b) is a Hasse diagram for a three-way factorial with factors A and B random, and factor C fixed. The denominator for ABC is error. Immediately below AB is the random interaction ABC. However, ABC is not an eligible denominator for AB because it includes the additional fixed factor C. Therefore, the denominator for AB is error. For AC and BC, the denominator will be ABC, because it is random, immediately below, and contains no additional fixed factor. Next consider main effects. We see two random terms immediately below A, the AB and AC interactions. However, AC is not an eligible denominator for A, because it includes the additional fixed factor C. Therefore, the denominator for A is AB. Similarly, the denominator for B is AB. Finally consider C. There are two random terms immediately below C (AC and BC), and both of these are eligible to be denominators for C because neither includes an additional fixed factor. Thus we have an approximate test for C: C and ABC in the numerator, AC and BC in the denominator. In general, with two eligible random terms below a term of interest, the second mean square added to the mean square of interest will be the leading eligible random term below both of the denominator terms.

Figure 11.4(c) is a Hasse diagram for a three-factor, fully-nested model,

with A and C random and B fixed. Nesting structure appears as a vertical chain, with one factor below another. Note that the B nested in A *term* is a random term, even though B is a fixed factor. This seems odd, but consider that there is a different set of B effects for every level of A; we have a random set of A levels, so we must have a random set of B levels, so B nested in A is a random term. The denominator for C is E, and the denominator for B is C. The next random term below A is B, but B contains the fixed factor B not found in A, so B is not an eligible denominator. The closest eligible random term below A is C, which is the denominator for A.

When all the nested effects are random, the denominator for any term is simply the term below it. A fixed factor nested in a random factor is something of an oddity—it is a random term consisting only of a fixed factor. It will never be an eligible denominator in the restricted model.

Example 11.7 Test denominators in the unrestricted model

Figure 11.4(a) shows a two-factor mixed-effects design. Using the unrestricted model, error is the denominator for AB, and AB is the denominator for both A and B. This is a change from the restricted model, which had error as the denominator for B.

Using the unrestricted model in the three-way mixed effects design shown in Figure 11.4(b), we find that error is the denominator for ABC, and ABC is the denominator for AB, BC, and AC; error was the denominator for AB in the restricted model. All three main effects have approximate tests, because there are two leading eligible random two-factor interactions below every main effect.

In the three-way nested design shown in Figure 11.4(c), the denominator for every term is the term immediately below it. This is again different from the restricted model, which used C as the denominator for A.

Example 11.8 Classical tests for the quartz crystal microbalance experiment

The full ANOVA for the crystal data, ignoring the fact that some terms are random, is here:

```
1 > anova(lm(log(y)~(type/crystal)*env*intensity,data=QCM))
                             Df Sum Sq Mean Sq F value Pr(>F)
  type                       1 84.190  84.190
  env                        1  1.365   1.365
  intensity                  7 95.711  13.673
  type:crystal               4  2.545   0.636
  type:env                   1  5.748   5.748
  type:intensity             7 14.848   2.121
  env:intensity              7  0.190   0.027
  type:crystal:env           4  1.551   0.388
  type:crystal:intensity    28  2.344   0.084
  type:env:intensity         7  2.216   0.317
  type:crystal:env:intensity 28  0.636   0.023
  Residuals                  0  0.000
```

Refer back to the Hasse diagram in Figure 11.1. The classical tests are:

- Type: $MS_{Type}/MS_{Crystal} = 84.19/.636 = 132.37$ with 1 and 4 degrees of freedom.

- Environment: $MS_{Env}/MS_{Env:Crystal} = 1.365/.388 = 3.518$ with 1 and 4 degrees of freedom.

- Intensity: $MS_{Intensity}/MS_{Intensity:Crystal} = 13.673/.084 = 162.77$ with 7 and 28 degrees of freedom.

- Type by environment: $MS_{TE}/MS_{EC} = 5.748/.388 = 14.81$ with 1 and 4 degrees of freedom.

- Type by intensity: $MS_{TI}/MS_{IC} = 2.121/.084 = 25.25$ with 7 and 28 degrees of freedom.

- Environment by intensity: $MS_{EI}/MS_{EIC} = .027/.023 = 1.17$ with 7 and 28 degrees of freedom.

- Type by environment by intensity: $MS_{TEI}/MS_{EIC} = .317/.023 = 13.78$ with 7 and 28 degrees of freedom.

The Kenward-Roger tests for the fixed effects in the crystal data were computed before, but copied here for clarity:

```
2 > car::Anova(fit2,test="F")
                         F Df Df.res    Pr(>F)
  type             132.3113  1      4 0.0003261 ***
  env                3.5194  1      4 0.1338965
  intensity        163.2944  7     28 < 2.2e-16 ***
  type:env          14.8249  1      4 0.0182978 *
  type:intensity    25.3333  7     28 1.617e-10 ***
  env:intensity      1.1943  7     28 0.3380488
  type:env:intensity 13.9297 7     28 1.248e-07 ***
```

A quick glance shows that the classical and REML/KR results are within rounding error of being identical.

We can also test random effects. Note that while the fixed effects tests would be the same under restricted or unrestricted assumptions, the tests of

1. The representative element for a random term is its variance component.

2. The representative element for a fixed term is a function Q equal to the sum of the squared effects for the term divided by the degrees of freedom.

3. The contribution of a term is the number of data values N, divided by the number of effects for that term (the superscript for the term in the Hasse diagram), times the representative element for the term.

4. The expected mean square for a term U is the sum of the contributions for U and all eligible random terms below U in the Hasse diagram.

5. In the unrestricted model, all random terms below U are eligible.

6. In the restricted model, all random terms below U are eligible except those that contains a fixed factor not found in U.

Display 11.2: Rules for computing expected mean squares in balanced factorials using the Hasse diagram.

random effects will differ. REML used unrestricted assumptions, so we will as well.

- Intensity by crystal: $\text{MS}_{\text{IC}}/\text{MS}_{\text{IEC}} = .084/.023 = 3.65$ with 28 and 28 degrees of freedom and p-value = .0005.

- Environment by crystal: $\text{MS}_{\text{EC}}/\text{MS}_{\text{IEC}} = .388/.023 = 16.87$ with 4 and 28 degrees of freedom and p-value 3.8×10^{-7}.

- Crystal: $(MS_C + \text{MS}_{\text{IEC}})/(\text{MS}_{\text{IC}} + \text{MS}_{\text{EC}}) = (.636 + .023)/(.084 + .388) = 1.40$. We must use Satterthwaite to get degrees of freedom. For the numerator, $(.636 + .023)^2/(.636^2/4 + .023^2/28) = 4.3$; for the denominator, $(.084 + .388)^2/(.084^2/28 + .388^2/4) = 5.9$. Together, the p-value is .34.

Recall that the corresponding three p-values from the REML analysis were .0007, 2×10^{-16}, and .28. The REML and ANOVA p-values for the random effects are not precisely the same (as was true for fixed effects), but the differences are not great.

One side effect of using the unrestricted model is that there are more approximate tests, because there are more eligible denominators. The unrestricted model also tends to be slightly more conservative.

Expected mean squares

The rules for computing expected mean squares are given in Display 11.2. The description of the representative element for a fixed term seems a little arcane, but we have seen this Q before in expected mean squares. For a fixed main effect A, the representative element is $\sum_i \alpha_i^2/(a-1) = Q(\alpha)$. For a fixed interaction AB, the representative element is $\sum_{ij}(\alpha\beta_{ij})^2/[(a-1)(b-1)] = Q(\alpha\beta)$. These are the same forms we saw in earlier chapters when discussing noncentrality parameters and power.

Example 11.9 Expected mean squares in the restricted model

Consider the term A in Figure 11.4(b). In the restricted model, the eligible random terms below A are AB and E; AC and ABC are ineligible due to the inclusion of the additional fixed factor C. Thus the expected mean square for A is

$$\sigma^2 + \frac{80}{20}\sigma_{\alpha\beta}^2 + \frac{80}{5}\sigma_\alpha^2 = \sigma^2 + 4\sigma_{\alpha\beta}^2 + 16\sigma_\alpha^2 \ .$$

For term C in Figure 11.4(b), all random terms below C are eligible, so the EMS for C is

$$\sigma^2 + \frac{80}{40}\sigma_{\alpha\beta\gamma}^2 + \frac{80}{8}\sigma_{\beta\gamma}^2 + \frac{80}{10}\sigma_{\alpha\gamma}^2 + \frac{80}{2}Q(\gamma) =$$

$$\sigma^2 + 2\sigma_{\alpha\beta\gamma}^2 + 10\sigma_{\beta\gamma}^2 + 8\sigma_{\alpha\gamma}^2 + 40Q(\gamma) \ .$$

For term A in Figure 11.4(c), the eligible random terms are C and E; B is ineligible. Thus the expected mean square for A is

$$\sigma^2 + \frac{80}{40}\sigma_\gamma^2 + \frac{80}{5}\sigma_\alpha^2 = \sigma^2 + 2\sigma_\gamma^2 + 16\sigma_\alpha^2 \ .$$

Example 11.10 Expected mean squares in the unrestricted model

We now recompute two of the expected mean squares from Example 11.9 using the unrestricted model. There are four random terms below A in Figure 11.4(b); all of these are eligible in the unrestricted model, so the expected mean square for A is

$$\sigma^2 + \frac{80}{40}\sigma_{\alpha\beta\gamma}^2 + \frac{80}{20}\sigma_{\alpha\beta}^2 + \frac{80}{10}\sigma_{\alpha\gamma}^2 + \frac{80}{5}\sigma_\alpha^2 =$$

$$\sigma^2 + 2\sigma_{\alpha\beta\gamma}^2 + 4\sigma_{\alpha\beta}^2 + 8\sigma_{\alpha\gamma}^2 + 16\sigma_\alpha^2 \ .$$

This includes two additional contributions that were not present in the restricted model.
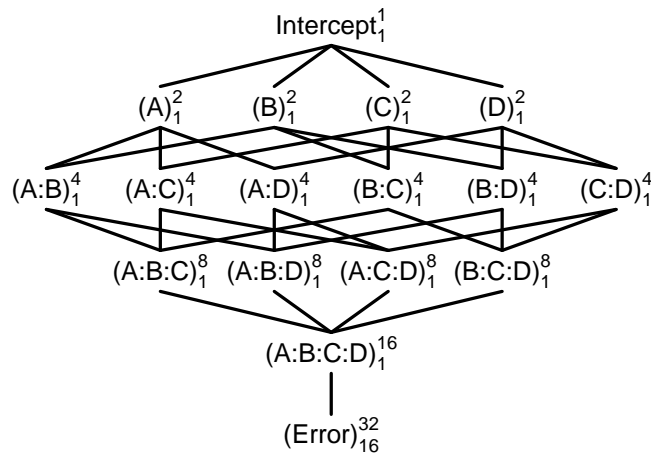
Figure 11.5: Hasse diagram for a four-way factorial with all random effects.

For term A in Figure 11.4(c), B, C, and E are all eligible random terms. Thus the expected mean square for A is

$$\sigma^2 + \frac{80}{40}\sigma_\gamma^2 + +\frac{80}{20}\sigma_\beta^2 + \frac{80}{5}\sigma_\alpha^2 = \sigma^2 + 2\sigma_\gamma^2 + 4\sigma_\beta^2 + 16\sigma_\alpha^2 \quad .$$

Term B contributes to the expected mean square of A in the unrestricted model.

We can figure out approximate tests by using the rules for expected mean squares and the Hasse diagram. Consider testing C in Figure 11.4(b). AC and BC are both eligible random terms below C, so both of their expected mean squares will appear in the EMS for C; thus both AC and BC need to be in the denominator for C. However, putting both AC and BC in the denominator double-counts the terms below AC and BC, namely ABC and error. Therefore, we add ABC to the numerator to match the double-counting.

Here is a more complicated example: testing a main effect in a four-factor model with all factors random. Figure 11.5 shows the Hasse diagram. Suppose that we wanted to test A. Terms AB, AC, and AD are all eligible random terms below A, so all would appear in the EMS for A, and all must appear in the denominator for A. If we put AB, AC, and AD in the denominator, then the expectations of ABC, ABD, and ACD will be double-counted there. Thus we must add them to the numerator to compensate. With A, ABC, ABD, and ACD in the numerator, ABCD and error are quadruple-counted in the numerator but only triple-counted in the denominator, so we must add ABCD to the denominator. We now have a numerator (A + ABC + ABD + ACD) and a denominator (AB + AC + AD + ABCD) with expectations that differ only by a multiple of σ_α^2.

Use Hasse diagrams to find approximate tests

Estimates of Variance Components

We can get point estimates for variance components by setting the observed mean squares for random terms equal to the formulae for their expected mean squares and solving the resulting linear equations for the unknown variance components. The resulting estimates of the variance components are unbiased (good), but they can be negative (embarrassing). In simple, balanced designs where all the estimates are positive, these ANOVA estimates of variance components often agree with REML estimates.

Example 11.11 ANOVA estimates of variance components for the quartz crystal data

Begin by equating the observed mean squares for random effects with their theoretical expectations:

$$
\begin{aligned}
.023 &= \sigma^2 + \sigma^2_{IEC} \\
.388 &= \sigma^2 + \sigma^2_{IEC} + 8\sigma^2_{EC} \\
.084 &= \sigma^2 + \sigma^2_{IEC} + 2\sigma^2_{IC} \\
.636 &= \sigma^2 + \sigma^2_{IEC} + 8\sigma^2_{EC} + 2\sigma^2_{IC} + 16\sigma^2_C
\end{aligned}
$$

This gives us

$$
\begin{aligned}
\widehat{\sigma}^2_{EC} &= (.388 - .023)/8 = .0456 \\
\widehat{\sigma}^2_{IC} &= (.084 - .023)/2 = .0305 \\
\widehat{\sigma}^2_C &= (.636 + .023 - .388 - .084)/16 = .0117
\end{aligned}
$$

In this balanced, relatively simple design, the ANOVA estimates agree with the REML estimates.

You can see a relationship between the formulae for variance component estimates and test numerators and denominators: mean squares in the test numerator are added in estimates, and mean squares in the test denominator are subtracted. Thus a variance component with an exact test will have an estimate that is just a difference of two mean squares.

Numerator MS's are added, denominator MS's are subtracted in estimates

11.2.3 Power

We are finally at a place where we can talk about power in mixed effects designs. Null hypotheses can be about either a fixed effect or a random effect. Let's begin with a fixed effect.

Consider a fixed effect for which there is an exact F-test as the ratio of two mean squares: $F = MS_1/MS_2$ with degrees of freedom ν_1 and ν_2. EMS_2 will be some linear combination of variance components with sum τ. EMS_1 will be τ plus some multiple of the Q contribution for the term being tested; that is, $EMS_1 = \tau + kQ$ for some multiplier k.

Noncentrality parameter

> The noncentrality parameter is $\nu_1 kQ/\tau$.

The measure of variability used in the denominator of the noncentrality parameter is the EMS of MS used in the denominator of the test; this will often not be σ^2.

Once we have the noncentrality parameter, degrees of freedom, and $\mathcal{E}$, we can compute power using software for the noncentral F distribution.

Example 11.12 Power for environment in the quartz crystal experiment

Consider an experiment with the structure of the quartz crystal experiment. There are a levels of intensity, b levels of environment, c levels of type, d crystals per type, and n measurements per crystal at each intensity-environment combination (in the actual crystal data, these values are 8, 2, 2, 3, and 1). With the unrestricted assumptions, the test for environment is $\text{MS}_{\text{Env}}/\text{MS}_{\text{Env:Cr}}$ with $b - 1$ and $(b - 1)c(d - 1)$ degrees of freedom. The EMS for environment is

$$\text{EMS}_{\text{Env}} = \sigma^2 + n\sigma_{IEC}^2 + na\sigma_{EC}^2 + nacdQ$$

and the noncentrality parameter is

$$\zeta = \frac{(b - 1)nacdQ}{\sigma^2 + n\sigma_{IEC}^2 + na\sigma_{EC}^2}$$

Suppose that we want power .95 under the following assumptions: $a = 8$, $b = 2$, $c = 2$, $\sigma^2 = .01$, $\sigma_{IEC}^2 = .02$, $\sigma_{EC}^2 = .05$, $Q = .01$, and $\mathcal{E} = .05$. In this case, the non-centrality parameter is

$$\zeta = \frac{(2 - 1)n \times 8 \times 2 \times d \times 1}{.01 + n \times .02 + 8n \times .05} = \frac{.16nd}{.01 + .02n + .4n} = \frac{.16nd}{.01 + .42n}$$

In Chapter 7, when we wanted to make the power arbitrarily close to 1, all we needed to make the noncentrality parameter arbitrarily large was to increase the sample size n. If we try increasing n here, ζ will never get any larger than

$$\zeta^\star = \frac{(b - 1)acdQ}{\sigma_{IEC}^2 + a\sigma_{EC}^2} = \frac{.16d}{.42}$$

Using $d = 3$ as in the real experiment, our largest possible non-centrality parameter is 1.14. Using this ζ, an F-test with 1 and 4 degrees of freedom and $\mathcal{E} = .05$ has power .13. That is not nearly big enough. Instead of sending n off to infinity, let's put n back to 1 and try increasing d (the number of crystals per type) instead. The non-centrality parameter is $.16d/.43 = .37d$, and the degrees of freedom will be 1 and $2(d - 1)$. Trying various values of d, we find that we need $d = 36$ to get power at least .95.

The function `cfcdae::mixed.power()` can automate much of this.

```
1 > mixed.power(~I*E*(T/C),c(8,2,2,3,1),
      list(E=.01,"I:E:C"=.02,Error=.01,"E:C"=.05),random="C",restrict=FALSE)
              num.ev den.ev num.df den.df power
  Intercept   0.43   0.43    1.0    4.0  0.05
  I                  0.03   0.03    7.0   28.0  0.05
  E                  0.91   0.43    1.0    4.0  0.13
  T                  0.43   0.43    1.0    4.0  0.05
  I:E                0.03   0.03    7.0   28.0  0.05
  C                  0.46   0.46    4.6    4.6  0.05
  I:T                0.03   0.03    7.0   28.0  0.05
  E:T                0.43   0.43    1.0    4.0  0.05
  I:C                0.03   0.03   28.0   28.0  0.05
  E:C                0.43   0.03    4.0   28.0  0.94
  I:E:T              0.03   0.03    7.0   28.0  0.05
  I:E:C              0.03   0.01   28.0    0.0   NaN
2 > mixed.power(~I*E*(T/C),c(8,2,2,3,1000),
      list(E=.01,"I:E:C"=.02,Error=.01,"E:C"=.05),random="C",restrict=FALSE)
              num.ev den.ev num.df  den.df power
  ...
  E           900.01 420.01    1.0     4.0  0.13
  ...
3 > mixed.power(~I*E*(T/C),c(8,2,2,35,1),
      list(E=.01,"I:E:C"=.02,Error=.01,"E:C"=.05),random="C",restrict=FALSE)
              num.ev den.ev num.df den.df power
  ...
  E             6.03   0.43    1.0   68.0  0.94
  ...
4 > mixed.power(~I*E*(T/C),c(8,2,2,36,1),
      list(E=.01,"I:E:C"=.02,Error=.01,"E:C"=.05),random="C",restrict=FALSE)
              num.ev den.ev num.df den.df power
  ...
  E             6.19   0.43    1.0   70.0  0.95
  ...
```

Line 1 shows the basic usage. The first argument is the right hand side of a formula, with the model expressed as you would assuming that everything is fixed. The second argument is number of levels for each factor. Take care that you enter the factors in the order that **R** enters them; `terms(~ formula)` can show you if you are unsure. The third argument is a list with named components for non-zero variance components or Q elements. Note that you need to quote the component names that contain a colon. The `random` argument is a (possibly NULL) character vector giving the names of random factors. By default this function uses the restricted model assumptions; you can select unrestricted by opting for `restrict=FALSE`. The default $\mathcal{E}$ is .05; you can change that with an `alpha=` argument.

The output of `mixed.power` gives the power for every term in the model except Error. The power will equal $\mathcal{E}$ unless you have set the representative element for that term to something greater than 0.

Line 1 shows the results when we use the number of levels in the actual crystal experiment. Power for environment is .13. Line 2 tries to increase power by increasing n to 1000; that does not work. In line 3 we have put n back to 1 and increased d to 35, and that is not quite enough. Line 4 shows we achieve our desired .95 power when we use $d = 36$.

> In mixed effects, you often need to increase the number of levels of a random factor in order to achieve the desired power.

If the term we are testing has an approximate test, you will need to compute the expected mean squares for all of the terms that are used, and you will need to compute approximate degrees of freedom for the numerator and denominator based on the expected mean squares instead of using the observed mean squares. Combining these, we compute power based on the noncentrality parameter $\nu_1 kQ/\tau$ where τ is the sum of the denominator expected mean squares and ν_1 is the approximate degrees of freedom for the numerator.

Power for random effects is similar to power for fixed effects, but it does not involve a noncentral F. Suppose that we wish to compute the power for testing the null hypothesis that $\sigma_\eta^2 = 0$, and that we have two mean squares with expectations $\text{EMS}_1 = \tau + k\sigma_\eta^2$ and $\text{EMS}_2 = \tau$ and degrees of freedom ν_1 and ν_2. The test for σ_η^2 is the F-ratio MS_1/MS_2.

Power for random effects uses central F

When the null hypothesis is true, the F-ratio has an F-distribution with ν_1 and ν_2 degrees of freedom. We reject the null when the observed F-statistic is greater than $F_{\mathcal{E},\nu_1,\nu_2}$. When the null hypothesis is false, the observed F-statistic is distributed as $(\tau + k\sigma_\eta^2)/\tau$ times a (central) F with ν_1 and ν_2 degrees of freedom. Thus the power is the probability that an F with ν_1 and ν_2 degrees of freedom exceeds $\tau/(\tau + k\sigma_\eta^2)F_{\mathcal{E},\nu_1,\nu_2}$.

When choosing the sample size to achieve a desired power for testing a random effect, you will again need to consider increasing the number of levels of one or more random effects; you cannot rely on increasing only n to get the power you need.

Example 11.13 Power for $\sigma_{Env:Crystal}^2$ in the quartz crystal experiment

Consider an experiment with the structure of the quartz crystal experiment. There are a levels of intensity, b levels of environment, c levels of type, d crystals per type, and n measurements per crystal at each intensity-environment combination (in the actual crystal data, these values are 8, 2, 2, 3, and 1). With the unrestricted assumptions, the test for the environment by crystal variance component is $\text{MS}_{\text{Env:Cr}}/\text{MS}_{\text{Int:Env:Cr}}$ with $(b-1)c(d-1)$ and $(a-1)(b-1)c(d-1)$ degrees of freedom. The EMS for environment by crystal is

$$\text{EMS}_{\text{Env:Cry}} = \sigma^2 + n\sigma_{IEC}^2 + na\sigma_{EC}^2$$

and the EMS for intensity by environment by crystal is

$$\text{EMS}_{\text{Int:Env:Cry}} = \sigma^2 + n\sigma_{IEC}^2$$

Suppose that we want power .95 under the following assumptions: $a = 8$, $b = 2$, $c = 2$, $\sigma^2 = .01$, $\sigma_{IEC}^2 = .02$, $\sigma_{EC}^2 = .005$, and $\mathcal{E} = .05$. The factor is

$$\frac{\text{EMS}_{EC}}{\text{EMS}_{IEC}} = \frac{.01 + n \times .02 + n \times 8 \times .005}{.01 + n \times .02} = \frac{.01 + .06n}{.01 + .02n}$$

The largest the multiplier can be is 3 (at infinite n). With a multiplier of 3 and $d = 3$ as in the real data set, the power is .47, which is much too small.

The other thing we can do is increase d; this does not change the multiplier, but it does change the degrees of freedom in the F distribution, and that will eventually be enough to achieve our power. With $n = 1$, the multiplier is 7/3 and the minimum d to get power .95 is $d = 19$; for $n = 2$, the multiplier is 13/5 and the minimum is $d = 15$; for $n = 3$ we need $d = 14$; and so on. We can also see this using `mixed.power`

```
5 > mixed.power(~I*E*(T/C),c(8,2,2,3,1),
ist("I:E:C"=.02,Error=.01,"E:C"=.005),random="C",restrict=FALSE)
          num.ev den.ev num.df den.df power
 ...
 E:C        0.07   0.03      4     28  0.35
 ...
5 > mixed.power(~I*E*(T/C),c(8,2,2,3,1000),
    list("I:E:C"=.02,Error=.01,"E:C"=.005),random="C",restrict=FALSE)
          num.ev den.ev num.df den.df power
 ...
 E:C       60.01  20.01      4     28  0.47
 ...
5 > mixed.power(~I*E*(T/C),c(8,2,2,19,1),
    list("I:E:C"=.02,Error=.01,"E:C"=.005),random="C",restrict=FALSE)
          num.ev den.ev num.df den.df power
 ...
 E:C        0.07   0.03   36.0  252.0  0.95
 ...
5 > mixed.power(~I*E*(T/C),c(8,2,2,15,2),
    list("I:E:C"=.02,Error=.01,"E:C"=.005),random="C",restrict=FALSE)
          num.ev den.ev num.df den.df power
 ...
 E:C        0.13   0.05   28.0  196.0  0.95
 ...
```

11.2.4 Variances of Means and Contrasts

REML software will compute the variance/covariance matrix of the estimated coefficients; from this you can compute the estimated variance of a mean or contrast from the data. This is what we need most of the time, but if you want to choose your sample size to get a confidence interval of a certain length, you need to be able to compute the variances and covariances of means without having any data. We can use the Hasse diagram to get these variances.

Treatment means make sense for combinations of fixed factors, but are generally less interesting for random effects. Consider the Hasse diagrams in Figure 11.6. All are three-way factorials with $a = 3$, $b = 4$, $c = 5$, and $n = 2$. In panels (a) and (c), factors A and B are fixed. Thus it makes sense to consider means for levels of factor A ($\overline{y}_{i\bullet\bullet\bullet}$), for levels of factor B ($\overline{y}_{\bullet j\bullet\bullet}$), and for AB combinations ($\overline{y}_{ij\bullet\bullet}$). In panel (b), only factor A is fixed, so only means $\overline{y}_{i\bullet\bullet\bullet}$ are usually of interest.

Look at treatment means for fixed factors

It is tempting to use the denominator mean square for A as the variance

1. Make a Hasse diagram for the model.

2. Identify the base term and base factors for the mean of interest.

3. The variance of the mean of interest will be the sum over all contributing terms T of

$$\sigma_T^2 \frac{\text{product of superscripts of all base factors above T}}{\text{superscript of term T}}$$

4. In the unrestricted model, all random terms contribute to the variance of the mean of interest.

5. In the restricted model, all random terms contribute to the variance of the mean of interest except those that contain a fixed factor not found in the base term.

Display 11.3: Steps for determining the variance of a marginal mean.

(a)   (b)   (c)
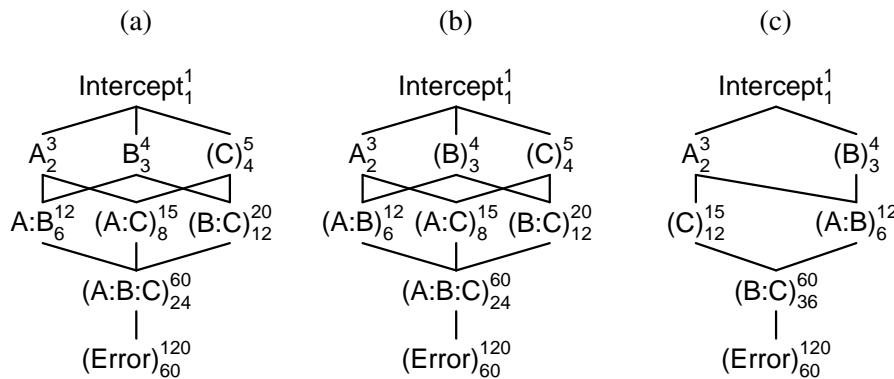
Figure 11.6: Hasse diagrams for three three-way factorials. (a) C random; (b) B and C random; (c) C random and nested in A.

for means $\bar{y}_{i\bullet\bullet\bullet}$. *This does not work!* We must go through the steps given in Display 11.3 to compute variances for means. We can use the denominator mean square for A when computing the variance for a *contrast* in factor A means; simply substitute the denominator mean square as an estimate of variance into the usual formula for the variance of a contrast. Similarly, we can use the denominator mean square for the AB interaction when we compute the variance of an AB interaction contrast, but this will not work for means $\bar{y}_{ij\bullet\bullet}$ or paired differences or other combinations that are not interaction contrasts.

Do not use denominator mean squares as variances for means

Display 11.3 gives the steps required to compute the variance of a mean. For a mean $\overline{y}_{i\bullet\bullet\bullet}$, the base term is A and the base factor is A; for a mean $\overline{y}_{ij\bullet\bullet}$, the base term is AB and the base factors are A and B.

Example 11.14 Variances of means

Let's compute variances for some means in the models of Figure 11.6 using restricted model assumptions. Consider first the mean $\overline{y}_{i\bullet\bullet\bullet}$. The base term is A, and the base factor is A. In panel (a), there will be contributions from C, AC, and E (but not BC or ABC because they contain the additional fixed factor B). The variance is

$$\sigma_\gamma^2 \frac{1}{5} + \sigma_{\alpha\gamma}^2 \frac{3}{15} + \sigma^2 \frac{3}{120} \quad .$$

In panel (b), there will be contributions from all random terms (A is the only fixed term). Thus the variance is

$$\sigma_\beta^2 \frac{1}{4} + \sigma_\gamma^2 \frac{1}{5} + \sigma_{\alpha\beta}^2 \frac{3}{12} + \sigma_{\alpha\gamma}^2 \frac{3}{15} + \sigma_{\beta\gamma}^2 \frac{1}{20} + \sigma_{\alpha\beta\gamma}^2 \frac{3}{60} + \sigma^2 \frac{3}{120} \quad .$$

Finally, in panel (c), there will be contributions from C and E (but not BC). The variance is

$$\sigma_\gamma^2 \frac{3}{15} + \sigma^2 \frac{3}{120} \quad .$$

Now consider a mean $\overline{y}_{\bullet j\bullet\bullet}$ in model (c). The contributing terms will be C, BC, and E, and the variance is

$$\sigma_\gamma^2 \frac{1}{15} + \sigma_{\beta\gamma}^2 \frac{4}{60} + \sigma^2 \frac{4}{120} \quad .$$

Finally, consider the variance of $\overline{y}_{ij\bullet\bullet}$; this mean does not make sense in panel (b). In panel (a), all random terms contribute to the variance, which is

$$\sigma_\gamma^2 \frac{1}{5} + \sigma_{\alpha\gamma}^2 \frac{3}{15} + \sigma_{\beta\gamma}^2 \frac{4}{20} + \sigma_{\alpha\beta\gamma}^2 \frac{3 \times 4}{60} + \sigma^2 \frac{3 \times 4}{120} \quad .$$

In panel (c), all random terms contribute, but the variance here is

$$\sigma_\gamma^2 \frac{3}{15} + \sigma_{\beta\gamma}^2 \frac{3 \times 4}{60} + \sigma^2 \frac{3 \times 4}{120} \quad .$$

The variance of a difference is the sum of the individual variances minus twice the covariance. We thus need to compute covariances of means in order to get variances of differences of means. Display 11.4 gives the steps for computing the covariance between two means, which are similar to those for variances, with the additional twist that we need to know which of the subscripts in the means agree and which disagree. For example, the factor A subscripts in $\overline{y}_{i\bullet\bullet\bullet} - \overline{y}_{i'\bullet\bullet\bullet}$ disagree, but in $\overline{y}_{ij\bullet\bullet} - \overline{y}_{ij'\bullet\bullet}, j \neq j'$, the factor A subscripts agree while the factor B subscripts disagree.

Need covariances to get variance of a difference

1. Identify the base term and base factors for the means of interest.

2. Determine whether the subscripts agree or disagree for each base factor.

3. The covariance of the means will be the sum over all contributing terms T of

$$\sigma_T^2 \frac{\text{product of superscripts of all base factors above T}}{\text{superscript of term T}}$$

4. In the unrestricted model, all random terms contribute to the covariance *except* those that are below a base factor with disagreeing subscripts.

5. In the restricted model, all random terms contribute to the covariance *except* those that contain a fixed factor not found in the base term and those that are below a base factor with disagreeing subscripts.

Display 11.4: Steps for determining the covariance between two marginal means.

Example 11.15 Covariances of means

Now compute covariances for some means in the models of Figure 11.6 using restricted model assumptions. Consider the means $\bar{y}_{i\bullet\bullet\bullet}$ and $\bar{y}_{i'\bullet\bullet\bullet}$. The base term is A, the base factor is A, and the factor A subscripts disagree. In model (a), only term C contributes to the covariance, which is

$$\sigma_\gamma^2 \frac{1}{5}$$

Using the variance for $\bar{y}_{i\bullet\bullet\bullet}$ computed in Example 11.14, we find

$$
\begin{aligned}
\text{Var}(\bar{y}_{i\bullet\bullet\bullet} - \bar{y}_{i'\bullet\bullet\bullet}) &= \text{Var}(\bar{y}_{i\bullet\bullet\bullet}) + \text{Var}(\bar{y}_{i'\bullet\bullet\bullet}) - 2 \times \text{Cov}(\bar{y}_{i\bullet\bullet\bullet}, \bar{y}_{i'\bullet\bullet\bullet}) \\
&= 2 \times (\sigma_\gamma^2 \frac{1}{5} + \sigma_{\alpha\gamma}^2 \frac{1}{5} + \sigma^2 \frac{1}{40}) - 2 \times \sigma_\gamma^2 \frac{1}{5} \\
&= 2 \times (\sigma_{\alpha\gamma}^2 \frac{1}{5} + \sigma^2 \frac{1}{40}) \\
&= \text{EMS}_{\text{AC}}(\frac{1}{40} + \frac{1}{40}) \ .
\end{aligned}
$$

The last line is what we would get by using the denominator for A and applying the usual contrast formulae with a sample size of 40 in each mean.

In model (b), B, C, and BC contribute to the covariance, which is

$$\sigma_\beta^2 \frac{1}{4} + \sigma_\gamma^2 \frac{1}{5} + \sigma_{\beta\gamma}^2 \frac{1}{20}$$

and leads to

$$
\begin{aligned}
\mathrm{Var}(\bar{y}_{i\bullet\bullet\bullet} - \bar{y}_{i'\bullet\bullet\bullet}) &= \mathrm{Var}(\bar{y}_{i\bullet\bullet\bullet}) + \mathrm{Var}(\bar{y}_{i'\bullet\bullet\bullet}) - 2 \times \mathrm{Cov}(\bar{y}_{i\bullet\bullet\bullet}, \bar{y}_{i'\bullet\bullet\bullet}) \\
&= 2 \times (\sigma_{\alpha\beta}^2 \frac{1}{4} + \sigma_{\alpha\gamma}^2 \frac{1}{5} + \sigma_{\alpha\beta\gamma}^2 \frac{1}{20} + \sigma^2 \frac{1}{40})
\end{aligned}
$$

In panel (c), all the random terms are below A, so none can contribute to the covariance, which is thus 0.

Consider now $\bar{y}_{\bullet j\bullet\bullet} - \bar{y}_{\bullet j'\bullet\bullet}$ in model (c). Only the term C contributes to the covariance, which is

$$
\sigma_\gamma^2 \frac{1}{15} \quad ;
$$

and leads to

$$
\begin{aligned}
\mathrm{Var}(\bar{y}_{\bullet j\bullet\bullet} - \bar{y}_{\bullet j'\bullet\bullet}) &= \mathrm{Var}(\bar{y}_{\bullet j\bullet\bullet}) + \mathrm{Var}(\bar{y}_{\bullet j'\bullet\bullet}) - 2 \times \mathrm{Cov}(\bar{y}_{\bullet j\bullet\bullet}, \bar{y}_{\bullet j'\bullet\bullet}) \\
&= 2 \times (\sigma_{\beta\gamma}^2 \frac{1}{15} + \sigma^2 \frac{1}{30}) \\
&= \frac{2}{30} \mathrm{EMS}_{\mathrm{BC}} \quad ;
\end{aligned}
$$

which is what would be obtained by using the denominator for B in the standard contrast formulae for means with sample size 30.

Things get a little more interesting with two-factor means, because we can have the first, the second, or both subscripts disagreeing, and we can get different covariances for each. Of course there are even more possibilities with three-factor means. Consider covariances for AB means in panel (a) of Figure 11.6. If the A subscripts differ, then only C and BC can contribute to the covariance; if the B subscripts differ, then C and AC contribute to the covariance; if both differ, then only C contributes to the covariance. In panel (c), if the A subscripts differ, then no terms contribute to covariance; if the B subscripts differ, then only C contributes to covariance. Table 11.2 summarizes the covariances and variances of differences of means for these cases.

11.3   Bayesian Analysis of Mixed Effects

Bayesians already assume that every unknown is a random variable, so there is little difference between fixed and random effects. When using `bglmm`, the principal difference is that fixed effects are forced to sum to zero, and random effects are not. The `bglmm` function also has the option to impose the zero sum restricted assumptions for all mixed effects.

Example 11.16 Quartz Crystal Microbalance, continued.

We begin in line 1 by doing a Bayesian fit model 2 using unrestricted model assumptions.

Table 11.2: Covariances and variances of differences of two-factor means $\overline{y}_{ij\bullet\bullet}$ for models (a) and (c) of Figure 11.6 as a function of which subscripts disagree.

		Covariance	Variance of difference
(a)	A	$\frac{1}{5}\sigma_\gamma^2 + \frac{1}{5}\sigma_{\beta\gamma}^2$	$2 \times (\frac{1}{5}\sigma_{\alpha\gamma}^2 + \frac{1}{5}\sigma_{\alpha\beta\gamma}^2 + \frac{1}{10}\sigma^2)$
(a)	B	$\frac{1}{5}\sigma_\gamma^2 + \frac{1}{5}\sigma_{\alpha\gamma}^2$	$2 \times (\frac{1}{5}\sigma_{\beta\gamma}^2 + \frac{1}{5}\sigma_{\alpha\beta\gamma}^2 + \frac{1}{10}\sigma^2)$
(a)	A and B	$\frac{1}{5}\sigma_\gamma^2$	$2 \times (\frac{1}{5}\sigma_{\alpha\gamma}^2 + \frac{1}{5}\sigma_{\beta\gamma}^2 + \frac{1}{5}\sigma_{\alpha\beta\gamma}^2 + \frac{1}{10}\sigma^2)$
(c)	A	0	$2 \times (\frac{1}{5}\sigma_\gamma^2 + \frac{1}{5}\sigma_{\beta\gamma}^2 + \frac{1}{10}\sigma^2)$
(c)	B	$\frac{1}{5}\sigma_\gamma^2$	$2 \times (\frac{1}{5}\sigma_{\beta\gamma}^2 + \frac{1}{10}\sigma^2)$
(c)	A and B	0	$2 \times (\frac{1}{5}\sigma_\gamma^2 + \frac{1}{5}\sigma_{\beta\gamma}^2 + \frac{1}{10}\sigma^2)$

```
1 > fit2bayes <- bglmm(log(y)~type*env*intensity+(1|type:crystal)+
    (1|intensity:type:crystal)+(1|env:type:crystal),data=QCM)
2 > summary(fit2bayes)[,c(1,3,4,8:10)]
                                mean      sd     2.5%    97.5% n_eff  Rhat
  (Intercept)                 3.89000 0.1840  3.52000  4.27000  1620 1.000
  type1                      -0.89600 0.2170 -1.24000 -0.42200   833 1.000
  env1                        0.11000 0.0890 -0.06320  0.29000  2060 1.000
  intensity1                 -1.58000 0.0829 -1.74000 -1.41000  3130 1.000
  intensity2                 -1.03000 0.0843 -1.19000 -0.85700  3150 1.000
  ...
  type1:env1                 -0.23500 0.0921 -0.40700 -0.03820  2100 1.000
  type1:intensity1           -0.63000 0.0818 -0.79200 -0.46800  3030 1.000
  type1:intensity2           -0.43500 0.0816 -0.59500 -0.27200  2640 1.000
  ...
  env1:intensity1             0.00960 0.0278 -0.04370  0.06910  4000 1.000
  env1:intensity2            -0.00471 0.0281 -0.06500  0.05150  4000 1.000
  ...
  type1:env1:intensity1      -0.30800 0.0423 -0.39100 -0.22300  4000 1.000
  type1:env1:intensity2      -0.06380 0.0408 -0.14300  0.01650  4000 1.000
  ...
  sigma0                      0.15700 0.0213  0.12300  0.20600   850 1.000
  ...
  sigma.type:crystal          0.30600 0.3140  0.03170  1.08000   518 1.010
  sigma.intensity:type:crystal 0.18200 0.0394  0.11300  0.26500   693 1.000
  sigma.env:type:crystal      0.28300 0.1180  0.13400  0.57800  1230 1.000
```

Line 2 shows an extract of the summary of the results. Comparing the estimates and standard errors for the fixed effects here and via REML in Example 11.1 we see that the estimates are generally very similar, but the standard errors for the intercept and the main effect of type are substantially larger in the Bayesian fit. The reason for this can be seen by examining the estimated random effects, where we see that the Bayesian approach has estimated a much larger crystal to crystal variance.

The `bglmm` function can also fit with restricted model assumptions for mixed terms, and we do that in line 3.

```
3 > fit2bayesrest <- bglmm(log(y)~type*env*intensity+(1|type:crystal)+
    (1|intensity:type:crystal)+(1|env:type:crystal),data=QCM,
    adapt_delta = .99,restrictmixed = TRUE)
4 > summary(fit2bayesrest)[,c(1,3,4,8:10)]
                                mean      sd     2.5%    97.5%  n_eff   Rhat
(Intercept)                  3.890000 0.0307  3.83000  3.9500   4000  1.000
type1                       -0.880000 0.4250 -1.59000  0.0396    243  1.010
env1                         0.117000 0.0299  0.05950  0.1770   4000  1.000
intensity1                  -1.580000 0.0794 -1.74000 -1.4300   4000  1.000
intensity2                  -1.030000 0.0767 -1.18000 -0.8770   4000  1.000
...
type1:env1                  -0.213000 0.2060 -0.57600  0.2230    327  1.010
type1:intensity1            -0.627000 0.0851 -0.79300 -0.4600   4000  0.999
type1:intensity2            -0.437000 0.0849 -0.60100 -0.2690   4000  1.000
...
env1:intensity1              0.004810 0.0397 -0.07700  0.0954   3550  0.999
env1:intensity2             -0.001170 0.0410 -0.08410  0.0893   4000  1.000
...
type1:env1:intensity1       -0.254000 0.0853 -0.42000 -0.0847   2890  1.000
type1:env1:intensity2       -0.053200 0.0686 -0.19000  0.0804   4000  1.000
...
sigma0                       0.291000 0.0271  0.24300  0.3490   2390  1.000
...
sigma.type:crystal           0.841000 0.9400  0.13400  3.3800    588  1.010
sigma.intensity:type:crystal 0.081600 0.0534  0.01090  0.2040    102  1.030
sigma.env:type:crystal       0.558000 0.6620  0.04820  2.4100    607  1.010
5 > bayes_factor(fit2bayes,fit2bayesrest)
The estimated Bayes factor in favor of x1 over x2 is equal to: 2058268
```

Line 4 gives an extract of the summary of the fit. Comparing this with the output of line 2 shows several differences. First, three of the four variances are estimated to be substantially larger in the restricted model. In fact, the 95% posterior intervals for σ_0 do not even overlap. The fitted fixed effects are very similar in these two Bayesian fits as well, but the standard errors are once again very different.

I would have guessed that the intensity:type:crystal and env:type:crystal interactions would, in a sense, be functions of the individual crystal, thus making the restricted assumptions appropriate. However, line 5 computes the Bayes factor for these two models, and the unrestricted model is overwhelmingly preferred.

11.4 Further Reading and Extensions

We have only scratched the surface of the subject of random effects. Searle (1971) provides a review, and Searle, Casella, and McCulloch (1992) provide book-length coverage.

In the single-factor situation, there is a simple formula for the EMS for treatments when the data are unbalanced: $\sigma^2 + n'\sigma_\alpha^2$, where

$$n' = \frac{1}{a-1}[N - \frac{1}{N}\sum_{i=1}^{a} n_i^2] \ .$$

The formula for n' reduces to n for balanced data.

Expected mean squares do not depend on normality, though the chi-square distribution for mean square and F-distribution for test statistics do depend on normality. Tukey (1956) and Tukey (1957b) work out variances for variance components, though the notation and algebra are rather heavy going.

The Satterthwaite formula is based on matching the mean and variance of an unknown distribution to that of an approximating distribution. There are quite a few other possibilities; Johnson and Kotz (1970) describe the major ones.

11.5   Problems

We wish to examine the average daily weight gain by calves sired by four bulls selected at random from a population of bulls. Bulls denoted A through D were mated with randomly selected cows. Average daily weight gain by the calves is given below (data set `Sires`). **Exercise 11.1**

A	B	C	D
1.46	1.17	.98	.95
1.23	1.08	1.06	1.10
1.12	1.20	1.15	1.07
1.23	1.08	1.11	1.11
1.02	1.01	.83	.89
1.15	.86	.86	1.12

a) Test the null hypothesis that there is no sire to sire variability in the response.

b) Find 90% interval estimates for the error variance and the sire to sire variance.

A 24-head machine fills bottles with vegetable oil. Five of the heads are chosen at random, and several consecutive bottles from these heads were taken from the line. The net weight of oil in these bottles is given in the following table (data set `Bottles`, originally from Swallow and Searle 1978): **Exercise 11.2**

		Head		
1	2	3	4	5
15.70	15.69	15.75	15.68	15.65
15.68	15.71	15.82	15.66	15.60
15.64		15.75	15.59	
15.60		15.71		
		15.84		

Is there any evidence for head to head variability? Estimate the head to head and error variabilities (both point and 99% interval estimates).

The burrowing mayfly *Hexagenia* can be used as an indicator of water **Exercise 11.3**

quality (it likes clean water). Before starting a monitoring program using *Hexagenia* we take three samples from each of ten randomly chosen locations along the upper Mississippi between Lake Peppin and the St. Anthony Lock and Dam. We use these data to estimate the within location and between location variability in *Hexagenia* abundance. An ANOVA follows; the data are in hundreds of insects per square meter.

	DF	SS	MS
Location	9	11.59	1.288
Error	20	1.842	0.0921

Give a point estimate for the between location variance in *Hexagenia* abundance.

Exercise 11.4

We are operating a bioreactor. In our experiment, we take three random runs of the reactor. From each run of the reactor we take three random samples and measure the product. Our interest focusses on the variability between runs and the variability between samples within run. Here is an ANOVA from the data.

	DF	SS	MS	F	P-value
Run	2	1.4283e+05	71415	3.00148	0.12491
Error	6	1.4276e+05	23793		

Estimate the run to run variance.

Exercise 11.5

I am curious about the role of the First Year Experience course (required of all freshmen) on student retention in our college. The 2450 incoming freshmen self select into 100 groups (half with 24 students and half with 25 students). The 100 sections are divided into 4 groups of 25 at random. These four groups are assigned to the factor level combinations of medium (online versus face to face) and freedom (student choice about which units to do or no student choice). Two years later, when students would be entering their third year of college, we determine how many of the 2450 students have returned for their third year.

How many error degrees of freedom does this design have? Justify your answer.

Exercise 11.6

Anecdotal evidence suggests that some individuals can tolerate alcohol better than others. As part of a traffic safety study, you are planning an experiment to test for the presence of individual to individual variation. Volunteers will be recruited who have given their informed consent for participation after having been informed of the risks of the study. Each individual will participate in two sessions one week apart. In each session, the individual will arrive not having eaten for at least 4 hours. They will take a hand-eye coordination test, drink 12 ounces of beer, wait 15 minutes, and then take a second hand-eye coordination test. The score for a session is the change in hand-eye coordination. There are two sessions, so $n = 2$. We believe that the individual to individual variation σ_α^2 will be about the same size as the error

σ^2. If we are testing at the 1% level, how many individuals should be tested to have power .9 for this setup?

Five tire types (brand/model combinations like Goodyear/Arriva) in the size 175/80R-13 are chosen at random from those available in a metropolitan area, and six tires of each type are taken at random from warehouses. The tires are placed (in random order) on a machine that will test tread durability and report a response in thousands of miles. The data follow (data set `Tires`):

Problem 11.1

Brand			Miles			
1	55	56	59	55	60	57
2	39	42	43	41	41	42
3	39	41	43	40	43	43
4	44	44	42	39	40	43
5	46	42	45	42	42	44

Compute a 99% interval estimate for the ratio of type to type variability to tire within type variability (σ_α^2/σ^2). Do you believe that this interval actually has 99% coverage? Explain.

Milk is tested after Pasteurization to assure that Pasteurization was effective. This experiment was conducted to determine variability in test results between laboratories, and to determine if the interlaboratory differences depend on the concentration of bacteria.

Problem 11.2

Five contract laboratories are selected at random from those available in a large metropolitan area. Four levels of contamination are chosen at random by choosing four samples of milk from a collection of samples at various stages of spoilage. A batch of fresh milk from a dairy was obtained and split into 40 units. These 40 units are assigned at random to the twenty combinations of laboratory and contamination sample. Each unit is contaminated with 5 ml from its selected sample, marked with a numeric code, and sent to the selected laboratory. The laboratories count the bacteria in each sample by serial dilution plate counts without knowing that they received four pairs, rather than eight separate samples. Data follow (colony forming units per μl, data set `Interlaboratory`):

	Sample			
Lab	1	2	3	4
1	2200	3000	210	270
	2200	2900	200	260
2	2600	3600	290	360
	2500	3500	240	380
3	1900	2500	160	230
	2100	2200	200	230
4	2600	2800	330	350
	4300	1800	340	290
5	4000	4800	370	500
	3900	4800	340	480

Analyze these data to determine if the effects of interest are present. If so, estimate them.

Problem 11.3

Composite materials used in the manufacture of aircraft components must be tested to determine tensile strength. A manufacturer tests five random specimens from each of five randomly selected batches, obtaining the following coded strengths (data set `Tensile`, originally from Vangel 1992).

Batch	Strength				
1	379	357	390	376	376
2	363	367	382	381	359
3	401	402	407	402	396
4	402	387	392	395	394
5	415	405	396	390	395

Compute point and interval estimates for the between batch and within batch variance components. If using Bayes methods, compute a 95% interval estimate for σ_α^2/σ^2.

Problem 11.4

Briefly describe the treatment structure you would choose for each of the following situations. Describe the factors, the number of levels for each, whether they are fixed or random, and which are crossed.

(a) One of the expenses in animal experiments is feeding the animals. A company salesperson has made the claim that their new rat chow (35% less expensive) is equivalent to the two standard chows on the market. You wish to test this claim by measuring weight gain of rat pups on the three chows. You have a population of 30 inbred, basically exchangeable female rat pups to work with, each with her own cage.

(b) Different gallons of premixed house paints with the same label color do not always turn out the same. A manufacturer of paint believes that color variability is due to three sources: supplier of tint materials, miscalibration of the devices that add the tint to the base paint, and uncontrollable random variation between gallon cans. The manufacturer wishes to assess the sizes of these sources of variation and is willing to

use 60 gallons of paint in the process. There are three suppliers of tint and 100 tint-mixing machines at the plant.

(c) Insect infestations in croplands are not uniform; that is, the number of insects present in meter-square plots can vary considerably. Our interest is in determining the variability at different geographic scales. That is, how much do insect counts vary from meter square to meter square within a hectare field, from hectare to hectare within a county, and from county to county? We have resources for at most 10 counties in southwestern Minnesota, and at most 100 total meter-square insect counts.

(d) The disposable diaper business is very competitive, with all manufacturers trying to get a leg up, as it were. You are a consumer testing agency comparing the absorbency of two brands of "newborn" size diapers. The test is to put a diaper on a female doll and pump body-temperature water through the doll into the diaper at a fixed rate until the diaper leaks. The response is the amount of liquid pumped before leakage. We are primarily interested in brand differences, but we are also interested in variability between individual diapers and between batches of diapers (which we can only measure as between boxes of diapers, since we do not know the actual manufacturing time or place of the diapers). We can afford to buy 32 boxes of diapers and test 64 diapers.

Dental fillings made with gold can vary in hardness depending on how the metal is treated prior to its placement in the tooth. Two factors are thought to influence the hardness: the gold alloy and the condensation method. In addition, some dentists doing the work are better at some types of fillings than others.

Problem 11.5

Five dentists were selected at random. Each dentist prepares 24 fillings (in random order), one for each of the combinations of method (three levels) and alloy (eight levels). The fillings were then measured for hardness using the Diamond Pyramid Hardness Number (big scores are better). The data follow (data set `Fillings`, originally from Xhonga 1971 via Brown 1975):

Dentist	Method	Alloy							
		1	2	3	4	5	6	7	8
1	1	792	824	813	792	792	907	792	835
	2	772	772	782	698	665	1115	835	870
	3	782	803	752	620	835	847	560	585
2	1	803	803	715	803	813	858	907	882
	2	752	772	772	782	743	933	792	824
	3	715	707	835	715	673	698	734	681
3	1	715	724	743	627	752	858	762	724
	2	792	715	813	743	613	824	847	782
	3	762	606	743	681	743	715	824	681
4	1	673	946	792	743	762	894	792	649
	2	657	743	690	882	772	813	870	858
	3	690	245	493	707	289	715	813	312
5	1	634	715	707	698	715	772	1048	870
	2	649	724	803	665	752	824	933	835
	3	724	627	421	483	405	536	405	312

Analyze these data to determine which factors influence the response and how they influence the response. (Hint: the dentist by method interaction can use close inspection.)

Problem 11.6

Eight 1-gallon containers of raw milk are obtained from a dairy and are assigned at random to four abuse treatments, two containers per treatment. Abuse consists of keeping the milk at 25°C for a period of time; the four abuse treatments are four randomly selected durations between 1 and 18 hours. After abuse, each gallon is split into five equal portions and frozen.

We have selected five contract laboratories at random from those available in the state. For each gallon, the five portions are randomly assigned to the five laboratories. The eight portions for a given laboratory are then placed in an insulated shipping container cooled with dry ice and shipped. Each laboratory is asked to provide duplicate counts of bacteria in each milk portion. Data follow (bacteria counts per μl, data set `AbusedMilk`).

	Abuse/Gallon							
	1		2		3		4	
Lab	1	2	3	4	5	6	7	8
1	7800	7000	870	490	1300	1000	31000	36000
	7500	7200	690	530	1200	980	35000	34000
2	8300	9700	900	930	2500	2300	27000	28000
	8200	10000	940	840	1900	2300	34000	32000
3	7300	7300	760	840	2100	2300	34000	34000
	7600	7900	790	780	2000	2200	34000	33000
4	5400	5500	520	750	1400	1100	16000	16000
	5700	5600	770	620	1300	1400	16000	15000
5	15000	12000	1200	800	4600	3500	41000	39000
	14000	12000	1100	600	4000	3600	40000	39000

Analyze these data. The main issues are the sources and sizes of variation, with an eye toward reliability of future measurements.

Problem 11.7

Cheese is made by bacterial fermentation of Pasteurized milk. Most of the bacteria are purposefully added to do the fermentation; these are the starter cultures. Some "wild" bacteria are also present in cheese; these are the nonstarter bacteria. One hypothesis is that nonstarter bacteria may affect the quality of a cheese, so that otherwise identical cheese making facilities produce different cheeses due to their different indigenous nonstarter bacteria.

Two strains of nonstarter bacteria were isolated at a premium cheese facility: R50#10 and R21#2. We will add these nonstarter bacteria to cheese to see if they affect quality. Our four treatments will be control, addition of R50, addition of R21, and addition of a blend of R50 and R21. Twelve cheeses are made, three for each of the four treatments, with the treatments being randomized to the cheeses. Each cheese is then divided into four portions, and the four portions for each cheese are randomly assigned to one of four aging times: 1 day, 28 days, 56 days, and 84 days. Each portion is measured for total free amino acids (a measure of bacterial activity) after it has aged for its specified number of days (data set `Nonstarters`, originally from Peggy Swearingen).

		Days			
Treatment	Cheese	1	28	56	84
Control	1	.637	1.250	1.697	2.892
	2	.549	.794	1.601	2.922
	3	.604	.871	1.830	3.198
R50	1	.678	1.062	2.032	2.567
	2	.736	.817	2.017	3.000
	3	.659	.968	2.409	3.022
R21	1	.607	1.228	2.211	3.705
	2	.661	.944	1.673	2.905
	3	.755	.924	1.973	2.478
R50+R21	1	.643	1.100	2.091	3.757
	2	.581	1.245	2.255	3.891
	3	.754	.968	2.987	3.322

We are particularly interested in the bacterial treatment effects and interactions, and less interested in the main effect of time.

Problem 11.8

As part of a larger experiment, researchers are looking at the amount of beer that remains in the mouth after expectoration. Ten subjects will repeat the experiment on two separate days. Each subject will place 10 ml or 20 ml of beer in his or her mouth for five seconds, and then expectorate the beer. The beer has a dye, so the amount of expectorated beer can be determined, and thus the amount of beer retained in the mouth (in ml, data set `Beer`, originally from Bréfort, Guinard, and Lewis 1989)

	10 ml		20 ml	
Subject	Day 1	Day 2	Day 1	Day 2
1	1.86	2.18	2.49	3.75
2	2.08	2.19	3.15	2.67
3	1.76	1.68	1.76	2.57
4	2.02	3.87	2.99	4.51
5	2.60	1.85	3.25	2.42
6	2.26	2.71	2.86	3.60
7	2.03	2.63	2.37	4.12
8	2.39	2.58	2.19	2.84
9	2.40	1.91	3.25	2.52
10	1.63	2.43	2.00	2.70

Compute interval estimates for the amount of beer retained in the mouth for both volumes.

Problem 11.9

One of the steps in a molecular biological analysis is the quantification of DNA. This can be done by measuring the absorbance of ultra-violet light at 260 nm (called the optical density). The absorption of light in the spectrophotometer should be proportional to concentration of DNA, but should not depend on the volume of sample used. However, there is a general belief that small samples, say less than 40 μl, lead to erroneous results. In addition, the proportionality cited above only holds over a range of concentrations; outside

that range nonlinear effects come into play. In theory, $OD_{260} = 0.02 \times C$, where C is the concentration in ng/μl.

This experiment measures the optical density at five concentrations (10, 30, 60, 120, 480 ng/μl) and six volumes (15, 20, 30, 40, 50, 100 μl). Three analysts are chosen at random from the 16 in the lab. Each of the analysts prepares two samples at each volume/concentration combination and measures the optical density.

The data in the table below are the optical densities $\times 100$; data set DNA, originally from S. Charaniya.

		Concentration								
Vol	User	10		30		60		120		480	
15	1	36	33	97	90	165	161	412	397	1040	1090
	2	34	30	85	80	171	156	365	376	930	1010
	3	31	28	87	82	165	159	318	341	1010	940
20	1	31	30	78	78	148	156	370	369	960	920
	2	25	31	92	97	175	167	321	332	980	892
	3	33	29	74	69	145	137	329	345	980	878
30	1	20	17	72	69	142	141	357	365	834	780
	2	23	26	76	71	160	146	305	298	824	846
	3	25	27	77	71	141	127	341	309	1050	965
40	1	16	14	66	67	137	132	347	357	730	882
	2	21	18	66	81	147	133	286	312	791	813
	3	23	23	75	64	136	126	322	296	868	794
50	1	28	27	76	80	138	138	350	347	785	833
	2	24	23	70	68	135	139	315	281	748	773
	3	17	21	62	67	146	132	294	286	734	692
100	1	19	19	67	68	127	130	336	340	642	605
	2	20	22	72	71	129	133	282	285	730	742
	3	20	21	62	59	128	126	285	265	715	718

Analyze these data. Which factors are important; does volume matter? What is the range of proportionality?

Problem 11.10

Consider a two-factor factorial design; factor A has a levels, factor B has b levels, and there are n units for each factor level combinations. Both factors are random. We want to make the power for testing A to be very high. Should we increase n or increase a or increase b? Justify your answer.

Problem 11.11

There is interest in whether gender differences exist in spatial reasoning and whether these differences are influenced by stress. From the students of a large psychology class, twenty women are selected at random, and twenty men are selected at random. All 40 subjects will be given two spatial reasoning tests in random order. One of the tests has a fixed time limit (the stress condition) and the other test is untimed (the no stress condition).

a) Draw a Hasse diagram for this experiment.
b) Determine the appropriate denominators for gender, stress, and the gender by stress interaction.

Why do you always wind up with the same number of numerator and denominator terms in approximate tests? **Question 11.1**

Derive the Satterthwaite approximate degrees of freedom for a sum of mean squares by matching the first two moments of the sum of mean squares to a multiple of a chi-square. **Question 11.2**

Consider a three-factor model with A and B fixed and C random. Show that the variance for the difference $\overline{y}_{ij\bullet} - \overline{y}_{i'j\bullet} - \overline{y}_{ij'\bullet} + \overline{y}_{i'j'\bullet}$ can be computed using the usual formula for contrast variance with the "denominator" expected mean square as error variance. **Question 11.3**

Chapter 12

Complete Block Designs

We now begin the study of *variance reduction design*. Experimental error makes inference difficult. As the variance of experimental error (σ^2) increases, confidence intervals get longer and test power decreases. All other things being equal, we would thus prefer to conduct our experiments with units that are homogeneous so that σ^2 will be small. Unfortunately, all other things are rarely equal. For example, there may be few units available, and we must simply take what we can get. Or we might be able to find homogeneous units, but using the homogeneous units would restrict our inference to a subset of the population of interest. Variance reduction designs can give us many of the benefits of small σ^2, without necessarily restricting us to a subset of the population of units.

12.1   Blocking

Variance reduction design deals almost exclusively with a technique called *blocking*. A *block* of units is a set of units that are homogeneous in some sense. Perhaps they are field plots located in the same general area, or are samples analyzed at about the same time, or are units that came from a single supplier. These similarities in the units themselves lead us to anticipate that units within a block may also have similar responses. So when constructing blocks, we try to achieve homogeneity of the units within blocks, but units in different blocks may be dissimilar.

Blocking designs are not completely randomized designs. The Randomized Complete Block design described in the next section is the first design we study that uses some kind of restricted randomization. When we design an experiment, we know the design we choose to use and thus the randomization that is used. When we look at an experiment designed by someone else, we can determine the design from the way the randomization was done, that is, from the kinds of restrictions that were placed on the randomization; we cannot determine the design based on the actual outcome of which units got which treatments.

There are many, many blocking designs, and we will only cover some of the more widely used designs. This chapter deals with *complete block designs* in which every treatment is used in every block; later chapters deal with *incomplete block designs* (not every treatment is used in every block) and some special block designs for treatments with factorial structure.

Complete blocks include every treatment

12.2   The Randomized Complete Block Design

The Randomized Complete Block design (RCB) is the basic blocking design. There are g treatments, and each treatment will be assigned to r units for a total of $N = gr$ units. We partition the N units into r groups of g units each; these r groups are our blocks. We make this partition into blocks in such a way that the units within a block are somehow alike; we anticipate that these alike units will have similar responses. In the first block, we randomly assign the g treatments to the g units; we do an independent randomization, assigning treatments to units in each of the other blocks. In effect, this is r single-replication completely randomized designs glued together. This is the RCB design.

RCB has r blocks of g units each

Block for homogeneity

Blocks exist at the time of the randomization of treatments to units. We cannot impose blocking structure on a completely randomized design after the fact; either the randomization was blocked or it was not.

Example 12.1 Mealybugs on cycads

Modern zoos try to reproduce natural habitats in their exhibits as much as possible. They therefore use appropriate plants, but these plants can be infested with inappropriate insects. Zoos need to take great care with pesticides, because the variety of species in a zoo makes it more likely that a sensitive species is present.

Cycads (plants that look vaguely like palms) can be infested with mealybug, and the zoo wishes to test three treatments: water (a control), horticultural oil (a standard no-mammalian-toxicity pesticide), and fungal spores in water (*Beauveria bassiana*, a fungus that grows exclusively on insects). Five infested cycads are removed to a testing area. Three branches are randomly chosen on each cycad, and two 3 cm by 3 cm patches are marked on each branch; the number of mealybugs in these patches is noted. The three branches on each cycad are randomly assigned to the three treatments. After three days, the patches are counted again, and the response is the change in the number of mealybugs (before − after). Data for this experiment are given in Table 12.1 (data from Scott Smith, data set `MealyBugs`).

How can we decode the experimental design from the description just given? *Follow the randomization!* Looking at the randomization, we see that the treatments were applied to the branches (or pairs of patches). Thus the branches (or pairs) must be experimental units. Furthermore, the randomization was done so that each treatment was applied once on each cycad. There was no possibility of two branches from the same plant receiving the same

Table 12.1: Changes in mealybug counts on cycads after treatment. Treatments are water, *Beauveria bassiana* spores, and horticultural oil.

	Plant				
	1	2	3	4	5
Water	-9	18	10	9	-6
	-6	5	9	0	13
Spores	-4	29	4	-2	11
	7	10	-1	6	-1
Oil	4	29	14	14	7
	11	36	16	18	15

treatment. This is a restriction on the randomization, with cycads acting as blocks. The patches are measurement units. When we analyze these data, we can take the average or sum of the two patches on each branch as the response for the branch. (Alternatively, we can use the data at the measurement unit level but add a random effect for experimental unit, using the final error in the model to represent measurement error.) To recap, there were $g = 3$ treatments applied to $N = 15$ units arranged in $r = 5$ blocks of size 3 according to an RCB design; there were two measurement units per experimental unit.

Why did the experimenter block? Experience and intuition lead the experimenter to believe that branches on the same cycad will tend to be more alike than branches on different cycads—genetically, environmentally, and perhaps in other ways. Thus blocking by plant may be advantageous.

It is important to realize that tables like Table 12.1 hide the randomization that has occurred. The table makes it appear as though the first unit in every block received the water treatment, the second unit the spores, and so on. This is not true. The table ignores the randomization for the convenience of a readable display. The water treatment may have been applied to any of the three units in the block, chosen at random.

You cannot determine the design used in an experiment just by looking at a table of results, you have to know the randomization. There may be many different designs that could produce the same data, and you will not know the correct analysis for those data without knowing the design. *Follow the randomization to determine the design.*

Follow the randomization to determine design

An important feature to note about the RCB is that we have placed no restrictions on the treatments. The treatments could simply be g treatments, or they could be the factor-level combinations of two or more factors. These factors could be fixed or random, crossed or nested. All of these treatment structures can be incorporated when we use blocking designs to achieve variance reduction.

General treatment structure

Example 12.2 Protein/amino acid effects on growing rats

Male albino laboratory rats (Sprague-Dawley strain) are used routinely

in many kinds of experiments. Proper nutrition for the rats is important. This experiment was conducted to determine the requirements for protein and the amino acid threonine. Specifically, this experiment will examine the factorial combinations of the amount of protein in diet and the amount of threonine in diet. The general protein in the diet is threonine deficient. There are eight levels of threonine (.2 through .9% of diet) and five levels of protein (8.68, 12, 15, 18, and 21% of diet), for a total of 40 treatments.

Two-hundred weanling rats were acclimated to cages. On the second day after arrival, all rats were weighed, and the rats were separated into five groups of 40 to provide groupings of approximately uniform weight. The 40 rats in each group were randomly assigned to the 40 treatments. Body weight and food consumption were measured twice weekly, and the response we consider is average daily weight gain over 21 days.

This is a randomized complete block design. Initial body weight is a good predictor of body weight in 3 weeks, so the rats were blocked by initial weight in an attempt to find homogeneous groups of units. There are 40 treatments, which have an eight by five factorial structure.

12.2.1 Why and when to use the RCB

We use an RCB to increase the power and precision of an experiment by decreasing the error variance. This decrease in error variance is achieved by finding groups of units that are homogeneous (blocks) and, in effect, repeating the experiment independently in the different blocks. The RCB is an effective design when there is a single source of extraneous variation in the responses that we can identify ahead of time and use to partition the units into blocks. Blocking is done at the time of randomization; you can't construct blocks after the experiment has been run.

Block when you can identify a source of variation

There is an almost infinite number of ways in which units can be grouped into blocks, but a few examples may suffice to get the ideas across. We would like to group into blocks on the basis of homogeneity of the responses, but that is not possible. Instead, we must group into blocks on the basis of other similarities that we think may be associated with responses.

Some blocking is fairly obvious. For example, you need milk to make cheese, and you get a new milk supply every day. Each batch of milk makes slightly different cheese. If your batches are such that you can make several types of cheese per batch, then blocking on batch of raw material is a natural.

Block on batch

Units may be grouped spatially. For example, some units may be located in one city, and other units in a second city. Or, some units may be in cages on the top shelf, and others in cages on the bottom shelf. It is common for units close in space to have more similar responses, so spatial blocking is also common.

Block spatially

Units may be grouped temporally. That is, some units may be treated or measured at one time, and other units at another time. For example, you may only be able to make four measurements a day, and the instrument may need

Block temporally

to be recalibrated every day. As with spatial grouping, units close in time may tend to have similar responses, so temporal blocking is common.

Age and gender blocking are common for animal subjects. Sometimes units have a "history." The number of previous pregnancies could be a blocking factor. In general, any source of variation that you think may influence the response and which can be identified prior to the experiment is a candidate for blocking.

Age, gender, and history blocks

12.2.2 The Generalized Randomized Complete Block

In some situations we may find that our blocks have more than g units. For example, we might have $2g$ or $3g$ units per block. In such a situation, we can create a Generalized Randomized Complete Block (GRCB) design by assigning each of the g treatments to two (or three, etc.) units per block, completely randomized within block. The mealy bug data of Table 12.1 are actually a GRCB.

If these large blocks are as homogeneous as we can make them, the GRCB is an effective design. On the other hand, if we can break the $2g$ units into two groups of g units that are still more homogeneous, then we would have more power by breaking up the large blocks into blocks of size g.

12.2.3 Analysis for the RCB

Now all the hard work in the earlier chapters studying analysis methods pays off. The design of an RCB is new, but there is nothing new in the analysis of an RCB. Once we have the correct model, we do point estimates, confidence intervals, multiple comparisons, testing, residual analysis, and so on in the same way as we have been doing.

Nothing new in analysis of RCB

Let y_{ij} be the response for the ith treatment in the jth block. The standard model for an RCB has a grand mean, a treatment effect, a block effect, and experimental error, as in

Blocks usually assumed additive

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \ .$$

This standard model says that treatments and blocks are additive, so that treatments have the same effect in every block, and blocks only serve to shift the mean response up or down.

Inference is done exactly as for a two-way factorial. In this the *model* we are using to analyze an RCB is just the same as a two-way factorial with replication $n = 1$, even though the *design* of an RCB is not the same. Blocks may be assumed to be fixed or random. With complete data, the inferential results for the treatments will be the same either way.

One difference between an RCB and a factorial is that we do not try to make inferences about blocks, even though the machinery of our model allows us to do so. The reason for this goes back to thinking of F-tests as

Do not test blocks—they were not randomized

approximations to randomization tests. Under the RCB randomization, units are assigned at random to treatments, but units always stay in the same block. Thus the block effects and sums of squares are not random, and there is no test for blocks; blocks simply exist. More pragmatically, we blocked because we believed that the units within blocks were more similar, so finding a block effect is not a major revelation.

Example 12.3 Mealybugs, continued

For a first analysis we take as our response the mean of the two measurements for each branch from Table 12.1.

```
1 > bybranch <- aggregate(change~plant+treatment,MealyBugs,mean)
2 > fit.fixed <- lm(change~plant+treatment,bybranch)
3 > summary(fit.fixed)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.867      1.087   8.157  3.8e-05 ***
plant1        -8.367      2.174  -3.848 0.004889 **
plant2        12.300      2.174   5.658 0.000477 ***
plant3        -0.200      2.174  -0.092 0.928966
plant4        -1.367      2.174  -0.629 0.547125
treatment1     7.533      1.537   4.900 0.001193 **
treatment2    -2.967      1.537  -1.930 0.089752 .
...
4 > anova(fit.fixed)
Analysis of Variance Table

Response: change
          Df Sum Sq Mean Sq F value   Pr(>F)
plant      4 686.40 171.600  9.6812 0.003708 **
treatment  2 432.03 216.017 12.1871 0.003729 **
Residuals  8 141.80  17.725

5 > pairwise(fit.fixed,treatment)

Pairwise comparisons ( hsd ) of treatment
               estimate signif diff     lower     upper
* Oil - Spores     10.5    7.608532  2.891468 18.108532
* Oil - Water      12.1    7.608532  4.491468 19.708532
  Spores - Water    1.6    7.608532 -6.008532  9.208532
```

Line 1 summarizes the data by the mean for each plant by treatment combination. Line 2 fits the additive model, and line 3 summarizes the fit. Lines 4 and 5 give an ANOVA for the data and make pairwise comparisons between the treatments. Note that **R** produces a p-value for plant (blocks). We know that this is an RCB, so we should ignore that test. Treatments are significant at the .0037 level, and the pairwise comparison results show that oil is better than water or spores, which cannot be distinguished from each other.

If we thought that these five plants represented a random sample of all cycads (or of those available to us), we could treat plant as a random effect.

Table 12.2: Average insect catch by three different traps over five periods. Data from Snedecor and Cochran (1967), data set `InsectCatch`.

			Period		
Trap	1	2	3	4	5
1	19.1	23.4	29.5	23.4	16.6
2	50.1	166.1	223.9	58.9	64.6
3	123.0	407.4	398.1	229.1	251.2

```
6 > fit.random <- lmer(change~treatment+(1|plant),bybranch)
7 > car::Anova(fit.random,test="F")
  Response: change
               F Df Df.res   Pr(>F)
  treatment 12.187  2      8 0.003729 **
8 > summary(fit.random)
  ...
  Random effects:
   Groups   Name        Variance Std.Dev.
   plant    (Intercept) 51.29    7.162
   Residual             17.72    4.210
  Number of obs: 15, groups:  plant, 5

  Fixed effects:
             Estimate Std. Error t value
  (Intercept)   8.867      3.382   2.621
  treatment1    7.533      1.537   4.900
  treatment2   -2.967      1.537  -1.93
```

Line 6 fits the mixed model with plant random; lines 7 and 8 provide sample post-fit information. We see that in this complete data situation, the results from assuming that block is random are identical to those with block as fixed.

Additivity is an assumption. In the real world, treatments could have different effects in different blocks. However, we cannot distinguish between random error and interaction in a single block/treatment combination of the RCB, because the RCB has only one observation for each treatment in each block. (That is, if you fit an interaction you will have zero degrees of freedom for estimating error.) A systematic pattern in the residuals versus fitted plot can indicate the presence of interaction. In some cases a transformation of the response can reduce the interaction. Reexpressing the data on the appropriate scale can make the data more additive. When the data are more additive, the term that we use as error contains less interaction and is a better surrogate for error. In other cases, a model for interaction such as the Tukey one degree of freedom model (Section 9.4.3) or the row-model of Section 9.6 might be better.

Are the data additive?

Transform for additivity

Example 12.4 Insect catch

Table 12.2 shows data for the average number of insects (macrolepi-

doptera) caught by three kinds of trap during five periods. The periods are not randomized and serve as blocks. Begin with a standard analysis of an RCB.

```
1 > fit1 <- lm(catch~period+trap,data=InsectCatch)
2 > anova(fit1)
    Response: catch
            Df Sum Sq Mean Sq F value    Pr(>F)
 period      4  52066   13016  3.4022 0.0661095 .
 trap        2 173333   86667 22.6528 0.0005073 ***
 Residuals   8  30607    3826
3 > plot(fit1,which=1)
```

The ANOVA seems to indicate a statistically significant trap effect, but the residual plot (line 4, shown in Figure 12.1 (a)) shows a distinctive pattern. In this case, the residuals are above 0 on the ends of the range and below zero in the middle. Although less distinct, the spread of the residuals is also greater on the left than on the right. I call this the "flopping fish." It can indicate an un-modeled interaction or a need for transformation of the response.

```
4 > car::boxCox(fit1)
5 > fit2 <- lm(-(catch^-.5)~period+trap,data=InsectCatch)
6 > plot(fit2,which=1)
7 > anova(fit2)
Response: -(catch^-0.5)
          Df    Sum Sq   Mean Sq  F value     Pr(>F)
period     4 0.005981 0.0014953   6.6082    0.01186 *
trap       2 0.059878 0.0299388 132.3083 7.416e-07 ***
Residuals  8 0.001810 0.0002263
```

Line 4 does a Box-Cox transformation; the plot (Figure 12.1 (b)) indicates that a reciprocal square root should help. Line 5 refits the model with transformed data, and the residual plot (line 6 and Figure 12.1 (c)) shows that the situation is muchly improved, although far from what we would hope to see. Line 7 shows that trap is considerably more significant after transformation.

The story is that in the first model the "error" term included interaction in addition to error. In the second model, the data were transformed so that what we use as error does not include transformable non-additivity.

```
8 > 30607/(30607+173333+52066)
1] 0.1195558
9 > .00181/(.00181+.05988+.00598)
1] 0.02674745
```

Line 8 shows that the error term was 12% of the total sum of squares in the first model, but line 9 shows that it is only 2.7% of total sum of squares after transformation.

For the generalized RCB, we have multiple experimental units for each treatment in each block. That allows us to estimate block by treatment interaction separately from pure error. The usual model for the GRCB is to assume that blocks are random and that there is a (random) block by treatment interaction.

Model for GRCB

(a)

(b)

(c)
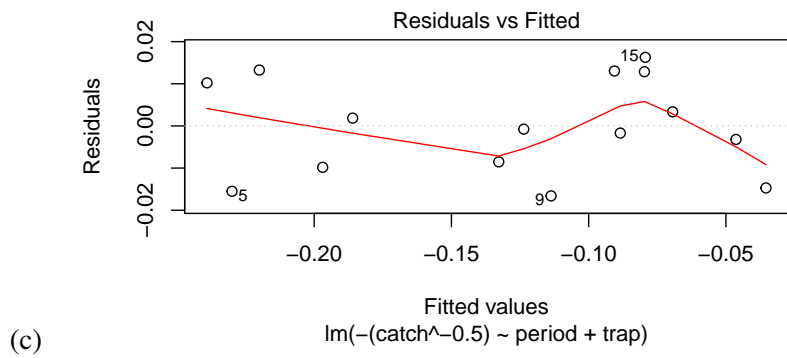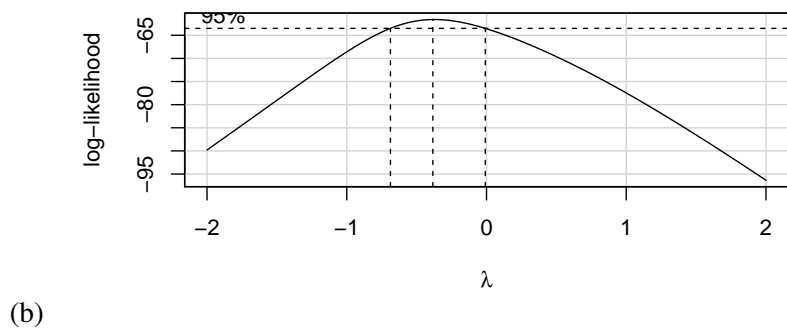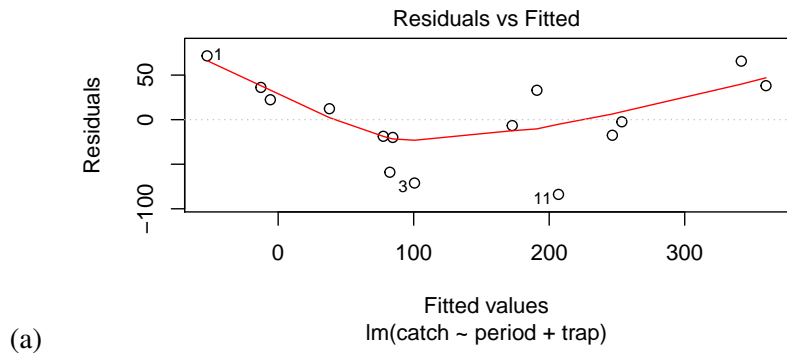
Figure 12.1: Plots for insect catch data. (a) Residuals versus predicted values; (b) Box-Cox transformation plot; (c) residuals versus predicted on the transformed scale.

Example 12.5 Mealybugs, continued

The mealy bug experiment is a GRCB, so we can use a random effect for each individual branch to model the plant by treatment interaction.

```
9 > fit.full <- lmer(change~treatment+(1|plant)+(1|branch),MealyBugs)
 boundary (singular) fit: see ?isSingular
0 > summary(fit.full)
 ...
 Random effects:
  Groups    Name         Variance Std.Dev.
  branch    (Intercept)  0.00     0.000
  plant     (Intercept) 49.64     7.046
  Residual              45.33     6.733
 Number of obs: 30, groups:  branch, 15; plant, 5

 Fixed effects:
             Estimate Std. Error t value
 (Intercept)    8.867      3.382   2.621
 treatment1     7.533      1.738   4.333
 treatment2    -2.967      1.738  -1.707
 ...

1 > pairwise(fit.full,treatment)
 Pairwise comparisons ( hsd ) of treatment
                 estimate signif diff     lower     upper
 * Oil - Spores     10.5      8.603748  1.896252 19.10375
 * Oil - Water      12.1      8.603748  3.496252 20.70375
   Spores - Water    1.6      8.603748 -7.003748 10.20375
```

Line 9 fits the model, and lines 10 and 11 provide summary information. Consider the estimated random effects from line 10. The branch (treatment by block) variance component is estimated to be zero, and the between patches within branch variance component is estimated at 45.33. The variance estimate for error in the summarized data (line 4 above) is 17.7; if the estimates were exactly equal to their theoretical values one would expect this value to be (at a minimum) the error variance from line 10 divided by 2 (average of two patches). That would be 45.33/2 or about 22.7, obviously more than 17.7. This explains why the branch variance was estimated at 0.

Furthermore, because the patch to patch variance is as large as it is, the standard errors for the effect estimates from line 10 and pairwise comparisons from line 11 are larger than in lines 3 and 5 above.

If the treatments have factorial structure, one can consider as random interactions either an all-treatment-combinations by block random interaction or multiple block by factor interactions, for example, A:block, B:block, or A:B:block. There are pros and cons to both approaches. Breaking the interaction out is more robust (meaning the computed p-values are likely to be more accurate). For example, if B:block is large but the other two are small, then B:block will only affect estimates and tests of the main effects of B. On the other hand, lumping the three together will mean using an error that is too small for B and too large for A and A:B. The disadvantage to breaking up the interaction is that the individual components of the interaction are es-

GRCB with
factorial
treatments

timated less precisely (fewer equivalent degrees of freedom), and this leads to lower power. A good compromise is to begin with the random interaction fully split out, but remove any random interaction terms (other than the full random interaction) that do not appear to be necessary.

12.2.4   How well did the blocking work?

The gain from using an RCB instead of a CRD is a decrease in error variance, and the loss is a decrease in error degrees of freedom by $(r - 1)$. This loss is only severe for small experiments. How can we quantify our gain or loss from an RCB? As discussed above, the "F-test" for blocks does not correspond to a valid randomization test for blocks. Even if it did, knowing simply that the blocks are not all the same does not tell us what we need to know: how much have we saved by using blocks? We need something other than the F-test to measure that gain.

Suppose that we have an RCB and a CRD to test the same treatments; both designs have the same total size N, and both use the same population of units. The efficiency of the RCB relative to the CRD is the factor by which the sample size of the CRD would need to be increased to have the same information as the RCB. (Information is a technical term; think of two designs with the same information as having approximately the same power or yielding approximately the same length of confidence intervals.) For example, if an RCB with fifteen units has relative efficiency 2, then a CRD using the same population of units would need 30 units to obtain the same information. Units almost always translate to time or money, so reducing N by blocking is one good way to save money.

Efficiency is denoted by E with a subscript to identify the designs being compared. The relative efficiency of an RCB to a CRD is given in the following formula:

$$E_{\text{RCB:CRD}} = \frac{(\nu_{rcb} + 1)(\nu_{crd} + 3)}{(\nu_{rcb} + 3)(\nu_{crd} + 1)} \frac{\sigma^2_{crd}}{\sigma^2_{rcb}} \ ,$$

where σ^2_{crd} and σ^2_{rcb} are the error variances for the CRD and RCB, $\nu_{rcb} = (r - 1)(g - 1)$ is the error degrees of freedom for the RCB design, and $\nu_{crd} = (r - 1)g$ is the error degrees of freedom for the CRD of the same size. The first part is a degrees of freedom adjustment; variances must be estimated and we get better estimates with more degrees of freedom. The second part is the ratio of the error variances for the two different designs. The efficiency is determined primarily by this ratio of variances; the degrees of freedom adjustment is usually close to 1.

We will never know the actual variances σ^2_{crd} or σ^2_{rcb}; we must estimate them. Suppose that we have conducted an RCB experiment. We can estimate σ^2_{rcb} using MS_{E} for the RCB design. We estimate σ^2_{crd} via

$$\widehat{\sigma}^2_{crd} = \frac{(r - 1)\text{MS}_{\text{Blocks}} + ((g - 1) + (r - 1)(g - 1))\text{MS}_{\text{E}}}{(r - 1) + (g - 1) + (r - 1)(g - 1)}$$

This is the weighted average of MS_{Blocks} and MS_E with MS_{Blocks} having weight equal to the degrees of freedom for blocks and MS_E having weight equal to the sum of the degrees of freedom for treatment and error. This is *not* the result of simply pooling sums of squares and degrees of freedom for blocks and error in the RCB.

Example 12.6 Mealybugs, continued

For the mealybug experiment, we have $g = 3$, $r = 5$, $\nu_{rcb} = (r-1)(g-1) = 8$, $\nu_{crd} = g(r-1) = 12$, $MS_{Blocks} = 171.6$, and $MS_E = 17.725$, so we get

$$\widehat{\sigma}^2_{crd} = \frac{4 \times 171.6 + (2+8) \times 17.725}{4+2+8} = 61.69 ,$$

$$\frac{(\nu_{rcb}+1)(\nu_{crd}+3)}{(\nu_{rcb}+3)(\nu_{crd}+1)} = \frac{9 \times 15}{11 \times 13} = .944 ,$$

$$\widehat{E}_{RCB:CRD} = \frac{(\nu_{rcb}+1)(\nu_{crd}+3)}{(\nu_{rcb}+3)(\nu_{crd}+1)} \frac{\widehat{\sigma}^2_{crd}}{MS_E} ,$$

$$= .944 \times \frac{61.69}{17.725} = 3.29 .$$

We had five units for each treatment, so an equivalent CRD would have needed $5 \times 3.29 = 16.45$, call it seventeen units per treatment. This blocking was rather successful. Observe that even in this fairly small experiment, the loss from degrees of freedom was rather minor.

12.2.5 Balance and missing data

The standard RCB is balanced, in the sense that each treatment occurs once in each block. Balance was helpful in factorials, and it is helpful in randomized complete blocks for the same reason: it makes the calculations and inference easier. When the data are balanced, simple formulae can be used, exactly as for balanced factorials. When the data are balanced, adding 1 million to all the responses in a given block does not change any contrast between treatment means.

Balance makes inference easier

Missing data in an RCB destroy balance. The approach to inference is to look at treatment effects adjusted for blocks. If the treatments are themselves factorial, we can compute whatever type of sum of squares we feel is appropriate, but we always adjust for blocks prior to treatments. The reason is that we believed, before any experimentation, that blocks affected the response. We thus allow blocks to account for any variability they can before examining any additional variability that can be explained by treatments. This "ordering" for sums of squares and testing does not affect the final estimated effects for either treatments or blocks.

Treatments adjusted for blocks

12.3 Latin Squares and Related Row/Column Designs

Randomized Complete Block designs allow us to block on a single source of variation in the responses. There are experimental situations with more than one source of extraneous variation, and we need designs for these situations.

Example 12.7 Addled goose eggs

The Canada goose (*Branta canadensis*) is a magnificent bird, but it can be a nuisance in urban areas when present in large numbers. One population control method is to addle eggs in nests to prevent them from hatching. This method may be harmful to the adult females, because the females fast while incubating and tend to incubate as long as they can if the eggs are unhatched. Would the removal of addled eggs at the usual hatch date prevent these potential side effects?

An experiment is proposed to compare egg removal and no egg removal treatments. The birds in the study will be banded and observed in the future so that survival can be estimated for the two treatments. It is suspected that geese nesting together at a site may be similar due to both environmental and interbreeding effects. Furthermore, we know older females tend to nest earlier, and they may be more fit.

We need to block on both site and age. We would like each treatment to be used equally often at all sites (to block on populations), and we would like each treatment to be used equally often with young and old birds (to block on age).

A Latin Square (LS) is a design that blocks for two sources of variation. A Latin Square design for g treatments uses g^2 units and is thus a little restrictive on experiment size. Latin Squares are usually presented pictorially. Here are examples of LS designs for $g = 2, 3,$ and 4 treatments:

LS has g^2 units for g treatments and blocks two ways

B	A
A	B

A	B	C
B	C	A
C	A	B

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

The g^2 units are represented as a square (what a surprise!). By convention, the letters A, B, and so on represent the g different treatments. There are two blocking factors in a Latin Square, and these are represented by the rows and columns of the square. Each treatment occurs once in each row and once in each column. Thus in the goose egg example, we might have rows one and two be different nesting sites, with column one being young birds and column two being older birds. This square uses four units, one young and one old bird from each of two sites. Using the two by two square above, treatment A is given to the site 1 old female and the site 2 young female, and treatment B is given to the site 1 young female and the site 2 old female.

Each treatment once in each row and column

Look a little closer at what the LS design is accomplishing. If you ignore the row blocking factor, the LS design is an RCB for the column blocking

factor (each treatment appears once in each column). If you ignore the column blocking factor, the LS design is an RCB for the row blocking factor (each treatment appears once in each row). The rows and columns are also balanced because of the square arrangement of units. A Latin Square blocks on both rows and columns *simultaneously*.

Rows and columns of LS form RCBs

We use Latin Squares because they allow blocking on two sources of variation, but Latin Squares do have drawbacks. First, a single Latin Square has exactly g^2 units. This may be too few or even too many units. Second, Latin Squares generally have relatively few degrees of freedom for estimating error; this problem is particularly serious for small designs. Third, it may be difficult to obtain units that block nicely on both sources of variation. For example, we may have two sources of variation, but one source of variation may only have $g - 1$ units per block.

12.3.1 The crossover design

One of the more common uses for a Latin Square arises when a sequence of treatments is given to a subject over several time periods. We need to block on subjects, because each subject tends to respond differently, and we need to block on time period, because there may be consistent differences over time due to growth, aging, disease progression, or other factors. A *crossover* design has each treatment given once to each subject, and has each treatment occurring an equal number of times in each time period. With g treatments given to g subjects over g time periods, the crossover design is a Latin Square. (We will also consider a more sophisticated view of and analysis for the crossover design in Chapter 16.)

Crossover design has subject and time period blocks

Example 12.8 Bioequivalence of drug delivery

Consider the blood concentration of a drug after the drug has been administered. The concentration will typically start at zero, increase to some maximum level as the drug gets into the bloodstream, and then decrease back to zero as the drug is metabolized or excreted. These time-concentration curves may differ if the drug is delivered in a different form, say a tablet versus a capsule. Bioequivalence studies seek to determine if different drug delivery systems have similar biological effects. One variable to compare is the area under the time-concentration curve. This area is proportional to the average concentration of the drug.

We wish to compare three methods for delivering a drug: a solution, a tablet, and a capsule. Our response will be the area under the time-concentration curve. We anticipate large subject to subject differences, so we block on subject. There are three subjects, and each subject will be given the drug three times, once with each of the three methods. Because the body may adapt to the drug in some way, each drug will be used once in the first period, once in the second period, and once in the third period. Table 12.3 gives the assignment of treatments and the responses (data from Selwyn and Hall 1984, data set `Bioequivalence`). This Latin Square is a crossover design.

Table 12.3: Area under the curve for administering a drug via
A—solution, B—tablet, and C—capsule. Table entries are treatments
and responses.

Period	Subject 1		Subject 2		Subject 3	
1	A	1799	C	2075	B	1396
2	C	1846	B	1156	A	868
3	B	2147	A	1777	C	2291

12.3.2 Randomizing the LS design

It is trivial to produce an LS for any number of treatments g. Assign the treatments in the first row in order. In the remaining rows, shift left all the treatments in the row above, bringing the first element of the row above around to the end of this row. The three by three square on page 435 was produced in this fashion. It is much less trivial to choose a square randomly. In principle, you assign treatments to units randomly, subject to the restrictions that each treatment occurs once in each row and once in each column, but effecting that randomization is harder than it sounds.

One LS is easy, random LS is harder

The recommended randomization is described in Fisher and Yates (1963). This randomization starts with *standard squares*, which are squares with the letters in the first row and first column in order. The three by three and four by four squares on page 435 are standard squares. For g of 2, 3, 4, 5, and 6, there are 1, 1, 4, 56, and 9408 standard squares. Appendix B contains several standard Latin Square plans.

Standard squares

The Fisher and Yates randomization goes as follows. For g of 3, 4, or 5, first choose a standard square at random. Then randomly permute all rows except the first, randomly permute all columns, and randomly assign the treatments to the letters. For g of 6, select a standard square at random, randomly permute all rows and columns, and randomly assign the treatments to the letters. For g of 7 or greater, choose any square, randomly permute the rows and columns, and randomly assign treatments to the letters.

Fisher-Yates randomization

12.3.3 Analysis for the LS design

The standard model for a Latin Square has a grand mean, effects for row and column blocks and treatments, and experimental error. Let y_{ijk} be the response from the unit given the ith treatment in the jth row block and kth column block. The standard model is

Additive treatment, row, and column effects

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk} \ ,$$

where α_i is the effect of the ith treatment, β_j is the effect of the j row block, and γ_k is the effect of the kth column block. Blocking factors could be random (for example, random subjects) or fixed (for example, a steady progression across treatment periods), but the inference for treatments will be the

same when the data are complete. As with the RCB, block effects are assumed to be additive, and assuming additivity does not make additivity true.

Here is something new: we do not observe all g^3 of the i, j, k combinations in an LS; we only observe g^2 of them. However, the LS is constructed so that we have balance when we look at rows and columns, rows and treatments, or columns and treatments. This balance implies that contrasts between rows, contrasts between columns, and contrasts between treatments are all orthogonal, and the standard calculations in fixed effects models for effects, sums of squares, contrasts, and so on work for the LS. Thus, for example,

$$\widehat{\alpha}_i = \overline{y}_{i\bullet\bullet} - \overline{y}_{\bullet\bullet\bullet}$$
$$\text{SS}_{\text{Trt}} = \sum_{i=1}^{g} g\widehat{\alpha}_i^2 \ .$$

Note that $\overline{y}_{\bullet\bullet\bullet}$ and $\overline{y}_{i\bullet\bullet}$ are means over g^2 and g units respectively. The sum of squares for error is the sum of the squared residuals, but it can also be found by subtracting the sums of squares for treatments, rows, and columns from the total sum of squares.

The Analysis of Variance table for a Latin Square design has sources for rows, columns, treatments, and error. We test the null hypothesis of no treatment effects via the F-ratio formed by mean square for treatments over mean square for error. As in the RCB, we do not test row or column blocking. Here is a schematic ANOVA table for a Latin Square:

Source	SS	DF	MS	F
Rows	SS_{Rows}	$g-1$	$\text{SS}_{\text{Rows}}/(g-1)$	
Columns	SS_{Cols}	$g-1$	$\text{SS}_{\text{Cols}}/(g-1)$	
Treatments	SS_{Trt}	$g-1$	$\text{SS}_{\text{Trt}}/(g-1)$	$\text{MS}_{\text{Trt}}/\text{MS}_{\text{E}}$
Error	SS_{E}	$(g-2)(g-1)$	$\text{SS}_{\text{E}}/[(g-2)(g-1)]$	

There is no intuitive rule for the degrees of freedom for error $(g-2)(g-1)$; we just have to do our sums. Start with the total degrees of freedom g^2 and subtract one for the constant and all the degrees of freedom in the model, $3(g-1)$. The difference is $(g-2)(g-1)$. Latin Squares can have few degrees of freedom for error.

Example 12.9 Bioequivalence, continued

Let's analyze the bioequivalence data from Table 12.3.

```
1 > biofit <- lm(area~period+subject+treatment,Bioequivalence)
2 > anova(biofit)
  Response: area
            Df Sum Sq Mean Sq F value   Pr(>F)
  period     2 928006  464003 103.231 0.009594 **
  subject    2 261115  130557  29.047 0.033282 *
  treatment  2 608891  304445  67.733 0.014549 *
  Residuals  2   8990    4495
3 > pairwise(biofit,treatment)

  Pairwise comparisons ( hsd ) of treatment
                    estimate signif diff     lower     upper
    soln - tablet    -85.0000       322.4625 -407.4625  237.4625
  * soln - capsule  -589.3333       322.4625 -911.7959 -266.8708
  * tablet - capsule -504.3333      322.4625 -826.7959 -181.8708
```

Line 1 fits the model with blocks (period and subject) and treatments. The ANOVA from line 2 shows reasonable evidence against the null hypothesis of no differences between treatments. The output shows F-tests for both period and subject. We should ignore these, because period and subject are unrandomized blocking factors. Line 3 shows that capsule seems to have a higher area than either solution or tablet, which we cannot tell apart. We would have obtained the same results with random subject effects.

Note that this three by three Latin Square has only 2 degrees of freedom for error. Even an F of 67.7 does not produce a tiny p-value.

12.3.4   Replicating Latin Squares

Increased replication gives us better estimates of error and increased power through averaging. We often need better estimates of error in LS designs, because a single Latin Square has relatively few degrees of freedom for error. Thus using multiple Latin Squares in a single experiment is common practice.

Replicate for better precision and error estimates

When we replicate a Latin Square, we may be able to "reuse" row or column blocks. For example, we may believe that the period effects in a crossover design will be the same in all squares; this reuses the period blocks across the squares. Replicated Latin Squares can reuse both row and column blocks, reuse neither row nor column blocks, or reuse one of the row or column blocks. Whether we reuse any or all of the blocks when replicating an LS depends on the experimental and logistical constraints. Some blocks may represent small batches of material or time periods when weather is fairly constant; these blocks may be unavailable or have been consumed prior to the second replication. Other blocks may represent equipment that could be reused in principle, but we might want to use several pieces of equipment at once to conclude the experiment sooner rather than later.

Some blocks can be reused

Reusability depends on experiment and logistics

From an analysis point of view, the advantage of reusing a block factor is that we will have more degrees of freedom for error. The risk when reusing a block factor is that the block effects will actually change, so that the assumption of constant block effects across the squares is invalid.

Example 12.10 Carbon monoxide emissions

Carbon monoxide (CO) emissions from automobiles can be influenced by the formulation of the gasoline that is used. In Minnesota, we use "oxygenated fuels" in the winter to decrease CO emissions. We have four gasoline blends, the combinations of factors A and B, each at two levels, and we wish to test the effects of these blends on CO emissions in nonlaboratory conditions, that is, in real cars driven over city streets. We know that there are car to car differences in CO emissions, and we suspect that there are route to route differences in the city (stop and go versus freeway, for example). With two blocking factors, a Latin Square seems appropriate. We will use three squares to get enough replication.

If we have only four cars and four routes, and these will be used in all three replications, then we are reusing the row and column blocking factors across squares. Alternatively, we might be using only four cars, but we have twelve different routes. Then we are reusing the row blocks (cars), but not the column blocks (routes). Finally, we could have twelve cars and twelve routes, which we divide into three sets of four each to create squares. For this design, neither rows nor columns is reused.

The analysis of a replicated Latin Square varies slightly depending on which blocks are reused. Let y_{ijkl} be the response for treatment i in row j and column k of square l. There are g treatments (and rows and columns in each block) and m squares. Consider the provisional model

Models depend on which blocks are reused

$$y_{ijkl} = \mu + \alpha_i + \beta_{j(l)} + \gamma_{k(l)} + \delta_l + \epsilon_{ijkl} \ .$$

This model has an overall mean μ, the treatment effects α_i, square effects δ_l, and row and column block effects $\beta_{j(l)}$ and $\gamma_{k(l)}$. As usual in block designs, block effects are additive.

This model has row and column effects nested in square, so that each square will have its own set of row and column effects. This model is appropriate when neither row nor column blocks are reused. The degrees of freedom for this model are one for the grand mean, $g - 1$ between treatments, $m - 1$ between squares, $m(g - 1)$ for each of rows and columns, and $(mg - m - 1)(g - 1)$ for error.

Df when neither rows nor columns reused

The model terms and degrees of freedom for the row and column block effects depend on whether we are reusing the row and/or column blocks. Suppose that we reuse row blocks, but not column blocks; reusing columns but not rows can be handled similarly. The model is now

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{k(l)} + \delta_l + \epsilon_{ijkl} \ ,$$

and the degrees of freedom are one for the grand mean, $g - 1$ between treatments, $m - 1$ between squares, $g - 1$ between rows, $m(g - 1)$ between columns, and $(mg - 2)(g - 1)$ for error. Finally, consider reusing both row and column blocks. Then the model is

Df when rows reused

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijkl} \ ,$$

and the degrees of freedom are one for the grand mean, $g - 1$ between treatments, rows and columns, $m - 1$ between squares, and $(mg + m - 3)(g - 1)$ for error.

Df when rows and columns reused

Example 12.11 CO emissions, continued

Consider again the three versions of the CO emissions example given above. The degrees of freedom for the sources of variation are

Source	4 cars, 4 routes DF	4 cars, 12 routes DF	12 cars, 12 routes DF
Squares	$(m - 1) = 2$	$(m - 1) = 2$	$(m - 1) = 2$
Cars	$(g - 1) = 3$	$(g - 1) = 3$	$m(g - 1) = 9$
Routes	$(g - 1) = 3$	$m(g - 1) = 9$	$m(g - 1) = 9$
Fuels	$(g - 1) = 3$	$(g - 1) = 3$	$(g - 1) = 3$
or A	1	1	1
B	1	1	1
AB	1	1	1
Error	$(mg + m - 3)(g - 1)$ $= 12 \times 3 = 36$	$(mg - 2)(g - 1)$ $= 10 \times 3 = 30$	$(mg - m - 1)(g - 1)$ $= 8 \times 3 = 24$
or			
Error	$47 - 11 = 36$	$47 - 17 = 30$	$47 - 23 = 24$

Note that we have computed error degrees of freedom twice, once by applying the formulae, and once by subtracting model degrees of freedom from total degrees of freedom. I usually obtain error degrees of freedom by subtraction.

We have presented the degrees of freedom for replicated Latin Squares in the context of all effects fixed. In practice, some effects may be random. REML analysis using random row or column blocks for balanced data will have the same effective degrees of freedom for tests as what we have seen for fixed effects.

Random blocks

Estimated effects follow the usual patterns, because even though we do not see all the $ijkl$ combinations, the combinations we do see are such that treatment, row, and column effects are orthogonal. So, for example,

Estimated effects and sums of squares follow the usual patterns

$$\widehat{\alpha}_i = \overline{y}_{i\bullet\bullet\bullet} - \overline{y}_{\bullet\bullet\bullet\bullet}$$
$$\widehat{\delta}_l = \overline{y}_{\bullet\bullet\bullet l} - \overline{y}_{\bullet\bullet\bullet\bullet} .$$

If row blocks are reused, we have

$$\widehat{\beta}_j = \overline{y}_{\bullet j\bullet\bullet} - \overline{y}_{\bullet\bullet\bullet\bullet} ,$$

and if row blocks are not reused we have

$$\widehat{\beta}_{j(l)} = \overline{y}_{\bullet j\bullet l} - \widehat{\delta}_l - \widehat{\mu}$$
$$= \overline{y}_{\bullet j\bullet l} - \overline{y}_{\bullet\bullet\bullet l} .$$

Table 12.4: Area under the curve for administering a drug via
A—solution, B—tablet, and C—capsule. Table entries are treatments
and responses.

		Period					
Subject		1		2		3	
1	A	1799	C	1846	B	2147
2	C	2075	B	1156	A	1777
3	B	1396	A	868	C	2291
4	B	3100	A	3065	C	4077
5	C	1451	B	1217	A	1288
6	A	3174	C	1714	B	2919
7	C	1430	A	836	B	1063
8	A	1186	B	642	C	1183
9	B	1135	C	1305	A	984
10	C	873	A	1426	B	1540
11	A	2061	B	2433	C	1337
12	B	1053	C	1534	A	1583

The rules for column block effects are analogous. In all cases, the sum of
squares for a source of variation is found by squaring an effect, multiplying
that by the number of responses that received that effect, and adding across
all levels of the effect.

When only one of the blocking factors (rows, for example) is reused, it is
fairly common to combine the terms for "between squares" ($m-1$ degrees of
freedom) and "between columns within squares" ($m(g-1)$ degrees of free-
dom) into an overall between columns factor with $gm-1$ degrees of freedom.
This is not necessary, but it sometimes makes the software commands easier.
Note that when neither rows nor columns is reused, you cannot get combined
$m(g-1)$ degrees of freedom terms for both rows and columns at the same
time. The "between squares" sums of squares and degrees of freedom comes
from contrasts between the means of the different squares and can be con-
sidered as either a row or column difference, but it cannot be combined into
both rows and columns in the same analysis.

Can combine
between squares
with columns

Example 12.12 Bioequivalence (continued)

Example 12.8 introduced a three by three Latin Square for comparing
delivery of a drug via solution, tablet, and capsule. In fact, this crossover
design included $m = 4$ Latin Squares. These squares involve twelve different
subjects, but the same three time periods. Data are given in Table 12.4, data
set `BioequivalenceFull`. The subject factor in this data set enumerates
the subjects from 1 through 12.

Line 1 fits the Latin Square model using random blocks. Line 2 gives summary information, and line 3 the ANOVA. Note that the complete data set is compatible with the null hypothesis of no treatment effects.

```
1 > fit.full <- lmer(area~period+(1|subject)+treatment,BioequivalenceFull)
2 > summary(fit.full)
  Random effects:
   Groups    Name         Variance  Std.Dev.
   subject   (Intercept)  428077    654.3
   Residual               205325    453.1
  Number of obs: 36, groups:  subject, 12

  Fixed effects:
              Estimate Std. Error t value
  (Intercept)  1693.44     203.41   8.325
  period1         34.31     106.80   0.321
  period2       -189.94     106.80  -1.778
  treatment1     -22.86     106.80  -0.214
  treatment2     -43.36     106.80  -0.406
3 > Anova(fit.full,test="F")
  Response: area
                F Df Df.res Pr(>F)
  period    1.7965  2     20 0.1916
  treatment 0.1984  2     20 0.8217
```

Those of you keeping score may recall from Example 12.9 that the data from just the first square seemed to indicate that there were differences between the treatments. Also the MS_E in the complete data is about 45 times bigger than for the first square. What has happened?

Here are two possibilities. First, the subjects may not have been numbered in a random order, so the early subjects could be systematically different from the later subjects. This can lead to some dramatic differences between analysis of subsets and complete sets of data, though we have no real evidence of that here.

Second, there could be subject by treatment interaction giving rise to different treatment effects for different subsets of the data. Our Latin Square blocking model is based on the assumption of additivity, but interaction could be present. The error term in our ANOVA contains any effects not explicitly modeled, so it would be inflated in the presence of subject by treatment interaction, and interaction could obviously lead to different treatment effects being estimated in different squares.

We explore this somewhat at line 4.

```
4 > square <- factor(rep(1:4,each=9))
5 > fit.full2 <- lmer(area~period+treatment+(1|square:treatment)+(1|subject),
+                     BioequivalenceFull)
 Warning message:
 In checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv,  :
 Model failed to converge with max|grad| = 0.00360962 (tol = 0.002, component 1)
6 > summary(fit.full2)
 ...
 Random effects:
  Groups           Name          Variance  Std.Dev.
  square:treatment (Intercept)       65.67    8.104
  subject          (Intercept) 427576.55  653.893
  Residual                     205372.05  453.180
 Number of obs: 36, groups:  square:treatment, 12; subject, 12

 Fixed effects:
             Estimate Std. Error t value
 (Intercept)  1693.44     203.33   8.329
 period1         34.31     106.82   0.321
 period2       -189.94     106.82  -1.778
 treatment1     -22.86     106.87  -0.214
 treatment2     -43.36     106.87  -0.406
 ...
```

We create a factor for square, and then in line 5 we include a treatment by square random term. Line 6 gives the summary. The variance of the interaction term is estimated to be quite small relative to the other random terms, and the standard errors of the treatment effects barely change. This does not give us confidence in the model with square by treatment interaction.

12.3.5   Efficiency of Latin Squares

We approach the efficiency of Latin Squares much as we did the efficiency of RCB designs. That is, we try to estimate by what factor the sample sizes would need to be increased in order for a simpler design to have as much information as the LS design. We can compare an LS design to an RCB by considering the elimination of either row or column blocks, or we can compare an LS design to a CRD by considering the elimination of both row and column blocks.

As with RCB's, our estimate of efficiency is the product of two factors, the first a correction for degrees of freedom for error and the second an estimate of the ratio of the error variances for the two designs. With g^2 units in a Latin Square, there are $\nu_{ls} = (g-1)(g-2)$ degrees of freedom for error; if either row or column blocks are eliminated, there are $\nu_{rcb} = (g-1)(g-1)$ degrees of freedom for error; and if both row and column blocks are eliminated, there are $\nu_{crd} = (g-1)g$ degrees of freedom for error.

The efficiency of a Latin Square relative to an RCB is

$$E_{\text{LS:RCB}} = \frac{(\nu_{ls}+1)(\nu_{rcb}+3)}{(\nu_{ls}+3)(\nu_{rcb}+1)} \frac{\sigma_{rcb}^2}{\sigma_{ls}^2} ,$$

and the efficiency of a Latin Square relative to a CRD is

$$E_{\text{LS:CRD}} = \frac{(\nu_{ls}+1)(\nu_{crd}+3)}{(\nu_{ls}+3)(\nu_{crd}+1)} \frac{\sigma^2_{crd}}{\sigma^2_{ls}}.$$

We have already computed the degrees of freedom, so all that remains is the estimates of variance for the three designs.

The estimated variance for the LS design is simply MS_E from the LS design. For the RCB and CRD we estimate the error variance in the simpler design with a weighted average of the MS_E from the LS and the mean squares from the blocking factors to be eliminated. The weight for MS_E is $(g-1)^2$, the sum of treatment and error degrees of freedom, and the weights for blocking factors are their degrees of freedom $(g-1)$. In formulae:

$$\widehat{\sigma}^2_{rcb} = \frac{(g-1)\text{MS}_\text{Rows} + ((g-1)+(g-1)(g-2))\text{MS}_\text{E}}{2(g-1)+(g-1)(g-2)}$$

$$= \frac{\text{MS}_\text{Rows}+(g-1)\text{MS}_\text{E}}{g} \qquad \text{(row blocks eliminated)},$$

or

$$\widehat{\sigma}^2_{rcb} = \frac{(g-1)\text{MS}_\text{Cols} + ((g-1)+(g-1)(g-2))\text{MS}_\text{E}}{2(g-1)+(g-1)(g-2)}$$

$$= \frac{\text{MS}_\text{Cols}+(g-1)\text{MS}_\text{E}}{g} \qquad \text{(column blocks eliminated)},$$

or

$$\widehat{\sigma}^2_{crd} = \frac{(g-1)(\text{MS}_\text{Rows}+\text{MS}_\text{col}+\text{MS}_\text{E})+(g-1)(g-2)\text{MS}_\text{E}}{3(g-1)+(g-1)(g-2)}$$

$$= \frac{\text{MS}_\text{Rows}+\text{MS}_\text{Cols}+(g-1)\text{MS}_\text{E}}{g+1} \qquad \text{(both eliminated)}.$$

The two versions of $\widehat{\sigma}^2_{rcb}$ are for eliminating row and column blocking, respectively.

Example 12.13 Bioequivalence, continued

Example 12.9 gave the ANOVA table for the first square of the bioequivalence data. The mean squares for subject, period, and error were 130,557; 464,003; and 4494.8 respectively. All three of these and treatments had 2 degrees of freedom each. Thus we have $\nu_{ls}=2$, $\nu_{rcb}=4$, and $\nu_{crd}=6$. The

estimated variances are

Blocking removed

$$\text{Neither} \quad \widehat{\sigma}_{ls}^2 \quad = \quad 4494.8$$

$$\text{Subjects} \quad \widehat{\sigma}_{rcb}^2 \quad = \quad \frac{130,557 + 2 \times 4494.8}{3} = 46516$$

$$\text{Periods} \quad \widehat{\sigma}_{rcb}^2 \quad = \quad \frac{464,003 + 2 \times 4494.8}{3} = 157664$$

$$\text{Both} \quad \widehat{\sigma}_{crd}^2 \quad = \quad \frac{130557 + 464,003 + 2 \times 4494.8}{4} = 150887 \quad .$$

The estimated efficiencies are

$$\text{Subjects} \quad E \quad = \quad \frac{(2+1)(4+3)}{(2+3)(4+1)} \frac{46516}{4494.8} = 8.69$$

$$\text{Periods} \quad E \quad = \quad \frac{(2+1)(4+3)}{(2+3)(4+1)} \frac{157664}{4494.8} = 29.46$$

$$\text{Both} \quad E \quad = \quad \frac{(2+1)(6+3)}{(2+3)(6+1)} \frac{150887}{4494.8} = 25.90 \quad .$$

Both subject and period blocking were effective, particularly the period blocking.

12.3.6   Designs balanced for residual effects

Crossover designs give all treatments to all subjects and use subjects and periods as blocking factors. The standard analysis includes terms for subject, period, and treatment. There is an implicit assumption that the response in a given time period depends on the treatment for that period, and not at all on treatments from prior periods. This is not always true. For example, a drug that is toxic and has terrible side effects may alter the responses for a subject, even after the drug is no longer being given. These effects that linger after treatment are called *residual effects* or *carryover effects*.

Residual effects affect subsequent treatment periods

There are experimental considerations when treatments may have residual effects. A *washout period* is a time delay inserted between successive treatments for a subject. The idea is that residual effects will decrease or perhaps even disappear given some time, so that if we can design this time into the experiment between treatments, we won't need to worry about the residual effects. Washout periods are not always practical or completely effective, so alternative designs and models have been developed.

A washout period may reduce residual effects

In an experiment with no residual effects, only the treatment from the current period affects the response. The simplest form of residual effect occurs when only the current treatment and the immediately preceding treatment

Balance for residual effects of preceding treatment

affect the response. A design balanced for residual effects, or carryover de-sign, is a crossover design with the additional constraint that each treatment follows every other treatment an equal number of times.

Look at these two Latin Squares with rows as periods and columns as subjects.

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

A	B	C	D
B	D	A	C
C	A	D	B
D	C	B	A

In the first square, A occurs first once, follows B twice, and follows D once. Other treatments have a similar pattern. The first square is a crossover design, but it is not balanced for residual effects. In the second square, A occurs first once, and follows B, C, and D once each. A similar pattern occurs for the other treatments, so the second square is balanced for residual effects. When g is even, we can find a design balanced for residual effects using g subjects; when g is odd, we need $2g$ subjects (two squares) to balance for residuals effects. A design that includes all possible orders for the treatments an equal number of times will be balanced for residual effects.

The model for a residual-effects design has terms for subject, period, direct effect of a treatment, residual effect of a treatment, and error. Specif-ically, let y_{ijkl} be the response for the kth subject in the lth time period; the subject received treatment i in period l and treatment j in period $l-1$. The indices i and l run from 1 to g, and k runs across the number of subjects. Use $j = 0$ to indicate that there was no earlier treatment (that is, when $l = 1$ and we are in the first period); j then runs from 0 to g. Our model is

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijkl}$$

where α_i is called the direct effect of treatment i, β_j is called the residual effect of treatment j, and γ_k and δ_l are subject and period effects as usual. We make the usual zero-sum assumptions for the block and direct treatment effects. For the β_j's we assume that $\beta_0 = 0$ and $\sum_{j=1}^{g} \beta_j = 0$. That is, we assume that there is a zero residual effect when in the first treatment period.

Direct treatment effects are orthogonal to block effects (we have a cross-over design), but residual effects are not orthogonal to direct treatment effects or subjects. Formulae for estimated effects and sums of squares are thus rather opaque, and it seems best just to let your statistical software do its work.

Example 12.14 Milk yield

Milk production in cows may depend on their feed. There is large cow to cow variation in production, so blocking on cow and giving all the treatments to each cow seems appropriate. Milk production for a given cow also tends to decrease during any given lactation, so blocking on period is important. This

Table 12.5: Milk production (pounds per 6 weeks) for eighteen cows fed A—roughage, B—limited grain, and C—full grain.

	Cow					
Period	1	2	3	4	5	6
1	A 1376	B 2088	C 2238	A 1863	B 1748	C 2012
2	B 1246	C 1864	A 1724	C 1755	A 1353	B 1626
3	C 1151	A 1392	B 1272	B 1462	C 1339	A 1010
	7	8	9	10	11	12
1	A 1655	B 1938	C 1855	A 1384	B 1640	C 1677
2	B 1517	C 1804	A 1298	C 1535	A 1284	B 1497
3	C 1366	A 969	B 1233	B 1289	C 1370	A 1059
	13	14	15	16	17	18
1	A 1342	B 1344	C 1627	A 1180	B 1287	C 1547
2	B 1294	C 1312	A 1186	C 1245	A 1000	B 1297
3	C 1371	A 903	B 1066	B 1082	C 1078	A 887
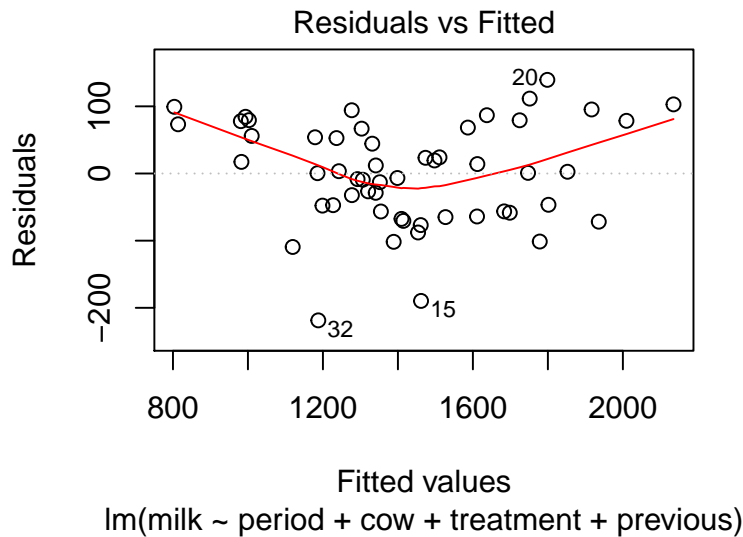
Figure 12.2: Residuals versus predicted values for the milk production data on the original scale.

leads us to a crossover design. The treatments of interest are A—roughage, B—limited grain, and C—full grain. The response will be the milk production during the six week period the cow is on a given feed. There was

insufficient time for washout periods, so the design was balanced for residual effects. Table 12.5 gives the data from Cochran, Autrey, and Cannon (1941) via Bellavance and Tardif (1995), data set `MilkProduction`.

We need to set up a factor to model the residual effect. The `previous` column of the data frame has the same levels as `treatment`, but adds "none" as a level for first observations for each cow. Because we are using the "treatment effects add to zero" parameterization, and because the contrast between the first row (which always has `previous` equal to "none") is colinear with the period effect, it works best to arrange the levels of `previous` so that "none" (or "first" or whatever we call it) is the next to last level of the factor. We do this in lines 2–3.

```
1 > data("MilkProduction")
2 > levels(MilkProduction$treatment)
  [1] "roughage"   "lim.grain"  "full.grain"
3 > newprev <- factor(MilkProduction$previous,
  +   levels=c("roughage","lim.grain","none","full.grain"))
4 > fit1 <- lm(milk~period+cow+treatment+newprev,MilkProduction)
5 > plot(fit1,which=1)
6 > car::boxCox(fit1)
7 > fit2 <- lm(log(milk)~period+cow+treatment+newprev,MilkProduction)
8 > plot(fit2,which=1)
```

Line 4 fits the model with period, subject, treatment, and residual effects, and line 5 plots the residuals (shown in Figure 12.2). The residuals show the flopping fish pattern. Box-Cox analysis (line 6, results not shown) indicate that a log transformation should help. We refit with transformed data in line 7 and plot the new residuals in line 8. This new residual plot (not shown) looks good except for a potential outlier.

The outlier test on line 9 indicates that observation 32 could well be an outlier, being substantially lower than the rest of the data would lead us to expect. We should go back to the original data collection notebooks to see if there is some reason to distrust observation 32. In the meantime, we construct an indicator variable for point 32 and refit with this indicator variable on lines 10-11. The residual plot for the transformed data with the indicator on line 12 looks good (not shown).

```
9 > car::outlierTest(fit2)
      rstudent unadjusted p-value Bonferroni p
 32 -4.699096          5.8507e-05     0.0031594
0 > pt32 <- rep(0,54);pt32[32]<-1
1 > fit3 <- lm(log(milk)~period+cow+treatment+newprev+pt32,MilkProduction)
2 > plot(fit3)
3 > summary(fit3)
  ...
              Estimate Std. Error  t value Pr(>|t|)
  (Intercept)  7.244693   0.006751 1073.079  < 2e-16 ***
  period1      0.151477   0.009467   16.001 6.27e-16 ***
  period2      0.011219   0.009467    1.185 0.245572
  cow1        -0.088886   0.027760   -3.202 0.003302 **
  cow2        -0.009089   0.027900   -0.326 0.746933
  ...
  cow17        0.161400   0.035077    4.601 7.67e-05 ***
  treatment1  -0.116120   0.010859  -10.694 1.41e-11 ***
  treatment2   0.003424   0.010859    0.315 0.754792
  newprev1    -0.040576   0.014108   -2.876 0.007471 **
  newprev2    -0.023829   0.014829   -1.607 0.118895
  newprev3          NA         NA       NA       NA
  pt32        -0.315417   0.067123   -4.699 5.85e-05 ***

4 > anova(fit3)
  Response: log(milk)
            Df  Sum Sq Mean Sq  F value     Pr(>F)
  period     2 0.99807 0.49903 209.8635  < 2.2e-16 ***
  cow       17 0.90727 0.05337  22.4437  2.933e-12 ***
  treatment  2 0.40999 0.20500  86.2095  6.242e-13 ***
  newprev    2 0.03378 0.01689   7.1034   0.003085 **
  pt32       1 0.05251 0.05251  22.0815  5.851e-05 ***
  Residuals 29 0.06896 0.00238
```

The summary on line 13 shows that the period and cow effects are large relative to the treatment and residual effects; we also see that in the ANOVA on line 14. Blocking really worked here. The roughage treatment reduces the response (on log scale) by .12, and it reduces the response in the following period by .04. On the other hand, using the zero sum constraint we see that the full grain treatment increases the response by about .11, and it increases the response by about .06 in the following period.

We could model with cow as a random effect.

```
15 > fit4 <- lmer(log(milk)~period+(1|cow)+ treatment+newprev+pt32,MilkProduction)
   fixed-effect model matrix is rank deficient so dropping 1 column / coefficient
16 > summary(fit4)
   ...
   Random effects:
    Groups   Name        Variance Std.Dev.
    cow      (Intercept) 0.017834 0.13354
    Residual             0.002377 0.04875
   Number of obs: 54, groups:  cow, 18

   Fixed effects:
                Estimate Std. Error t value
   (Intercept)  7.244475   0.032191 225.045
   period1      0.151696   0.009462  16.032
   period2      0.011438   0.009462   1.209
   treatment1  -0.116682   0.010828 -10.776
   treatment2   0.003776   0.010828   0.349
   newprev1    -0.040955   0.014024  -2.920
   newprev2    -0.023426   0.014708  -1.593
   pt32        -0.303641   0.066206  -4.586
   ...
```

Because the residual effect is not orthogonal to the other effects and additionally the outlier indicator variable breaks orthogonality, we see that the estimated effects in the random subject model are just slightly different from those in the fixed subject model.

When resources permit an additional test period for each subject, considerable gain can be achieved by repeating the last treatment for each subject. For example, if cow 13 received the treatments A, B, and C, then the treatment in the fourth period should also be C. With this structure, every treatment follows every treatment (including itself) an equal number of times, and every residual effect occurs with every subject. These conditions permit more precise estimation of direct and residual treatment effects.

Repeat last treatment

12.4   Graeco-Latin Squares

Randomized Complete Blocks allow us to control one extraneous source of variability in our units, and Latin Squares allow us to control two sources. The Latin Square design can be extended to control for three sources of extraneous variability; this is the Graeco-Latin Square. For four or more sources of variability, we use Latin Hyper-Squares. Graeco-Latin Squares allow us to test g treatments using g^2 units blocked three different ways. Graeco-Latin Squares don't get used very often, because they require a fairly restricted set of circumstances to be applicable.

Graeco-Latin Squares block three ways

The Graeco-Latin Square is represented as a g by g table or square. Entries in the table correspond to the g^2 units. Rows and columns of the square correspond to blocks, as in a Latin Square. Each entry in the table has one Latin letter and one Greek letter. Latin letters correspond to treatments, as in a Latin Square, and Greek letters correspond to the third blocking factor. The Latin letters occur once in each row and column (they form a Latin Square),

Treatments occur once in each blocking factor

and the Greek letters occur once in each row and column (they also form a Latin Square). In addition, each Latin letter occurs once with each Greek letter. Here is a four by four Graeco-Latin Square:

A α	B γ	C δ	D β
B β	A δ	D γ	C α
C γ	D α	A β	B δ
D δ	C β	B α	A γ

Each treatment occurs once in each row block, once in each column block, and once in each Greek letter block. Similarly, each kind of block occurs once in each other kind of block.

If two Latin Squares are superimposed and all g^2 combinations of letters from the two squares once, the Latin Squares are called *orthogonal*. A Graeco-Latin Square is the superposition of two orthogonal Latin Squares.

Orthogonal Latin Squares

Graeco-Latin Squares do not exist for all values of g. For example, there are Graeco-Latin Squares for g of 3, 4, 5, 7, 8, 9, and 10, but *not* for g of 6. Appendix B lists orthogonal Latin Squares for $g = 3, 4, 5, 7$, from which a Graeco-Latin Square can be built.

No GLS for $g = 6$

The usual model for a Graeco-Latin Square has terms for treatments and row, column, and Greek letter blocks and assumes that all these terms are additive. The balance built into these designs allows us to use our standard methods for estimating effects and computing sums of squares, contrasts, and so on, just as for a Latin Square.

Additive blocks plus treatments

The Latin Square/Graeco-Latin Square family of designs can be extended to have more blocking factors. These designs, called Hyper-Latin Squares, are rare in practice.

Hyper Squares

12.5   Further Reading and Extensions

Our discussion of the RCB has focused on its standard form, where we have g treatments and blocks of size g. There are several other possibilities. For example, we may be able to block our units, but there may not be enough units in each block for each treatment. This leads us to incomplete block designs, which we will consider in Chapter 13.

Another possibility is that units are expensive (so we do not want to waste any), but the block sizes are not a nice multiple of the number of treatments. Here, we can combine an RCB (or GRCB) with one of the incomplete block designs from Chapter 13. For example, with three treatments (A, B, and C) and three blocks of size 5, we could use (A, B, C, A, B) in block 1, (A, B, C, A, C) in block 2, and (A, B, C, B, C) in block 3. So each block has one full complement of the treatments, plus two more according to an incomplete block design.

The final possibility that we mention is that we can have blocks with different numbers of units; that is, some blocks have more units than others.

Standard designs assume that all blocks have the same number of units, so we must do something special. The most promising approach is probably *optimal design* via special design software. Optimal design (see Chapter 14) allocates treatments to units in such a way as to optimize some criterion; for example, we may wish to minimize the average variance of the estimated treatment effects. See Silvey (1980). The algorithms that do the optimization are complicated, but software exists that will do what is needed. See Cook and Nachtsheim (1989). Oh yes, in case you were worried, most standard designs such as RCB's are also "optimal" designs; we just don't need the fancy software in the standard situations.

12.6 Problems

Winter road treatments to clear snow and ice can lead to cracking in the pavement. An experiment was conducted comparing four treatments: sodium chloride, calcium chloride, a proprietary organic compound, and sand. Traffic level was used as a blocking factor and a randomized complete block experiment was conducted. One observation is missing, because the spreader in that district was not operating properly. The response is new cracks per mile of treated roadway (data set `Cracks`).

Exercise 12.1

	A	B	C	D
Block 1		32	27	36
Block 2	38	40	43	33
Block 3	40	63	14	27

Our interest is in the following comparisons: chemical versus physical (A,B,C versus D), inorganic versus organic (A,B versus C), and sodium versus calcium (A versus B). Which of these comparisons seem large?

Grains or crystals adversely affect the sensory qualities of foods using dried fruit pulp. A factorial experiment was conducted to determine which factors affect graininess. The factors were drying temperature (three levels), acidity (pH) of pulp (two levels), and sugar content (two levels). The experiment has two replications, with each replication using a different batch of pulp. Response is a measure of graininess (data set `Graininess`).

Exercise 12.2

Temp.	Rep.	Sugar low pH low	Sugar low pH high	Sugar high pH low	Sugar high pH high
1	1	21	12	13	1
	2	21	18	14	8
2	1	23	14	13	1
	2	23	17	16	11
3	1	17	20	16	14
	2	23	17	17	5

Analyze these data to determine which factors effect graininess, and which combination of factors leads to the least graininess.

The data below are from a replicated Latin Square with four treatments; row blocks were reused, but column blocks were not (data set `RLS`). Test for treatment differences and use Tukey HSD with level .01 to analyze the pairwise treatment differences.

Exercise 12.3

```
D 44   B 26   C 67   A 77    B 51   D 62   A 71   C 49
C 39   A 45   D 71   B 74    C 63   A 74   D 67   B 47
B 52   D 49   A 81   C 88    A 74   C 75   B 60   D 58
A 73   C 58   B 76   D 100   D 82   B 79   C 74   A 68
```

Consider replicating a six by six Latin Square three times, where we use the same row blocks but different column blocks in the three replicates. The six treatments are the factorial combinations of factor A at three levels and factor B at two levels. Give the sources and degrees of freedom for the Analysis of Variance of this design.

Exercise 12.4

Disk drive substrates may affect the amplitude of the signal obtained during readback. A manufacturer compares four substrates: aluminum (A), nickel-plated aluminum (B), and two types of glass (C and D). Sixteen disk drives will be made, four using each of the substrates. It is felt that operator, machine, and day of production may have an effect on the drives, so these three effects were blocked. The design and responses (in microvolts $\times 10^{-2}$) are given in the following table (data from Nelson 1993, data set `Substrates`); Greek letters indicate day:

Exercise 12.5

Machine	Operator							
	1		2		3		4	
1	Aα	8	Cγ	11	Dδ	2	Bβ	8
2	Cδ	7	Aβ	5	Bα	2	Dγ	4
3	Dβ	3	Bδ	9	Aγ	7	Cα	9
4	Bγ	4	Dα	5	Cβ	9	Aδ	3

Analyze these data and report your findings, including a description of the design.

Briefly describe the experimental design you would choose for each of the following situations. Report on treatments, units, blocks, and so on.

Problem 12.1

(a) We need to evaluate three "formulations" for a breakfast cereal targeted to boys. In fact, the cereals are all the same, they differ only in the cover art on the box. A market research firm arranges to have 150 boys between the ages of 5 and 8 years to serve as judges. Each boy is served a bowl of cereal with the box left on the table for his viewing. After the cereal is eaten, the boy is interviewed for his liking of the cereal. We do not anticipate age differences within this narrow age range.

(b) Land use along stream banks can dramatically affect the stream-water quality, so farmers have been adopting various buffers between their fields and streams. The Fish and Wildlife Service wishes to compare how three different buffers affect macro invertebrate populations 100 meters downstream from the buffer. Six different streams have been selected for

experimentation, and we expect considerable stream to stream variation in the invertebrate populations. We have funds to implement fifteen different buffers and measure the invertebrate populations downstream. We may put up to five (well-separated) buffers on a given stream.

(c) Carbon nanotubes can produce electricity when heated. We wish to determine how this feature might be used to produce batteries. We wish to vary two factors: the temperature to which the tubes are heated (low and high) and the speed at which the tubes are brought to temperature (slow and fast). Tubes are available from four different laboratories, and there could be differences between the tubes produced in different labs. In addition, these tubes are darned expensive, so we will need to reuse them in our experiments. However, no one really knows how multiple applications of heat treatments will affect the electricity produced; it might change over uses.

(d) Currently all counties place children under the care of the state into foster homes, but there is some support for reviving residential (group) homes. Suppose that the state wishes to conduct an experiment to compare outcomes under the two models of child care. The experiment will be conducted in six counties: Anoka and Washington (two suburban counties in the Twin Cities area), St. Louis and Sterns (two counties with moderate cities), and Murray and Jackson (two rural counties in southwest Minnesota). A county as a whole must continue to use foster homes or switch to residential homes. The response will be measured after 10 years by looking at the outcomes of the children in the counties.

(e) The polymerase chain reaction is used to make enough copies of a tiny segment of DNA to enable the segment to be studied. This enables DNA fingerprinting, the determination of paternity, and other applications. For use in field biology, the procedure needs to be adapted to each species. Here we wish to determine paternity in shrikes (small raptors with the lovely habit of storing their prey by impaling it on thorns or barbed wire). We have four procedural steps that can be altered, and we have chosen two alternatives for each step. We can afford 32 attempts in this preliminary study. Each attempt corresponds to a blood sample taken from a feather of a shrike, and we can take no more than one feather per bird.

(f) We wish to study the effects of three factors on corn yields: nitrogen added, planting depth, and planting date. The nitrogen and depth factors have two levels, and the date factor has three levels. There are 24 plots available: twelve are in St. Paul, MN, and twelve are in Rosemount, MN.

(g) You manage a french fry booth at the state fair and wish to compare four brands of french fry cutters for amount of potato wasted. You sell a lot of fries and keep four fry cutters and their operators going constantly. Each day you get a new load of potatoes, and you expect some day to day variation in waste due to size and shape of that day's load. Different operators may also produce different amounts of waste. A full day's usage is needed to get a reasonable measure of waste, and you would like to finish in under a week.

(h) Ruminant animals may not be able to quickly utilize protein in their diets, so we are interested in dietary changes that make the protein available. We can vary the cereal source (oats or hay) and the protein source (soy or fish meal) in the diets. There are twelve lambs available for the experiment, and we expect fairly large animal to animal differences. Each diet must be fed to a lamb for at least 1 week before the protein uptake measurement is made. The measurement technique is safe and benign, so we may use each lamb more than once. We do not expect any carry-over (residual) effects from one diet to the next, but there may be effects due to the aging of the lambs.

(i) A Health Maintenance Organization wishes to test the effect of substituting generic drugs for name brand drugs on patient satisfaction. Satisfaction will be measured by questionnaire after the study. They decide to start small, using only one drug (a decongestant for which they have an analogous generic) and twenty patients at each of their five clinics. The patients at the different clinics are from rather different socioeconomic backgrounds, so some clinic to clinic variation is expected. Drugs may be assigned on an individual basis.

(j) A sociologist is developing a new questionaire and response scale (a weighted combination of the answers to the questions) to assess where an individual lies on the liberal to conservative spectrum in social attitudes. The new scale is supposed to match an existing scale, and we need to conduct an experiment to test the equality of average scores.

The best experiment would give both questionaires to many people, but that is infeasible; each subject will only receive one questionaire. Subjects will be students in introductory sociology classes at the U of M, and we have resources to question 80 students. We anticipate that students planning to major in sociology may have different attitudes from nonmajors. We also anticipate that older, nontraditional students could have different attitudes from traditional students. (Assume that there is no problem with finding subjects, obtaining their consent, or obtaining their answers.)

(k) Air flow through heating and air conditioning vents can become noisy if the vent system is not properly designed. This can be a problem for concert halls and similar rooms. The noise seems to depend mostly on the kind of "bend" or "elbow" that is used to form turns in the vent. Unfortunately, the noise of a given vent also seems to depend rather delicately on just exactly how the vent was assembled, not simply the overall design, so we can't tell how a design will work from a single vent. We have four designs to compare, and can afford to make 20 vents and measure them for noise.

(l) A consumer testing agency wishes to compare three brands of home bread-making machines for the quality and consistency of the bread that they produce. The machines will be used with premixed ingredient packets, and at least nine loaves will be needed from each machine. Testers believe that brand of ingredient premix may affect quality (there are three

brands available locally). Testers also believe that day of baking also affects quality (due to temperature, humidity and related environmental factors). All baking is done in the morning, so each machine can only be used once a day.

(m) One of the constant issues in retailing is price. Raising a price lowers sales volume. This might increase total revenue if volume decreases only slightly, or it might decrease total revenue if volume decreases more. Web retailers have the ability to offer different prices to different customers by constructing different web pages with different prices. We want to compare sales volumes for a popular video game when we set the price at four different levels. There are probably many ways in which customers differ, but we don't see the customers so we don't know how they differ (inter-site web tracking has made this assumption less and less likely, but make the assumption anyway). We want to have at least 100 page accesses for each of the four prices.

For each of the following, describe the design that was used, give a skeleton ANOVA, and indicate how you would test the various terms in the model.

Problem 12.2

(a) House plants add a touch of nature to the indoors. They can also be expensive, so we would like to find conditions that best enable them to grow. This experiment considers four treatments, which are the factor/level combinations of water (unfiltered or filtered) and fertilizer (recommended amount or 75% of recommended amount), on the growth of "snake" plants. Thirty-two identical pots with identical soils are laid out in a 4 by 8 (rows by columns) pattern on a table in the greenhouse. Thirty-two approximately equal height snake plants are then randomly placed in the pots. The four treatments are then randomized to the pots such that each treatment occurs once in each column and twice in each row. After eight weeks, the leaf area of each plant is measured as a response.

(b) A Collison nebulizer is an instrument that produces aerosols from a liquid solution. Compressed air draws the liquid up into airborne drops, from which the water then evaporates to form aerosolized solid-phase particles. We will study the average particle size as a function of the pressure of the compressed air and the concentration of salt in the solution. Air pressure is at two levels (10 or 15 psi), and concentration is at three levels (5g, 10g, or 15g of KCl in 50g of water).

The following procedure is repeated each Monday afternoon for three consecutive weeks. Six half-hour time slots are randomly assigned to the six combinations of pressure and concentration. The nebulizer is used with that combination to produce an aerosol, and we measure the average particle diameter for the setting. At the end of three weeks, we thus have 18 responses.

(c) St. John's Wort (SJW) is sold as an herbal supplement that "improves mood"; FDA rules do not permit vendors to claim that SJW cures depression, as the clinical evidence is still too skimpy. The following experiment is conducted. One hundred patients with previously untreated

clinical depression are divided at random into two groups of 50. Patients in the first group will be given SJW daily at the recommended dose (or portion as it is known in the herbal supplement literature) for six months. At the end of six months, the patients are evaluated for depression by a psychologist using a standard measurement scheme and yielding a "depression score". These patients will then switch to a daily dose of Prozac, a commercial antidepressant. After six months of Prozac therapy, the patients are again evaluated for depression. The other 50 patients have six months of Prozac followed by six months of SJW, again with an evaluation at the end of each six month period.

(d) All crude oil contains some volatile compounds, but the crude oil from the Bakken oil field in North Dakota is especially known for its large fraction of volatiles. In particular, sometimes Bakken crude seems prone to explosion and fire; this is widely considered to be a bad thing. We want to harvest the volatiles from the crude, and we have four different methods for doing so. Each week we get a new batch of crude oil from North Dakota, randomly divide it into four equally sized tanks, and then randomly assign our four methods to extract volatiles to the four tanks. As a response, we measure the mass of volatiles that we can extract. We do this for ten consecutive weeks.

(e) Recent political tensions make detection of radioisotopes in the atmosphere very exciting. However, recent detections of cesium-137 have been traced back to fallout from atmospheric bomb tests prior to the atmospheric test ban treaty. In particular, forest fires in areas that were subjected to fallout decades ago can put the radioisotopes back into the atmosphere. (Cesium acts like phosphorus, which is a natural plant nutrient and is absorbed into trees.)

We want to compare cesium-137 concentrations in forest fire smoke for different kinds of fires in pine forests: surface fires and crown fires. We have 48 experimental plots, two plots in each of 24 national forests. These forests are in different directions and distances from the Nevada and Utah atomic test ranges. At each forest, we randomly assign the two fire types to the two plots. During the burn, we sample the smoke for cesium-137.

(f) Researchers are concerned that food colorings and food additives may affect activity of children. There are 300 children in this experiment, all aged 8 years. Every day for 7 weeks, the children will receive a drink of grape juice (naturally purple). On weeks two through seven, the parents will fill out activity diaries to quantify the level of activity. For two randomly chosen weeks out of weeks two through seven, the purple drink will also contain purple food color and sodium benzoate (a preservative).

(g) A common stereotype is that "beautiful is good;" that is, attractive people are stereotyped as having good personality qualities and unattractive people are stereotyped as having bad personality qualities. To test this hypothesis in employment screening, four fictitious and reasonably

equivalent résumés are created (all are for women). We manipulate the "beautiful" nature of the résumés by including applicant pictures. We will use three pictures that have been rated as attractive, neutral, and unattractive, plus a control with no picture. These four résumés will be sent to six different personnel officers for initial screening for a job. For each personnel officer, the four résumés will be randomly assigned to the four picture treatments. The response is the rating given by the officer to each applicant.

(h) Birds will often respond to other birds that invade their territory. We are interested in the time it takes nesting red-shouldered hawks to respond to invading calls, and want to know if that time varies according to the type of intruder. We have two state forests that have red-shouldered hawks nesting. In each forest, we choose ten nests at random from the known nesting sites. At each nest, we play two prerecorded calls over a loud-speaker (several days apart). One call is a red-shouldered hawk call; the other call is a great horned owl call. The response we measure is the time until the nesting hawks leave the nest to drive off the intruder.

(i) The food science department conducts an experiment to determine if the level of fiber in a muffin affects how hungry subjects perceive themselves to be. There are twenty subjects—ten randomly selected males and ten randomly selected females—from a large food science class. Each subject attends four sessions lasting 15 minutes. At the beginning of the session, they rate their hunger on a 1 to 100 scale. They then eat the muffin. Fifteen minutes later they again rate their hunger. The response for a given session is the decrease in hunger. At the four sessions they receive two low-fiber muffins and two high-fiber muffins in random order.

(j) One of the problems encountered when restoring a wetland is that reed canary grass will take over and crowd out all other vegetation. We wish to compare eight treatments for their efficacy in keeping the fraction of reed canary grass down. The treatments are the factorial combinations of burning (yes or no), tilling (yes or no), and herbicide (yes or no). We have 16 plots, eight in a site that is always wet and eight in a site that sometimes gets a little dry. At each site we randomly assign the eight treatments to plots.

(k) A consumer testing agency is trying to compare four over-the-counter acne medications (creams). They have obtained 96 teenagers as subjects, and they expect considerable subject to subject variation. To combat this variation, they want each subject to use more than one medication. They feel that it is unrealistic to divide the faces into four small patches with a different cream for each patch, so they just divide each face into left- and right-hand halves. Each subject then uses two medications, one for the right-hand side of the face, and one for the left side. They keep a record of blemishes, and the response for each side of the face will be the total number of blemishes on that side in a six-week study period. The medications are assigned to the face halves at random subject to the

restrictions that each pair of medications is used for the same number of subjects.

(l) "Fat cats" does not apply just to politicians and businessmen; many domestic house cats are obese due to lack of activity. One possibility for increasing cat activity is to hide their food in toys. Four house cats were fitted with activity monitors on their collars and tested over two days. On one of the two days, a cat would be fed from a dish as normal. On the other day, the cat's food would be hidden in toy "mice" that the cat would need to find and tip over to get the food. Two of the cats were randomly selected to use dish feeding on the first day, then toys on the second day. The remaining two cats had the opposite order. It was suspected that the activity monitor might bother the cat on the first day while it was getting used to it.

Suppose that we have a Latin Square with the following ANOVA: **Problem 12.3**

```
              DF           SS           MS            F
columns       15       17.406       1.1604      4.98166
rows           3       7.1707       2.3902     10.26148
treatments     3       4.8905       1.6302      6.99845
Error         42       9.7832      0.23293
```

What is the relative efficiency of this design compared to an RCB that only blocked on rows?

Many people enjoy dipping crackers in their soup. This experiment explores the amount of soup that is absorbed by, or adhered to, dipping crackers. We consider three different types of soup: a condensed, cream-based soup (Cream of Mushroom, need to add water); a condensed water-based soup (Tomato, need to add water); and a ready-to-eat pre-made soup (Steak and Potatoes, no additional water needed). We also consider three serving temperatures: 100, 110, and 120 degrees F. **Problem 12.4**

In this experiment, nine bowls of soup will be prepared in random order, one each for the combinations of soup type and serving temperature. For each bowl, 200g (approximately) of the soup is added and the combination of soup and bowl is weighed. Then five unbroken saltine crackers are placed in the soup. After two minutes, the crackers are removed (along with any soup that clings to them). The soup and bowl are again weighed, the decrease in weight (in g) is response. The experiment is then repeated the next day.

The data for this experiment are in the table below (data from S. Kleba, data set `SoupCrackers`). Analyze these data to understand the effects of the different factors on absorbance.

	Replication 1			Replication 2		
	Cream	Water	Ready	Cream	Water	Ready
100°	47	39	38	42	40	38
110°	40	37	42	43	39	41
120°	47	39	42	49	36	42

Many professions have board certification exams. Part of the certification **Problem 12.5**

process for bank examiners involves a "work basket" of tasks that the examinee must complete in a satisfactory fashion in a fixed time period. New work baskets must be constructed for each round of examinations, and much effort is expended to make the workbaskets comparable (in terms of average score) from exam to exam. This year, two new work baskets (A and B) are being evaluated. We have three old work baskets (C, D, and E) to form a basis for comparison. We have ten paid examinees (1 through 6 are certified bank examiners, 7 through 9 are noncertified bank examiners nearing the end of their training, and 10 is a public accountant with no bank examining experience or training) who will each take all five tests. There are five graders who will each grade ten exams. We anticipate differences between the examinees and the graders; our interest is in the exams, which were randomized so that each examinee took each exam and each grader grades two of each exam.

The data follow (data set `BankExaminers`). The letter indicates exam. Scores are out of 100, and 60 is passing. We want to know if either or both of the new exams are equivalent to the old exams.

Student	Grader				
	1	2	3	4	5
1	68 D	65 A	76 E	74 C	76 B
2	68 A	77 E	84 B	65 D	75 C
3	73 C	85 B	72 D	68 E	62 A
4	74 E	76 C	57 A	79 B	64 D
5	80 B	71 D	76 C	59 A	68 E
6	69 D	75 E	81 B	68 A	68 C
7	60 C	62 D	62 E	66 B	40 A
8	70 B	55 A	62 C	57 E	40 D
9	61 E	67 C	53 A	63 D	69 B
10	37 A	53 B	31 D	48 C	33 E

Problem 12.6

Cell engineering attempts to insert new genes into DNA so that the daughter cells have certain properties, for example, production of therapeutic proteins. However, gene insertion does not occur in all of the cells, and antibiotics are used to select those cells where insertion was successful. This is done by inserting a gene for antibiotic resistance along with the gene of interest. After treatment, only those cells with successful insertion will survive.

When multiple insertions are done, different kinds of antibiotic resistance may be utilized corresponding to different antibiotics. One selection approach is to select cells surviving after antibiotic A, and then next select from those survivors the cells that survive antibiotic B. Alternatively, one may put all the cells together with both antibiotics A and B. The issue is that antibiotics work using different mechanisms, and their effects might interact in unexpected ways.

This experiment studies how puromycin and hygromycin work together to kill cells. We look at nine treatments, the factor/level combinations of puromycin at 0, 2, and 4 μg/mL (suggested concentration is 2) and hygromycin at 0, 150, and 300 μg/mL (suggested concentration is 200). On week 1 we prepare nine wells on a plate with HEK293 cells and randomly

assign these wells to the nine treatments. The plate is incubated, and then the live cells are separated and counted with the response reported as 1,000s of live cells per mL. Low densities mean that the antibiotics were effective. This process is repeated on week 2 with a second plate and on week 3 with a third plate.

Data are shown in the table below (data set `Antibiotics`, data from Min Lu). Analyze these data to determine the effects of the antibiotics on cell density.

	Puromycin/Hydromycin								
	0			2			4		
Plate	0	150	300	0	150	300	0	150	300
1	2180	2380	1790	471	877	828	166	760	801
2	2420	2030	1520	500	721	838	178	721	821
3	2260	2250	1650	367	928	767	193	826	787

Problem 12.7

An experiment was conducted to see how variety of soybean and crop rotation practices affect soybean productivity. There are two varieties used, Hodgson 78 and BSR191. These varieties are each used in four different 5-year rotation patterns with corn. The rotation patterns are (1) four years of corn and then soybeans (C-C-C-C-S), (2) three years of corn and then two years of soybeans (C-C-C-S-S), (3) soybean and corn alternation (S-C-S-C-S), and (4) five years of soybeans (S-S-S-S-S). Here we only analyze data from the fifth year.

This experiment was conducted twice in Waseca, MN, and twice in Lamberton, MN. Two groups of eight plots were chosen at each location. The first group of eight plots at each location was randomly assigned to the variety-rotation treatments in 1983. The second group was then assigned in 1984. Responses were measured in 1987 and 1988 (the fifth years) for the two groups.

The response of interest is the weight (g) of 100 random seeds from soybean plants (data from Whiting 1990, data set `Rotations`). Analyze these data and report your findings.

		Rotation pattern			
Location-Year	Variety	1	2	3	4
W87	1	155	151	147	146
	2	153	156	159	155
W88	1	170	159	157	168
	2	164	170	162	169
L87	1	142	135	139	136
	2	146	138	135	133
L88	1	170	155	159	173
	2	167	162	153	162

Problem 12.8

An experiment was conducted to determine how different soybean varieties compete against weeds. There were sixteen varieties of soybeans and three weed treatments: no herbicide, apply herbicide 2 weeks after planting

the soybeans, and apply herbicide 4 weeks after planting the soybeans. The measured response is weed biomass in kg/ha. There were two replications of the experiment—one in St. Paul, MN, and one in Rosemount, MN—for a total of 96 observations (data from Bussan 1995, data set `Herbicides`):

Variety	Herb. 2 weeks		Herb. 4 weeks		No herb.	
	R	StP	R	StP	R	StP
Parker	750	1440	1630	890	3590	740
Lambert	870	550	3430	2520	6850	1620
M89-792	1090	130	2930	570	3710	3600
Sturdy	1110	400	1310	2060	2680	1510
Ozzie	1150	370	1730	2420	4870	1700
M89-1743	1210	430	6070	2790	4480	5070
M89-794	1330	190	1700	1370	3740	610
M90-1682	1630	200	2000	880	3330	3030
M89-1946	1660	230	2290	2210	3180	2640
Archer	2210	1110	3070	2120	6980	2210
M89-642	2290	220	1530	390	3750	2590
M90-317	2320	330	1760	680	2320	2700
M90-610	2480	350	1360	1680	5240	1510
M88-250	2480	350	1810	1020	6230	2420
M89-1006	2430	280	2420	2350	5990	1590
M89-1926	3120	260	1360	1840	5980	1560

Analyze these data for the effects of herbicide and variety.

Plant shoots can be encouraged in tissue culture by exposing the cotyledons of plant embryos to cytokinin, a plant growth hormone. However, some shoots become watery, soft, and unviable; this is vitrification. An experiment was performed to study how the orientation of the embryo during exposure to cytokinin and the type of growth medium after exposure to cytokinin affect the rate of vitrification. There are six treatments, which are the factorial combinations of orientation (standard and experimental) and medium (three kinds). On a given day, the experimenters extract embryos from white pine seeds and randomize them to the six treatments. The embryos are exposed using the selected orientation for 1 week, and then go onto the selected medium. The experiment was repeated 22 times on different starting days. The response is the fraction of shoots that are normal (data from David Zlesak, data set `PlantShoots`):

Problem 12.9

Day	Medium 1		Medium 2		Medium 3	
	Exp.	Std.	Exp.	Std.	Exp.	Std.
1	.67	.34	.46	.26	.63	.40
2	.70	.42	.69	.42	.74	.17
3	.86	.42	.89	.33	.80	.17
4	.76	.53	.74	.60	.78	.53
5	.63	.71	.50	.29	.63	.29
6	.65	.60	.95	1.00	.90	.40
7	.73	.50	.83	.88	.93	.88
8	.94	.75	.94	.75	.80	1.00
9	.93	.70	.77	.50	.90	.80
10	.71	.30	.48	.40	.65	.30
11	.83	.20	.74	.00	.69	.30
12	.82	.50	.72	.00	.63	.30
13	.67	.67	.67	.25	.90	.42
14	.83	.50	.94	.40	.83	.33
15	1.00	1.00	.80	.33	.90	1.00
16	.95	.75	.76	.25	.96	.63
17	.47	.50	.71	.67	.67	.50
18	.83	.50	.94	.67	.83	.83
19	.90	.33	.83	.67	.97	.50
20	1.00	.50	.69	.25	.92	1.00
21	.80	.63	.63	.00	.70	.50
22	.82	.60	.57	.40	1.00	.50

Analyze these data and report your conclusions on how orientation and medium affect vitrification.

Problem 12.10 We have all seen videos of Mentos® inserted into bottles of Diet Coke®: instant geyser of carbonated stickiness. This experiment explores how the number of Mentos® tablets (1, 4, or 7) and time since opening the bottle (0 or 20 minutes after opening) affect the amount of beverage that is ejected.

We have 36 .5L bottles of soft drink; 12 each of Diet Sunkist®, Diet 7up®, and 7up®. It is possible that there are differences between type of beverage. The 12 bottles of each type are randomly assigned to the six factor/level combinations of time and number of tablets, two bottles to each combination. Bottles were held at the same temperature and handled carefully. In random order, each bottle was given its assigned treatment, and the amount (mL) of beverage remaining 30 seconds after the tablet drop was measured as response.

Data are in the table below (data set `MentosGeyser`, data from T. Steichen). Analyze these data for the effects of delay time and number of tablets.

					Tablets		
Type	Elapsed	1	1	4	4	7	7
DS	0	343	355	236	255	208	215
D7	0	269	298	178	172	165	159
7	0	359	346	225	225	206	199
DS	20	252	249	246	241	208	222
D7	20	387	398	249	265	253	250
7	20	395	407	269	256	225	230

An army rocket development program was investigating the effects of slant range and propellant temperature on the accuracy of rockets. The overall objective of this phase of the program was to determine how these variables affect azimuth error (that is, side to side as opposed to distance) in the rocket impacts.

Problem 12.11

Three levels were chosen for each of slant range and temperature. The following procedure was repeated on 3 days. Twenty-seven rockets are grouped into nine sets of three, which are then assigned to the nine factor-level combinations in random order. The three rockets in a group are fired all at once in a single volley, and the azimuth error recorded. (Note that meteorological conditions may change from volley to volley.) The data follow (Bicking 1958) (data set `Rockets`):

				Slant range/Days					
		1			2			3	
Temp.	1	2	3	1	2	3	1	2	3
1	-10	-22	-9	-5	-17	-4	11	-10	1
	-13	0	7	-9	6	13	-5	10	20
	14	-5	12	21	0	20	22	6	24
2	-15	-25	-15	-14	-3	14	-9	8	14
	-17	-5	2	15	-1	5	-3	-2	18
	7	-11	5	-11	-20	-10	20	-15	-2
3	-21	-26	-15	-18	-8	0	13	-5	-8
	-23	-8	-5	5	5	-13	-9	-18	3
	0	-10	0	-10	-10	3	-13	-3	12

Analyze these data and determine how slant range and temperature affect azimuth error. (Hint: how many experimental units per block?)

An experiment is conducted to study the effect of alfalfa meal in the diet of male turkey poults (chicks). There are nine treatments. Treatment 1 is a control treatment; treatments 2 through 9 contain alfalfa meal. Treatments 2 through 5 contain alfalfa meal type 22; treatments 6 through 9 contain alfalfa meal type 27. Treatments 2 and 6 are 2.5% alfalfa, treatments 3 and 7 are 5% alfalfa, treatments 4 and 8 are 7.5% alfalfa. Treatments 5 and 9 are also 7.5% alfalfa, but they have been modified to have the same calories as the control treatment.

Problem 12.12

The randomization is conducted as follows. Seventy-two pens of eight birds each are set out. Treatments are separately randomized to pens grouped 1–9, 10–18, 19–27, and so on. We do not have the response for pen 66. The response is average daily weight gain per bird for birds aged 7 to 14 days in g/day (data from Turgay Ergul, data set `TurkeyPoults`):

| | Pen Groups | | | | | | | |
Trt	1–9	10–18	19–27	28–36	37–45	46–54	55–63	64–72
1	23.63	19.86	24.00	22.11	25.38	24.18	23.43	18.75
2	20.70	20.02	23.95	19.13	21.21	20.89	23.55	22.89
3	19.95	18.29	17.61	19.89	23.96	20.46	22.55	17.30
4	21.16	19.02	19.38	19.46	20.48	19.54	19.96	20.71
5	23.71	16.44	20.71	20.16	21.70	21.47	20.44	22.51
6	20.38	18.68	20.91	23.07	22.54	21.73	25.04	23.22
7	21.57	17.38	19.55	19.79	20.77	18.36	20.32	21.98
8	18.52	18.84	22.54	19.95	21.27	20.09	19.27	20.02
9	23.14	20.46	18.14	21.70	22.93	21.29	22.49	

Analyze these data to determine the effects of the treatments on weight gain.

Problem 12.13

Implantable pacemakers contain a small circuit board called a substrate. Multiple substrates are made as part of a single "laminate." In this experiment, seven laminates are chosen at random. We choose eight substrate locations and measure the length of the substrates at those eight locations on the seven substrates. Here we give coded responses $(10,000 \times [response - 1.45]$, data from Todd Kerkow, data set `Laminates`).

| | Laminate | | | | | | |
Location	1	2	3	4	5	6	7
1	28	20	23	29	44	45	43
2	11	20	27	31	33	38	36
3	26	26	14	17	41	36	36
4	23	26	18	21	36	36	39
5	20	21	30	28	45	31	33
6	16	19	24	23	33	32	39
7	37	43	49	33	53	49	32
8	04	09	13	17	39	29	32

Analyze these data to determine the effect of location. (Hint: think carefully about the design.)

Problem 12.14

The oleoresin of trees is obtained by cutting a tapping gash in the bark and removing the resin that collects there. Acid treatments can also improve collection. In this experiment, four trees (*Dipterocarpus kerrii*) will be tapped seven times each. Each of the tappings will be treated with a different strength of sulfuric acid (0, 2.5, 5, 10, 15, 25, and 50% strength), and the resin collected from each tapping is the response (in grams, data from Bin Jantan, Bin Ahmad, and Bin Ahmad 1987, data set `Oleoresin`):

| | Acid strength (%) | | | | | | |
Tree	0	2.5	5	10	15	25	50
1	3	108	219	276	197	171	166
2	2	100	198	319	202	173	304
3	1	43	79	182	123	172	194
4	.5	17	33	78	51	41	70

Determine the effect of acid treatments on resin output; if acid makes a difference, which treatments are best?

Problem 12.15 Hormones can alter the sexual development of animals. This experiment studies the effects of growth hormone (GH) and follicle-stimulating hormone (FSH) on the length of the seminiferous tubules in pigs. The treatments are control, daily injection of GH, daily injection of FSH, and daily injection of GH and FSH. Twenty-four weanling boars are used, four from each of six litters. The four boars in each litter are randomized to the four treatments. The boars are castrated at 100 days of age, and the length (in meters!) of the seminiferous tubules determined as response (data from Swanlund *et al.* 1995, data set `Tubules`).

| | Litter | | | | | |
	1	2	3	4	5	6
Control	1641	1290	2411	2527	1930	2158
GH	1829	1811	1897	1506	2060	1207
FSH	3395	3113	2219	2667	2210	2625
GH+FSH	1537	1991	3639	2246	1840	2217

Analyze these data to determine the effects of the hormones on tubule length.

Problem 12.16 Shade trees in coffee plantations may increase or decrease the yield of coffee, depending on several environmental and ecological factors. Robusta coffee was planted at three locations in Ghana. Each location was divided into four plots, and trees were planted at densities of 185, 90, 70, and 0 trees per hectare. Data are the yields of coffee (kg of fresh berries per hectare) for the 1994-95 cropping season (data from Amoah, Osei-Bonsu, and Oppong 1997, data set `ShadedCoffee`):

Location	185	90	70	0
1	3107	2092	2329	2017
2	1531	2101	1519	1766
3	2167	2428	2160	1967

Analyze these data to determine the effect of tree density on coffee production.

Problem 12.17 A sensory experiment was conducted to determine if consumers have a preference between regular potato chips (A) and reduced-fat potato chips (B). Twenty-four judges will rate both types of chips; twelve judges will rate the chips in the order regular fat, then reduced fat; and the other twelve will have the order reduced fat, then regular fat. We anticipate judge to judge differences and possible differences between the first and second chips tasted.

The response is a liking scale, with higher scores indicating greater liking (data from Monica Coulter, data set `PotatoChips`):

							Judge					
Period	Chip	1	2	3	4	5	6	7	8	9	10	11	12
1	A	8	5	7	8	7	7	4	9	8	7	7	7
2	B	6	6	8	8	4	7	8	9	9	7	5	3

						Judge						
		13	14	15	16	17	18	19	20	21	22	23	24
1	B	4	6	6	7	6	4	8	6	7	6	8	7
2	A	7	8	7	8	4	8	7	7	7	8	8	8

Analyze these data to determine if there is a difference in liking between the two kinds of potato chips.

Question 12.1 Find conditions under which the estimated variance for a CRD based on RCB data is less than the naive estimate pooling sums of squares and degrees of freedom for error and blocks. Give a heuristic argument, based on randomization, suggesting why your relationship is true.

Question 12.2 The inspector general is coming, and an officer wishes to arrange some soldiers for inspection. In the officer's command are men and women of three different ranks, who come from six different states. The officer is trying to arrange 36 soldiers for inspection in a six by six square with one soldier from each state-rank-gender combination. Furthermore, the idea is to arrange the soldiers so that no matter which rank or file (row or column) is inspected by the general, the general will see someone from each of the six states, one woman of each rank, and one man of each rank. Why is this officer so frustrated?

Chapter 13

Incomplete Block Designs

Block designs group similar units into blocks so that variation among units within the blocks is reduced. Complete block designs, such as RCB and LS, have each treatment occurring once in each block. Incomplete block designs also group units into blocks, but the blocks do not have enough units to accommodate all the treatments.

Incomplete block designs share with complete block designs the advantage of variance reduction due to blocking. The drawback of incomplete block designs is that they do not provide as much information per experimental unit as a complete block design with the same error variance. Thus complete blocks are preferred over incomplete blocks when both can be constructed with the same error variance.

Example 13.1 Eyedrops

Eye irritation can be reduced with eyedrops, and we wish to compare three brands of eyedrops for their ability to reduce eye irritation. (There are problems here related to measuring eye irritation, but we set them aside for now.) We expect considerable subject to subject variation, so blocking on subject seems appropriate. If each subject can only be used during one treatment period, then we must use one brand of drop in the left eye and another brand in the right eye. We are forced into incomplete blocks of size two, because our subjects have only two eyes.

Suppose that we have three subjects that receive brands (A and B), (A and C), and (B and C) respectively. How can we estimate the expected difference in responses between two treatments, say A and B? We can get some information from subject 1 by taking the difference of the A and B responses; the subject effect will cancel in this difference. This first difference has variance $2\sigma^2$. We can also get an estimate of A-B by subtracting the B-C difference in subject three from the A-C difference in subject two. Again, subject effects cancel out, and this difference has variance $4\sigma^2$. Similar approaches yield estimates of A-C and B-C using data from all subjects.

If we had had two complete blocks (three-eyed subjects?) with the same

unit variance, then we would have had two independent estimates of A-B each with variance $2\sigma^2$. Thus the incomplete block design has more variance in its estimates of treatment differences than does the complete block design with the same variance and number of units.

There are many kinds of incomplete block designs. This chapter will cover only some of the more common types. Several of the incomplete block designs given in this chapter have "balanced" in their name. It is important to realize that these designs are not balanced in the sense that all block and factor-level combinations occur equally often. Rather they are balanced using somewhat looser criteria that will be described later.

Two general classes of incomplete block designs are *resolvable* designs and *connected* designs. Suppose that each treatment is used r times in the design. A resolvable design is one in which the blocks can be arranged into r groups, with each group representing a complete set of treatments. Resolvable designs can make management of experiments simpler, because each replication can be run at a different time or a different location, or entire replications can be dropped if the need arises. The eyedrop example is not resolvable.

<div style="text-align: right">Resolvable designs split into replications</div>

A design is *disconnected* if you can separate the treatments into two groups, with no treatment from the first group ever appearing in the same block with a treatment from the second group. A *connected* design is one that is not disconnected. In a connected design you can estimate all treatment differences. You cannot estimate all treatment differences in a disconnected design; in particular, you cannot estimate differences between treatments in different groups. Connectedness is obviously a very desirable property.

<div style="text-align: right">Connected designs can estimate all treatment differences</div>

13.1   Balanced Incomplete Block Designs

The Balanced Incomplete Block Design (BIBD) is the simplest incomplete block design. We have g treatments, and each block has k units, with $k < g$. Each treatment will be given to r units, and we will use b blocks. The total number of units N must satisfy $N = kb = rg$. The final requirement for a BIBD is that all pairs of treatments must occur together in the same number of blocks. The BIBD is called "balanced" because the variance of the estimated

<div style="text-align: right">BIBD</div>

Table 13.1: Plates washed before foam disappears. Letters indicate treatments; data set `DirtyDishes`.

					Session						
1	2	3	4	5	6	7	8	9	10	11	12
A 19	D 6	G 21	A 20	B 17	C 15	A 20	B 16	C 13	A 20	B 17	C 14
B 17	E 26	H 19	D 7	E 26	F 23	E 26	F 23	D 7	F 24	D 6	E 24
C 11	F 23	J 28	G 20	H 19	J 31	J 31	G 21	H 20	H 19	J 29	G 21

difference of treatment effects $\widehat{\alpha}_i - \widehat{\alpha}_j$ is the same for all pairs of treatments i, j.

Example 13.1 is the simplest possible BIBD. There are $g = 3$ treatments, with blocks of size $k = 2$. Each treatment occurs $r = 2$ times in the $b = 3$ blocks. There are $N = 6$ total units, and each pair of treatments occurs together in one block.

We may use the BIBD design for treatments with factorial structure. For example, suppose that we have three factors each with two levels for a total of $g = 8$ treatments. If we have $b = 8$ blocks of size $k = 7$, then we can use a BIBD with $r = 7$, with each treatment left out of one block and each pair of treatments occurring together six times.

Example 13.2 Dish detergent

John (1961) gives an example of a BIBD. Nine different dishwashing solutions are to be compared. The first four consist of base detergent I and 3, 2, 1, and 0 parts of an additive; solutions five through eight consist of base detergent II and 3, 2, 1, and 0 parts of an additive; the last solution is a control. There are three washing basins and one operator for each basin. The three operators wash at the same speed during each test, and the response is the number of plates washed when the foam disappears. The speed of washing is the same for all three detergents used at any one session, but could differ from session to session.

Table 13.1 gives the design and the results. There are $g = 9$ treatments arranged in $b = 12$ incomplete blocks of size $k = 3$. Each treatment appears $r = 4$ times, and each pair of treatments appears together in one block.

The requirement that all pairs of treatments occur together in an equal number of blocks is a real stickler. Any given treatment occurs in r blocks, and there are $k - 1$ other units in each of these blocks for a total of $r(k - 1)$ units. These must be divided evenly between the $g - 1$ other treatments. Thus $\lambda = r(k - 1)/(g - 1)$ must be a whole number for a BIBD to exist. For the eyedrop example, $\lambda = 2(2 - 1)/(3 - 1) = 1$, and for the dishes example, $\lambda = 4(3 - 1)/(9 - 1) = 1$.

Treatment pairs occur together λ times

A major impediment to the use of the BIBD is that no BIBD may exist for your combination of $kb = rg$. For example, you may have $g = 5$ treatments and $b = 5$ blocks of size $k = 3$. Then $r = 3$, but $\lambda = 3(3 - 1)/(5 - 1) = 3/2$ is not a whole number, so there can be no BIBD for this combination of r, k, and g. Unfortunately, λ being a whole number is not sufficient to guarantee

that a BIBD exists, though one usually does.

A BIBD always exists for every combination of $k < g$. For example, you can always generate a BIBD by using all combinations of the g treatments taken k at a time. Such a BIBD is called *unreduced*. The problem with this approach is that you may need a lot of blocks for the design. For example, the unreduced design for $g = 8$ treatments in blocks of size $k = 4$ requires $b = 70$ blocks. Appendix B contains a list of some BIBD plans for $g \leq 9$. Fisher and Yates (1963) and Cochran and Cox (1957) contain much more extensive lists.

<div style="text-align:right">*Unreduced BIBD has all combinations*</div>

<div style="text-align:right">*BIBD tables*</div>

If you have a plan for a BIBD with g, k, and b blocks, then you can construct a plan for g treatments in b blocks of $g - k$ units per block simply by using in each block of the second design the treatments *not* used in the corresponding block of the first design. The second design is called the *complement* of the first design. When $b = g$ and $r = k$, a BIBD is said to be *symmetric*. The eyedrop example above is symmetric; the detergent example is not symmetric.

<div style="text-align:right">*Design complement*</div>

<div style="text-align:right">*Symmetric BIBD*</div>

Randomization of a BIBD occurs in three steps. First, randomize the assignment of physical blocks to subgroups of treatment letters (or numbers) given in the design. Second, randomize the assignment of these treatment letters to physical units within blocks. Third, randomize the assignment of treatment letters to treatments.

<div style="text-align:right">*BIBD randomization*</div>

13.1.1 Analysis for the BIBD

The model for the BIBD looks just like the model for the RCB. Let y_{ij} be the response for treatment i in block j. The difference for the BIBD is that we do not observe all i, j combinations. Use the model

<div style="text-align:right">*BIBD model*</div>

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \ .$$

Treatment effects α_i may be random or fixed, and block effects β_j may be random or fixed. As with the RCB, the assumption of additivity needs to be assessed.

Analysis of the BIBD differs from complete block designs such as RCB and LS in one important way: whether blocks were fixed or random did not change the results of the analysis for complete block designs, but the two approaches will lead to different results for incomplete block designs such as the BIBD. However, the differences between the two are minor in most instances.

<div style="text-align:right">*Fixed or random blocks?*</div>

When blocks are assumed to be fixed, the analysis is called an *intrablock* analysis. In effect, all of the information regarding treatment differences is assembled from differences taken between observations within a block. The discussion in Example 13.1 is a case of intrablock analysis: the estimate for treatment A minus treatment B is constructed from A-B within block 1 together with A-C within block 2 and B-C within block 3.

<div style="text-align:right">*Intrablock analysis*</div>

From an ANOVA perspective, intrablock analysis is simple: treatments adjusted for blocks. However, the simple formulae we have used for estimating treatment effects and sums of squares do *not* work for the BIBD; see the next section for what insight we can gain from the formulae that do work. This is analogous to the situation of an RCB with missing data: examine treatments adjusted for blocks, and let the software do the calculations.

Special formulae
for BIBD

When blocks are assumed to be random, there is additional information about treatments that can be extracted from the total response in each block; this is the *interblock* information. Combining the interblock information with the intrablock analysis is called interblock recovery. Interblock recovery happens automatically when you do REML analysis with random blocks[1]. The variability of the estimates from interblock information is almost always substantially greater than that for intrablock estimates leading to the combined estimates of treatment effects usually being pretty close to the intrablock estimates, and the variability of estimates after interblock recovery being only slightly smaller than the intrablock estimates. In fact, the better your blocking works, the less there is to be gained from interblock recovery (that is, most of the information is intrablock when block to block variance is high).

Interblock
recovery

Example 13.3 Dish detergent, continued

There are nine treatments and twelve blocks of size three in a BIBD. The basic intrablock analysis is treatments adjusted for blocks. We fit the model in line 1. The plots from line 2 (not shown) reveal somewhat short tails but no major issues.

---

[1] In ye olde days, we would do the intrablock analysis, then regress the block totals on dummy variables for the treatments included in the blocks, then estimate block variance relative to error variance, then form a weighted average of intrablock and interblock estimates.

```
1 > fit1 <- lm(plates~session+treat,DirtyDishes)
2 > plot(fit1)
3 > anova(fit1)
 Response: plates
            Df  Sum Sq Mean Sq F value    Pr(>F)
 session   11  412.75  37.523  45.533 6.028e-10 ***
 treat      8 1086.81 135.852 164.854 6.809e-14 ***
 Residuals 16   13.19   0.824
4 > summary(fit1)
 ...
 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)  19.41667    0.15130 128.334  < 2e-16 ***
 session1     -1.04630    0.55932  -1.871 0.079800 .
 ...
 session11    -0.41667    0.55932  -0.745 0.467106
 treat1        0.33333    0.49414   0.675 0.509574
 treat2       -2.22222    0.49414  -4.497 0.000366 ***
 treat3       -6.22222    0.49414 -12.592 1.02e-09 ***
 treat4      -12.88889    0.49414 -26.084 1.54e-14 ***
 treat5        5.88889    0.49414  11.918 2.27e-09 ***
 treat6        3.55556    0.49414   7.196 2.13e-06 ***
 treat7        1.66667    0.49414   3.373 0.003876 **
 treat8       -0.22222    0.49414  -0.450 0.658946
```

The ANOVA at line 3 shows treatments adjusted for blocks to be highly significant; it also shows a p-value for blocks, which we should ignore. The summary at line 4 shows that the standard error for an estimated treatment effect is 0.494.

```
5 > linear.contrast(fit1,treat,rep(c(.125,-1),c(8,1)))
 estimates        se   t-value       p-value   lower-ci   upper-ci
   -11.375 0.5559027 -20.46222 6.725878e-13 -12.55346 -10.19654
6 > linear.contrast(fit1,treat,c(.2,.2,.2,.2,-.2,-.2,-.2,-.2,0))
 estimates        se   t-value       p-value  lower-ci  upper-ci
 -6.377778 0.2964814 -21.51156 3.102838e-13 -7.00629 -5.749265
7 > sidelines(pairwise(fit1,treat))

  b1-0 -12.889
  b1-1  -6.222
  b1-2  -2.222 |
  b2-0  -0.222 | |
  b1-3   0.333 | |
  b2-1   1.667   | |
  b2-2   3.556     | |
  b2-3   5.889       |
  ctrl  10.111
```

We are likely interested in whether the experimental treatments differ from control. The contrast at line 5 shows that the new treatments average 11.4 plates less than the control, a difference that is highly significant. What about the two different bases in the experimental treatments? The contrast at line 6 shows that the average for base 1 is 6.4 plates less than that for base 3, again highly significant. The pairwise comparisons at line 7 show that the bases barely overlap, and control is better than all of the experimental treatments.

Interblock analysis requires an assumption of random blocks, which cer-

tainly seems reasonable for this experiment.

```
 8 > fit2 <- lmer(plates~(1|session)+treat,DirtyDishes)
 9 > summary(fit2)
...
Random effects:
 Groups    Name         Variance Std.Dev.
 session   (Intercept)  0.05635  0.2374
 Residual               0.80437  0.8969
Number of obs: 36, groups:  session, 12

Fixed effects:
            Estimate Std. Error t value
(Intercept)  19.4167     0.1644 118.079
treat1        0.3333     0.4323   0.771
treat2       -2.6061     0.4323  -6.029
treat3       -6.1742     0.4323 -14.283
treat4      -12.9129     0.4323 -29.872
treat5        6.0569     0.4323  14.012
treat6        3.7955     0.4323   8.780
treat7        1.3787     0.4323   3.189
treat8       -0.1742     0.4323  -0.403
...
10 > Anova(fit2,test="F")
Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)

Response: plates
          F Df Df.res    Pr(>F)
treat 192.96  8  24.71 < 2.2e-16 ***
```

In line 8 we fit the model with block as random. The summary at line 9 shows that the variance between blocks is actually fairly small relative to the residual variance. However, even with the relatively small block to block variance, the improvement with interblock recovery is small: the standard error of a treatment effect is .432, which is only 13% less than the value of .494 from the intrablock analysis. Nevertheless, this is a less expensive improvement than collecting more data. The interblock ANOVA at line 10 shows a more significant effect (higher F and higher denominator degrees of freedom) than we saw for intrablock.

13.1.2 Efficiency for the BIBD

Just as we consider the relative efficiency of RCB to CRD, we can consider the relative efficiency of BIBD to RCB. Define $E_{\text{BIBD:RCB}}$ to be

Efficiency of BIBD to RCB

$$E_{\text{BIBD:RCB}} = \frac{g(k-1)}{(g-1)k} \ ,$$

where g is the number of treatments and k is the number of units per block. Observe that $E_{\text{BIBD:RCB}} < 1$, because $k < g$ in the BIBD. For the detergent example, $E_{\text{BIBD:RCB}} = 9 \times 2/(8 \times 3) = 3/4$.

The value $E_{\text{BIBD:RCB}}$ is the relative efficiency of the BIBD to an RCB with the same variance. One way to think about $E_{\text{BIBD:RCB}}$ is that every unit

in a BIBD is only worth $E_{\text{BIBD:RCB}}$ units worth of information in an RCB with the same variance. Thus while each treatment is used r times in a BIBD, the effective sample size is only $rE_{\text{BIBD:RCB}}$.

Note that we have defined $E_{\text{BIBD:RCB}}$ with the idea that we *could* create an RCB with the same error variance. If we could create an actual RCB with the same error variance, we would never use the BIBD, because the RCB would give us more power and narrower confidence intervals for the same sample size.

In practice, we can often find incomplete blocks with a smaller variance σ_{bibd}^2 than can be attained using complete blocks σ_{rcb}^2. We prefer the BIBD design over the RCB if

$$\frac{\sigma_{bibd}^2}{rE_{\text{BIBD:RCB}}} < \frac{\sigma_{rcb}^2}{r}$$

or

$$\frac{\sigma_{bibd}^2}{\sigma_{rcb}^2} < E_{\text{BIBD:RCB}} \ ;$$

in words, we prefer the BIBD if the reduction in variance more than compensates for the loss of efficiency. This comparison ignores adjustments for error degrees of freedom.

The relative efficiency also plays a role in the formulae for hand calculation for intrablock analysis in the BIBD; this role of efficiency gives us more insight into the BIBD. (The availability of these simple formulae helped make the BIBD attractive before computers.) The hand-calculation formulae for the BIBD use the effective sample size in place of the actual sample size. Let $\overline{y}_{\bullet j}$ be the mean response in the jth block; let $v_{ij} = y_{ij} - \overline{y}_{\bullet j}$ be the data with block means removed; and let $v_{i\bullet}$ be the sum of the v_{ij} values for treatment i (there are r of them). Then we have

$$\widehat{\alpha}_i = \frac{v_{i\bullet}}{rE_{\text{BIBD:RCB}}} \ ,$$

$$\text{SS}_{\text{Trt}} = \sum_{i=1}^{g}(rE_{\text{BIBD:RCB}})\widehat{\alpha}_i{}^2 \ ,$$

and

$$Var(\sum_i w_i\widehat{\alpha}_i) = \sigma^2 \sum_i \frac{w_i^2}{rE_{\text{BIBD:RCB}}} \ .$$

We can also use pairwise comparison procedures with the effective sample size.

13.2   Row and Column Incomplete Blocks

We use Latin Squares and their variants when we need to block on two sources of variation in complete blocks. We can use *Youden Squares* when

we need to block on two sources of variation, but cannot set up the complete blocks for LS designs. I've always been amused by this name, because Youden Squares are not square.

The simplest example of a Youden Square starts with a Latin Square and deletes one of the rows (or columns). The resulting arrangement has g columns and $g - 1$ rows. Each row is a complete block for the treatments, and the columns form an unreduced BIBD for the treatments. Here is a simple Youden Square formed from a four by four Latin Square:

Youden Squares are incomplete Latin Squares

A	B	C	D
B	A	D	C
C	D	A	B

A more general definition of a Youden Square is a rectangular arrangement of treatments, with the columns forming a BIBD and all treatments occurring an equal number of times in each row. In particular, any symmetric BIBD ($b = g$) can be rearranged into a Youden Square. For example, here is a symmetric BIBD with $g = b = 7$ and $r = k = 3$ arranged as a Youden Square:

Youden Square is BIBD on columns and RCB on rows

A	B	C	D	E	F	G
B	C	D	E	F	G	A
D	E	F	G	A	B	C

In Appendix B, thoses BIBD's that can be arranged as Youden Squares are so arranged.

The analysis of a Youden Square is a combination of the Latin Square and BIBD, as might be expected. Because both treatments and columns appear once in each row, row contrasts are orthogonal to treatment and column contrasts, and this makes computation a little easier. Youden Squares are also called *row orthogonal* for this reason. The intrablock ANOVA has terms for rows, columns, treatments (adjusted for columns), and error. Row effects and sums of squares are computed via the standard formulae, ignoring columns and treatments. Column sums of squares (unadjusted) are computed ignoring rows and treatments. Intrablock treatment effects and sums of squares are computed as for a BIBD with columns as blocks. Error sums of squares are found by subtraction. Interblock analysis of the Youden Square and the combination of inter- and intrablock information are exactly like the BIBD.

Row orthogonal designs

Intrablock analysis adjusts for rows and columns

Interblock analysis similar to BIBD

Example 13.4 Lithium in blood

We wish to compare the blood concentrations of lithium 12 hours after administering lithium carbonate, using either a 300 mg capsule, 250 mg capsule, 450 mg time delay capsule, or 300 mg solution. There are twelve subjects, each of whom will be used twice, 1 week apart. We anticipate that the responses will be different in the second week, so we block on subject and week. The response is the serum lithium level as shown in Table 13.2 (data from Westlake 1974, data set `Lithium`).

There are $g = 4$ treatments in $b = 12$ blocks of size $k = 2$, so that $r = 6$. We have $\lambda = 2$, $E = 2/3$, and each treatment appears three times in each week for a Youden Square.

It is probably safe to treat subjects as random. On the other hand, we might expect consistent differences from week 1 to 2, so we treat period as fixed. Line 1 fits the model. The plot from line 2 (not shown) reals a hint of non-additivity (this is actually more clear in the model that treats subjects as fixed). Line 3 refits with a log transformation of the response. The plot from line 4 shows the non-additivity has been fixed, but there is a hint of decreasing variance.

```
1 > fit <- lmer(concentration~period+treatment+(1|subject),Lithium)
2 > plot(fit)
3 > fit2 <- lmer(log(concentration)~period+treatment+(1|subject),Lithium)
4 > plot(fit2)
5 > summary(fit2)
...
 Random effects:
  Groups   Name        Variance Std.Dev.
  subject  (Intercept) 0.01759  0.1326
  Residual             0.03763  0.1940
 Number of obs: 24, groups:  subject, 12

 Fixed effects:
             Estimate Std. Error t value
 (Intercept)  5.33448    0.05508  96.842
 period1      0.15151    0.03960   3.826
 treatment1   0.03158    0.07488   0.422
 treatment2   0.09152    0.07488   1.222
 treatment3  -0.17819    0.07488  -2.380
6 > car::Anova(fit2,test="F")
 Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)

 Response: log(concentration)
               F Df  Df.res   Pr(>F)
 period   14.6393  1  8.1023 0.004927 **
 treatment 1.7051  3 12.4656 0.217005
```

The summary at line 5 shows us that subject variability is of the same order as residual variability, and the ANOVA at line 6 shows us that there is no evidence for differences between treatments (a pairwise comparison would say the same).

13.3   Partially Balanced Incomplete Blocks

BIBD's are great, but their balancing requirements may imply that the smallest possible BIBD for a given g and k is too big to be practical. For example, let's look for a BIBD for $g = 12$ treatments in incomplete blocks of size $k = 7$. To be a BIBD, $\lambda = r(k-1)/(g-1) = 6r/11$ must be a whole number; this implies that r is some multiple of 11. In addition, $b = rg/k = (11 \times m) \times 12/7$ must be a whole number, and that implies that b is a multiple of $11 \times 12 = 132$. So the smallest possible BIBD has $r = 77$,

BIBD's are too big for some g and k

Table 13.2: Serum levels of lithium (μEq/l) 12 hours after administration. Treatments are 300 mg and 250 mg capsules, 450 mg time delay capsule, and 300 mg solution.

			Subject			
Week	1	2	3	4	5	6
1	A 200	D 267	C 156	B 280	D 333	D 233
2	B 160	C 178	A 200	C 178	A 167	B 200
	7	8	9	10	11	12
1	B 320	B 320	C 111	A 333	A 233	C 244
2	A 200	D 200	D 133	D 200	C 178	B 160

$b = 132$, and $N = 924$. This is a bigger experiment that we are likely to run.

Partially Balanced Incomplete Block Designs (PBIBD) allow us to run incomplete block designs with fewer blocks than may be required for a BIBD. The PBIBD has g treatments and b blocks of k units each; each treatment is used r times, and there is a total of $N = gr = bk$ units. The PBIBD does not have the requirement that each pair of treatments occurs together in the same number of blocks. This in turn implies that not all differences $\widehat{\alpha}_i - \widehat{\alpha}_j$ have the same variance in a PBIBD.

PBIBD has $N = gr = bk$; some treatment pairs more frequent

Here is a sample PBIBD with $g = 12$, $k = 7$, $r = 7$, and $b = 12$. In this representation, each row is a block, and the numbers in the row indicate which treatments occur in that block.

Sample PBIBD

Block			Treatments				
1	1	2	3	4	5	8	10
2	2	3	4	5	6	9	11
3	3	4	5	6	7	10	12
4	1	4	5	6	7	8	11
5	2	5	6	7	8	9	12
6	1	3	6	7	8	9	10
7	2	4	7	8	9	10	11
8	3	5	8	9	10	11	12
9	1	4	6	9	10	11	12
10	1	2	5	7	10	11	12
11	1	2	3	6	8	11	12
12	1	2	3	4	7	9	12

We see, for example, that treatment 1 occurs three times with treatments 5 and 9, and four times with all other treatments.

The design rules for a PBIBD are fairly complicated:

Requirements for PBIBD

1. There are g treatments, each used r times. There are b blocks of size $k < g$. Of course, $bk = gr$. No treatment occurs more than once in a block.

2. There are m *associate classes*. Any pair of treatments that are ith associates appears together in λ_i blocks. We usually arrange the λ_i

Associate classes

values in decreasing order, so that first associates appear together most frequently.

3. All treatments have the same number of ith associates, namely ρ_i. ρ_i ith associates

4. Let A and B be two treatments that are ith associates, and let p^i_{jk} be the number of treatments that are jth associates of A and kth associates of B. This number p^i_{jk} does not depend on the pair of ith associates chosen. In particular, $p^i_{jk} = p^i_{kj}$.

The PBIBD is partially balanced, because the variance of $\widehat{\alpha}_i - \widehat{\alpha}_j$ depends upon whether i, j are first, second, or mth associates. The randomization of a PBIBD is just like that for a BIBD.   Randomize PBIBD like BIBD

Let's check the design given above and verify that it is a PBIBD. First note that $g = 12$, $k = 7$, $r = 7$, $b = 12$, and no treatment appears twice in a block. Next, there are two associate classes, with first associates appearing together four times and second associates appearing together three times. The pairs (1,5), (1,9), (2,6), (2,10), (3,7), (3,11), (4,8), (4,12), (5,9), (6,10), (7,11), and (8,12) are second associates; all other pairs are first associates. Each treatment has nine first associates and two second associates. For any pair of first associates, there are six other treatments that are first associates of both, four other treatments that are first associates of one and second associates of the other (two each way), and no treatments that are second associates of both. We thus have

$$\{p^1_{ij}\} = \left[\begin{array}{cc} 6 & 2 \\ 2 & 0 \end{array} \right] \quad .$$

For any pair of second associates, there are nine treatments that are first associates of both, and one treatment that is a second associate of both, so that

$$\{p^2_{ij}\} = \left[\begin{array}{cc} 9 & 0 \\ 0 & 1 \end{array} \right] \quad .$$

Thus all the design requirements are met, and the example design is a PBIBD.

One historical advantage of the PBIBD was that the analysis could be done by hand. That is, there are (relatively) simple expressions for the various intra- and interblock analyses. With computers, that particular advantage is no longer very important. The intrablock analysis of the PBIBD is simply treatments adjusted for blocks, as with the BIBD.   Intrablock analysis is treatments adjusted for blocks

The efficiency of a PBIBD is actually an average efficiency. The variance of $\widehat{\alpha}_i - \widehat{\alpha}_j$ depends on whether treatments i and j are first associates, second associates, or whatever. So to compute efficiency $E_{\text{PBIBD:RCB}}$, we divide the variance obtained in an RCB for a pairwise difference ($2\sigma^2/r$) by the average of the variances of all pairwise differences in the PBIBD. There is an algorithm to determine $E_{\text{PBIBD:RCB}}$, but there is no simple formula. We can say that the efficiency will be less than $g(k-1)/[(g-1)k]$, which is the efficiency of a BIBD with the same block size and number of treatments.   PBIBD less efficient on average than BIBD

There are several extensive catalogues of PBIBD's, including Bose, Clatworthy, and Shrikhande (1954) (376 separate designs) and Clatworthy (1973).

13.4   Cyclic Designs

Cyclic designs are easily constructed incomplete block designs that permit the study of g treatments in blocks of size k. We will only examine the simplest situation, where the replication r for each treatment is a multiple of k, the block size. So $r = mk$, and $b = mg$ is the number of blocks. Cyclic designs include some BIBD and PBIBD designs.

<div style="float:right; text-align:right;">Cyclic designs
are simple</div>

A cycle of treatments starts with an initial treatment and then proceeds through the subsequent treatments in order. Once we get to treatment g, we go back down to treatment 1 and start increasing again. For example, with seven treatments we might have the cycle (4, 5, 6, 7, 1, 2, 3).

<div style="float:right; text-align:right;">Cycles of
treatments</div>

Cyclic construction starts with an initial block and builds $g - 1$ more blocks from the initial block by replacing each treatment in the initial block by its successor in the cycle. Additional sets of g blocks are constructed from new initial blocks. Thus all we need to know to build the design are the initial blocks.

<div style="float:right; text-align:right;">Proceed through
cycles from initial
block</div>

Write the initial block in a column, and write the cycles for each treatment in the initial block in rows, obtaining a k by g arrangement. The columns of this arrangement are the blocks. For example, suppose we have seven treatments and the initial block [1,4]. The cyclic design has blocks (columns):

1	2	3	4	5	6	7
4	5	6	7	1	2	3

Each row is a cycle started by a treatment in the initial block. Cycles are easy, so cyclic designs are easy, once you have the initial block.

But wait, there's more! Not only do we have an incomplete block design with the columns as blocks, we have a complete block design with the rows as blocks. Thus cyclic designs are row orthogonal designs (and may be Youden Squares if the cyclic design is BIBD).

<div style="float:right; text-align:right;">Cyclic designs
are row
orthogonal</div>

Appendix B.3 contains a table of initial blocks for cyclic designs for k from 2 through 10 and g from 6 through 15. Several initial blocks are given for the smaller designs, depending on how many replications are required. For example, for $k = 3$ the table shows initial blocks for 3, 6, and 9 replications. Use the first initial block if $r = 3$, use the first and second initial blocks if $r = 6$, and use all three initial blocks if $r = 9$. For $g = 10$, $k = 3$, and $r = 6$, the initial blocks are (1,2,5) and (1,3,8), and the plan is

1	2	3	4	5	6	7	8	9	10
2	3	4	5	6	7	8	9	10	1
5	6	7	8	9	10	1	2	3	4

1	2	3	4	5	6	7	8	9	10
3	4	5	6	7	8	9	10	1	2
8	9	10	1	2	3	4	5	6	7

As with the PBIBD, there is an algorithm to compute the (average) efficiency of a cyclic design, but there is no simple formula. The initial blocks

given in Appendix B.3 were chosen to make the cyclic designs as efficient as possible.

13.5   Square, Cubic, and Rectangular Lattices

Lattice designs work when the number of treatments g and the size of the blocks k follow special patterns. Specifically,

Lattice designs for special g, k combinations

- A Square Lattice can be used when $g = k^2$.

- A Cubic Lattice can be used when $g = k^3$.

- A Rectangular Lattice can be used when $g = k(k + 1)$.

These lattice designs are resolvable and are most useful when we have a large number of treatments to be run in small blocks.

We illustrate the Square Lattice when $g = 9 = 3^2$. Arrange the nine treatments in a square; for example:

$$
\begin{array}{ccc}
1 & 2 & 3 \\
4 & 5 & 6 \\
7 & 8 & 9
\end{array}
$$

There is nothing special about this pattern; we could arrange the treatments in any way. The first replicate of the Square Lattice consists of blocks made up of the rows of the square: here (1, 2, 3), (4, 5, 6), and (7, 8, 9). The second replicate consists of blocks made from the columns of the square: (1, 4, 7), (2, 5, 8), and (3, 6, 9). A Square Lattice must have at least these two replicates to be connected, and a Square Lattice with only two replicates is called a *simple lattice*.

A simple lattice has two replications made of rows and columns of the square

We add a third replication using a Latin Square. A Square Lattice with three replicates is called a *triple lattice*. Here is a three by three Latin Square:

Triple lattice uses Latin Square for third replicate

$$
\begin{array}{ccc}
A & B & C \\
B & C & A \\
C & A & B
\end{array}
$$

Assign treatments to blocks using the letter patterns from the square. The three blocks of the third replicate are (1, 6, 8), (2, 4, 9), and (3, 5, 7).

You can construct additional replicates for every Latin Square that is orthogonal to those already used. For example, the following square

Additional replicates use orthogonal Latin Squares

$$
\begin{array}{ccc}
A & B & C \\
C & A & B \\
B & C & A
\end{array}
$$

is orthogonal to the first one used. Our fourth replicate is thus (1, 5, 9), (2, 6, 7), and (3, 4, 8). Recall that there are no six by six Graeco-Latin Squares (six by six orthogonal Latin Squares), so only simple and triple lattices are possible for $g = 6^2$.

For $g = k^2$, there are at most $k-1$ orthogonal Latin Squares. The Square Lattice formed when $k - 1$ Latin Squares are used has $k + 1$ replicates; is called a *balanced lattice*; and is a BIBD with $g = k^2$, $b = k(k+1)$, $r = k+1$, $\lambda = 1$, and $E = k/(k + 1)$. The BIBD plan for $g = 9$ treatments in $b = 12$ blocks of size $k = 3$, given in Appendix B, is exactly the balanced lattice constructed above.

Balanced Lattice ($k + 1$ replicates) is a BIBD

The (average) efficiency of a Square Lattice relative to an RCB is

$$E_{\text{SL:RCB}} = \frac{(k + 1)(r - 1)}{(k + 1)(r - 1) + r} \quad.$$

This is the best possible efficiency for any resolvable design.

The *Rectangular Lattice* is closely related to the Square Lattice. Arrange the $g = k(k+1)$ treatments in an $(k + 1) \times (k + 1)$ square with the diagonal blank, for example:

Rectangular Lattice is subset of a square

$$
\begin{array}{cccc}
\bullet & 1 & 2 & 3 \\
4 & \bullet & 5 & 6 \\
7 & 8 & \bullet & 9 \\
10 & 11 & 12 & \bullet
\end{array}
$$

As with the Square Lattice, the first two replicates are formed from the rows and columns of this arrangement, ignoring the diagonal: (1, 2, 3), (4, 5, 6), (7, 8, 9), (10, 11, 12), (4, 7, 10), (1, 8, 11), (2, 5, 12), (3, 6, 9). Additional replicates are formed from the letters of orthogonal Latin Squares that satisfy the extra constraints that all the squares have the same diagonal and all letters appear on the diagonal; for example:

Rows, columns, and Latin Squares for a Rectangular Lattice

$$
\begin{array}{cccc|cccc}
A & B & C & D & A & C & D & B \\
C & D & A & B & B & D & C & A \\
D & C & B & A & C & A & B & D \\
B & A & D & C & D & B & A & C
\end{array}
$$

These squares are orthogonal and share the same diagonal containing all treatments. The next two replicates for this Rectangular Lattice design are thus (5, 9, 11), (1, 6, 10), (2, 4, 8), (3, 7, 12) and (6, 8, 12), (3, 4, 11), (1, 5, 7), (2, 9, 10).

The *Cubic Lattice* is a generalization the Square Lattice. In the Square Lattice, each treatment can be indexed by two subscripts i, j, with $1 \leq i \leq k$ and $1 \leq j \leq k$. The subscript i indexes rows, and the subscript j indexes columns. The first row in the Square Lattice is all those treatments with $i = 1$. The second column is all those treatments with $j = 2$. The blocks of the first replicate of a Square Lattice are rows; that is, treatments are the same block if they have the same i. The blocks of the second replicate of the Square Lattice are columns; that is, treatments are in the same block if they have the same j.

Cubic Lattice for k^3 treatments in blocks of k

For the Cubic Lattice, we have $g = k^3$ treatments that we index with three subscripts i, j, l, with $1 \leq i \leq k$, $1 \leq j \leq k$, and $1 \leq l \leq k$. Each replicate of the Cubic Lattice will be k^2 blocks of size k. In the first

Form blocks by keeping two subscripts constant

replicate of a Cubic Lattice, treatments are grouped so that all treatments in a block have the same values of i and j. In the second replicate, treatments in the same block have the same values of i and l, and in the third replicate, treatments in the same block have the same values of j and l. For example, when $g = 8 = 2^3$, the cubic lattice will have four blocks of size two in each replicate. These blocks are as follows (using the ijl subscript to represent a treatment):

Replicate 1	Replicate 2	Replicate 3
(111, 112)	(111, 121)	(111, 211)
(121, 122)	(112, 122)	(112, 212)
(211, 212)	(211, 221)	(121, 221)
(221, 222)	(212, 222)	(122, 222)

Cubic Lattice designs can have 3, 6, 9, and so forth replicates by repeating this pattern.

The intrablock Analysis of Variance for a Square, Cubic, or Rectangular Lattice is analogous to that for the BIBD; namely, treatments should be adjusted for blocks.

<div style="text-align: right;">Treatments adjusted for blocks</div>

13.6   Alpha Designs

Alpha Designs allow us to construct resolvable incomplete block designs when the number of treatments g or block size k does not meet the strict requirements for one of the lattice designs. Alpha Designs require that the number of treatments be a multiple of the block size $g = mk$, so that there are m blocks per replication and $b = rm$ blocks in the complete design.

<div style="text-align: right;">Alpha Designs are resolvable with $g = mk$</div>

We construct an Alpha Design in three steps. First we obtain the "generating array" for k, m, and r. This array has k rows and r columns. Next we expand each column of the generating array to m columns using a cyclic pattern to obtain an "intermediate array" with k rows and mr columns. Finally we add m to the second row of the intermediate array, $2m$ to the third row, and so on. Columns of the final array are blocks.

<div style="text-align: right;">Three-step construction</div>

Section B.4 has generating arrays for m from 5 to 15, k at least four but no more than the minimum of m and $100/m$, and r up to four. The major division is by m, so first find the full array for your value of m. We only need the first k rows and r columns of this full tabulated array.

<div style="text-align: right;">Finding the generating array</div>

For example, suppose that we have $g = 20$ treatments and blocks of size $k = 4$, and we desire $r = 2$ replications. Then $m = 5$ and $b = 10$. The full generating array for $m = 5$ from Section B.4 is

$$
\begin{array}{cccc}
1 & 1 & 1 & 1 \\
1 & 2 & 5 & 3 \\
1 & 3 & 4 & 5 \\
1 & 4 & 3 & 2 \\
1 & 5 & 2 & 4 \\
\end{array}
$$

We only need the first $k = 4$ rows and $r = 2$ columns, so our generating array is

$$
\begin{array}{cc}
1 & 1 \\
1 & 2 \\
1 & 3 \\
1 & 4
\end{array}
$$

Step two takes each column of the generating array and does cyclic substitution with $1, 2, \ldots, m$, to get m columns. So, for our array, we get

Construct intermediate array

$$
\begin{array}{ccccc@{\qquad}ccccc}
1 & 2 & 3 & 4 & 5 & 1 & 2 & 3 & 4 & 5 \\
1 & 2 & 3 & 4 & 5 & 2 & 3 & 4 & 5 & 1 \\
1 & 2 & 3 & 4 & 5 & 3 & 4 & 5 & 1 & 2 \\
1 & 2 & 3 & 4 & 5 & 4 & 5 & 1 & 2 & 3
\end{array}
$$

The first five columns are from the first column of the generating array, and the last five columns are from the last column of the generating array. This is the intermediate array.

Finally, we take the intermediate array and add $m = 5$ to the second row, $2m = 10$ to the third row, and $3m = 15$ to the last row, obtaining

Add multiples of m to rows

$$
\begin{array}{ccccc@{\qquad}ccccc}
1 & 2 & 3 & 4 & 5 & 1 & 2 & 3 & 4 & 5 \\
6 & 7 & 8 & 9 & 10 & 7 & 8 & 9 & 10 & 6 \\
11 & 12 & 13 & 14 & 15 & 13 & 14 & 15 & 11 & 12 \\
16 & 17 & 18 & 19 & 20 & 19 & 20 & 16 & 17 & 18
\end{array}
$$

This is our final design, with columns being blocks and numbers indicating treatments.

The Alpha Designs constructed from the tables in Section B.4 are with a few exceptions the most efficient Alpha Designs possible. The average efficiencies for these Alpha Designs are very close to the theoretical upper bound for average efficiency of a resolvable design, namely

$$
E_{\alpha:\text{RCB}} \leq \frac{(g-1)(r-1)}{(g-1)(r-1) + r(m-1)} \ .
$$

13.7   Further Reading and Extensions

Incomplete block designs have been the subject of a great deal of research and theory; we have mentioned almost none of it. Two excellent sources for more theoretical discussions of incomplete blocks are John (1971) and John and Williams (1995). Among the topics relevant to this chapter, John (1971) describes recovery of interblock information for BIBD, PBIBD, and general incomplete block designs; existence and construction of BIBD's; classification, existence, and construction of PBIBD's; and efficiency. John and Williams (1995) is my basic reference for Cyclic Designs, Alpha Designs, and incomplete block efficiencies; and it has a good deal to say about row column designs, interblock information, and other topics as well.

Most of the designs described in this chapter are not recent. Many of these incomplete block designs were introduced by Frank Yates in the late 1930's, including BIBD's (Yates 1936a), Square Lattices (Yates 1936b), and Cubic Lattices (Yates 1939), as well other designs such as Lattice Squares (different from a Square Lattice, Yates 1940). PBIBD's first appear in Bose and Nair (1939). Alpha Designs are the relative newcomers, first appearing in Patterson and Williams (1976).

John and Williams (1995) provide a detailed discussion of the efficiencies of incomplete block designs, including a proof that the BIBD has the highest possible efficiency for equally replicated designs with equal block sizes. Section 3.3 of their book gives an expression for the efficiency of a cyclic design; Sections 2.8 and 4.10 give a variety of upper bounds for the efficiencies of blocked designs and resolvable designs. Chapter 12 of John (1971) and Chapter 1 of Bose, Clatworthy, and Shrikhande (1954) describe efficiency of PBIBD's.

Some experimental situations will not fit into any of the standard design categories. For example, different treatments may have different replication, or blocks may have different sizes. Computer software exists that will search for "optimal" allocations of the treatments to units. *Optimal* can be defined in several ways; for example, you could choose to minimize the average variance for pairwise comparisons. See Silvey (1980) and Cook and Nachtsheim (1989).

13.8   Problems

Consider the following incomplete block experiment with nine treatments (A-I) in nine blocks of size three (data set IBD.   **Exercise 13.1**

			Block					
1	2	3	4	5	6	7	8	9
C 54	B 35	A 48	G 46	D 61	C 52	A 54	B 45	A 31
H 56	G 36	G 42	H 56	E 61	I 53	H 59	I 46	B 28
D 53	D 40	E 43	I 59	F 54	E 48	F 62	F 47	C 25

(a) Identify the type of design.
(b) Analyze the data for differences between the treatments.

Chemical yield may be influenced by the temperature, pressure, and/or time in the reactor vessel. Each of these factors may be set at a high or a low level. Thus we have a 2^3 experiment. Unfortunately, the process feedstock is highly variable, so batch to batch differences in feedstock are expected; we must start with new feedstock every day. Furthermore, each batch of feedstock is only big enough for seven runs (experimental units). We have enough money for eight batches of feedstock. We decide to use a BIBD, with each of the eight factor-level combinations missing from one of the blocks.   **Exercise 13.2**

Give a skeleton ANOVA (source and degrees of freedom only), and describe an appropriate randomization scheme.

Briefly describe the following incomplete block designs (BIBD, or PBIBD with what associate classes, and so on).

Exercise 13.3

(a)

Block	1	2	3	4
	A	A	B	A
	B	C	C	B
	C	D	D	D

(b)

Block	1	2	3	4	5
	A	A	A	B	C
	B	B	C	D	D
	C	D	E	E	E

(c)

Block	1	2	3	4
	1	3	1	2
	2	4	3	4

We wish to compare the average access times of five brands of half-height computer disk drives (denoted A through E). We would like to block on the computer in which they are used, but each computer will only hold four drives. Average access times and the design are given in the following table (data from Nelson 1993, data set `DiskSpeed`):

Exercise 13.4

		Computer		
1	2	3	4	5
A 35	A 41	B 40	A 32	A 40
B 42	B 45	C 42	C 33	B 38
C 31	D 32	D 33	D 35	C 35
D 30	E 40	E 39	E 36	E 37

Analyze these data and report your findings, including a description of the design.

Briefly describe the experimental design you would choose for each of the following situations, and why.

Problem 13.1

(a) We wish to study "sensory specific satiety." This is the phenomenon wherein if you eat a lot of some food, then that food and similar foods become less appealing. In our case we are investigating four kinds of potato chips: classic, sour cream, barbecue, and cheese. Each subject will participate in several sessions. At each session a subject will eat a load food (one of the four kinds of chips). After eating the load food, the subject will rate his or her liking of each of the four kinds of chips. We anticipate large subject to subject differences. We also anticipate that ratings could differ from session to session (for example, we suspect that first session ratings could be higher than last session ratings). Each subject will be available for two sessions, and we have 24 subjects. Choose an appropriate design for this experiment.

(b) Animal waste (manure) management is an increasingly important problem in dairy farming. One current proposal is to extract methane from the manure before further processing the manure for other uses. (Sale of the methane as a fuel makes the whole process economically viable.) The methane extraction process needs to be tuned. In the present situation, we wish to study three temperatures to use in the extraction, and three moisture levels for the manure before its injection into the extractor (basically the manure is allowed to dry until it reaches the appropriate

moisture content). We have an essentially infinite supply of manure from each of several farms. However, we wish to finish the tuning using no more than 27 runs, and we anticipate that there may be yield differences in the manure depending on its farm of origin.

(c) It has long been known that mothers pass antibodies to their infants through breast milk. More recently, attention has focused on the antibacterial properties of oligosaccharides, which are indigestible sugars that are present in surprisingly large quantities in breast milk. It is believed that bacteria in the gut bind to the oligosaccharides rather than to the intestinal wall, thus reducing incidence of disease.

This experiment wishes to compare the effects of oligosaccharides on the incidence of disease (diarrhea) among human infants fed "formula" instead of breast milk. There will be three kinds of formula: control, control plus oligosaccharides extracted from human milk, and control plus oligosaccharides produced in the lab. We expect enormous genetic and environmental variability, so we'd really like to use identical triplets. However, those are exceedingly rare. What we can use is 18 pairs of identical twins.

(d) Competition cuts tree growth rate, so we wish to study the effects on tree growth of using four herbicides on the competition. There are many study sites available, but each site is only large enough for three plots. Resources are available for 24 plots (that is, eight sites with three plots per site). Large site differences are expected.

(e) Three treatments are being studied for the rehabilitation of acidified lakes. Unfortunately, there is tremendous lake to lake variability, and we only have six lakes on which we are allowed to experiment. We may treat each lake as a whole, or we may split each lake in two using a plastic "curtain" and treat the halves separately. Sadly, the technology does not allow us to split each lake into three.

(f) A retail bookstore has two checkouts, and thus two checkout advertising displays. These displays are important for enticing impulse purchases, so the bookstore would like to know which of the four types of displays available will lead to the most sales. The displays will be left up for one week, because it is expensive to change displays and you really need a full week to get sufficient volume of sales and overcome day-of-week effects; there are, however, week to week differences in sales. The store wishes to complete the comparison in at most 8 and preferably fewer weeks.

(g) We wish to compare four "dog collars." The thought is that some collars will lead to faster obedience than others. The response we measure will be the time it takes a dog to complete a walking course with lots of potential distractions. We have 24 dogs that can be used, and we expect large dog to dog variability. Dogs can be used more than once, but if they are used more than once there should be at least 1 week between trials. Our experiment should be completed in less than 3 weeks, so no dog could possibly be used more than three times.

(h) My family of four suffers from allergies, so we all take antihistamines of one sort or another. Our doctors have suggested four different potential medications, but we would like to choose one drug for all four of us to use. We (I) want to run an experiment to choose that drug optimally. Some constraints on the design include (a) we should each try all the drugs, (b) the doctors say that we need to take a drug for a month or so to get a reasonable idea of how well it works, (c) allergens change over time, and (d) we should complete the experiment in under six months. The response will be an "allergy symptom" index scored over the last week of usage for each mediation.

(i) Some trumpets sound better than others, and there are groups that claim that temperature treatments will improve the sound of a trumpet. Some groups advocate cryogenic freezing, whereas other groups advocate a heat treatment. We wish to compare the freezing treatment, the heat treatment, and a control of no treatment. A professional musician will play the instruments, which will be judged for sound by a panel of experts; the average of the experts scores will be the response for any unit.

Without a doubt, different models of trumpet sound different. Some instrument manufacturers have loaned us twelve trumpets, two from each of six models. We also have the time constraint that we can only use each instrument once.

(j) Recent research suggests that a mixture of caffeine and alcohol injected into the blood after stroke can reduce stroke damage by 80% (my wife suggests prophylaxis via Irish coffee). We wish to replicate their experiment and study their mixture, caffeine alone, alcohol alone, and a control. We can use 60 inbred rats, in which we can artificially induce stroke.

For each of the following, describe the experimental design that was used, and give a skeleton ANOVA.

Problem 13.2

(a) Tissue engineering attempts to mimic live tissue using a constructed product. In this case, we are producing tubular constructs by allowing free floating cells in a suspension to deposit on a fibrin gel wrapped around a tube. Once the cells are deposited, the tube is placed in a nutrient broth that allows the construct to grow to completion. We then measure tensile strength as a response.

In this experiment we use three concentrations of fibrin and two concentrations of cells for a total of six treatments, and we use each treatment twice for a total of twelve constructs. The jars of nutrients can only hold four tubes, so the experiment is run in three nutrient jars. Jar one holds treatments 1, 2, 3, and 4; jar two holds treatments 1, 2, 5, and 6; and jar three holds treatments 3, 4, 5, and 6.

(b) We wish to study how hives of bees react to odors of different concentrations. The idea is that we place an odor attractant of a given concentration 100 meters from the bee hive. We then count the number of bees that

visit the attractant in the first hour after it is put in place. The study uses six hives and is conducted on three consecutive days. Two hives get each of the three concentrations on each day, and all three concentrations are used for each hive; otherwise, the assignment of treatments is random.

(c) Does the cost of a gift reflect how much it is appreciated? We have three gifts: a CD (inexpensive), a coupon for dinner for two (moderate), and an iPhone (expensive). We use engaged couples. The partners are separated; each partner is asked to complete a few neutral tasks that have nothing to do with the experiment, and then each is told that their (other) partner has selected a thank you gift for them, which will be one of the three above. The response is how much each partner appreciates the gift. For each couple, we use two different gifts, randomly assigned to the partners. This experiment used 30 engaged couples (60 individuals in total), and each pair of gifts is used 10 times.

(d) It has been believed that there is a link between the tuberculosis in cattle and tuberculosis in badgers (little furry ones, not necessarily from Wisconsin). The British government sponsored a multiyear experiment to test this hypothesis. Suppose that the experiment went as follows. Forty pairs of pastures (eighty total fields) are selected around Britain. The members of any pair of pastures are geographically near to each other, approximately equal in size, approximately equal is slope and aspect, and have approximately equal cattle densities (animals per hectare) grazing. One pasture in each pair is selected at random, and in that pasture traps are set out to reduce the badger population by 75%; this will need to be done over and over again over time. After six years, we take as response for each pasture the fraction of cattle that have been infected with TB. (Note: badgers have been a protected species in Britain for nearly 40 years, so the real experiment was not undertaken lightly. Predictably, it's results were apparently ignored by the government.)

(e) *The Fellowship of the Ring*, *The Mummy*, and *The Phantom Menace* are three movies that feature hordes of computer generated bad guys. In an effort to get a real comparison rating, 24 teenage boys will screen the movies and give ratings. However, the movies are long and so each boy will only watch two movies, one in the afternoon and one in the evening. The movies are assigned at random subject to the restrictions that each pair of movies is used the same number of times and each movie is used an equal number of times in the afternoon and evening.

(f) Most faucets are made of brass, which is an alloy that contains lead. Lead is toxic, and we don't want any leaching out of the plumbing fixture into our water (actually, up to 11 ppb is allowed). We wish to determine if disinfectant (chlorine or chloramine) or alkalinity (high or low levels of sodium bicarbonate) affect the amount of lead leaching out of the faucet. We go to the building supply center and buy four faucets, one from each of four manufacturers. We anticipate that there may be manufacturer to manufacturer differences. We also anticipate that early use of the faucet may have different lead levels from later use.

(g) Plant breeders wish to study six varieties of corn. They have 24 plots available, four in each of six locations. The varieties are assigned to location as follows (there is random assignment of varieties to plot within location):

		Locations			
1	2	3	4	5	6
A	B	A	A	B	A
B	C	C	B	C	C
D	E	D	D	E	D
E	F	F	E	F	F

(h) We wish to study gender bias in paper grading. We have 12 "lower" level papers and 12 "advanced" level papers. There are four paid graders who do not know the students or their names. Each paper is submitted for grading exactly once (that is, no paper is graded by more than one grader). We examine gender bias by the name put on the paper: either a male first name, a female first name, or just initials. The twelve lower-level papers are assigned at random to the combinations of grader and name gender, as are the advanced-level papers. The response we measure is the grade given (on a 0-100 scale).

(i) Song bird abundance can be measured by sending trained observers to a site to listen for the calls of the birds and make counts. Consider an experiment on the effects of three different forest harvesting techniques on bird abundance. There are six forests and two observers, and there will be two harvests in each of the six forests. The harvest techniques were assigned in the following way:

			Forest			
Observer	1	2	3	4	5	6
1	A	C	B	B	A	C
2	C	A	A	C	B	B

(d) Wafer board is a manufactured wood product made from wood chips. One potential problem is warping. Consider an experiment where we compare three kinds of glue and two curing methods. All six combinations are used four times, once for each of four different batches of wood chips. The response is the amount of warping.

Japanese beetles ate the Roma beans in our garden last year, so we ran an experiment this year to learn the best pesticide. We have six garden beds with beans, and the garden store has three different sprays that claim to keep the beetles off the beans. Sprays drift on the wind, so we cannot spray very small areas. We divide each garden bed into two plots and use a different spray on each plot. Below are the numbers of beetles per plot (data set `JapaneseBeetles`).

Problem 13.3

| | Bed | | | | | |
	1	2	3	4	5	6
	19 A	9 A	25 B	9 A	26 A	13 B
	21 B	16 C	30 C	11 B	33 C	18 C

Analyze these data to determine the effects of sprays. Which one should we use?

Milk can be strained through filter disks to remove dirt and debris. Filters are made by surface-bonding fiber webs to both sides of a disk. This experiment is concerned with how the construction of the filter affects the speed of milk flow through the filter.

Problem 13.4

We have a 2^4 factorial structure for the filters. The factors are fiber weight (normal or heavy), loft (thickness of the filter, normal or low), bonding solution on bottom surface (A or B), and bonding solution on top surface (A or B). Note the unfortunate fact that the "high" level of the second factor, loft, is low loft. Treatments 1 through 16 are the factor-level combinations in standard order.

These are speed tests, so we pour a measured amount of milk through the disk and record the filtration time as the response. We expect considerable variation from farm to farm, so we block on farm. We also expect variation from milking to milking, so we want all measurements at one farm to be done at a single milking. However, only three filters can be satisfactorily used at a single milking. Thus we must use incomplete blocks of size three.

Sixteen farms were selected. At each farm there will be three strainings at one milking, with the milk strained first with one filter, then a second, then a third. Each treatment will be used three times in the design: once as a first filter, once as second, and once as third. The treatments and responses for the experiment are given below (data from Connor 1958, data set `MilkFiltration`):

Treatments and Responses

Filtration time

Farm	First		Second		Third	
1	10	451	7	457	16	343
2	11	260	8	418	13	320
3	12	464	5	317	14	315
4	9	306	6	462	15	291
5	13	381	4	597	6	491
6	14	362	1	325	7	449
7	15	292	2	402	8	576
8	16	431	3	477	5	394
9	7	329	9	261	4	430
10	8	389	10	413	1	272
11	5	368	11	244	2	447
12	6	398	12	517	3	354
13	2	490	16	311	9	278
14	3	467	13	429	10	486
15	4	735	14	642	11	474
16	1	402	15	380	12	589

What type of design is this? Analyze the data and report your findings on the influence of the treatment factors on straining time.

Problem 13.5

The State Board of Education has adopted basic skills tests for high school graduation. One of these is a writing test. The student writing samples are graded by professional graders, and the board is taking some care to be sure that the graders are grading to the same standard. We examine grader differences with the following experiment. There are 25 graders. We select 30 writing samples at random; each writing sample will be graded by five graders. Thus each grader will grade six samples, and each pair of graders will have a test in common (data set `BasicSkills`.

Exam	Grader					Score					Exam	Graders					Scores				
1	1	2	3	4	5	60	59	51	64	53	16	1	9	12	20	23	61	67	69	68	65
2	6	7	8	9	10	64	69	63	63	71	17	2	10	13	16	24	78	75	76	75	72
3	11	12	13	14	15	84	85	86	85	83	18	3	6	14	17	25	67	72	72	75	76
4	16	17	18	19	20	72	76	77	74	77	19	4	7	15	18	21	84	81	76	79	77
5	21	22	23	24	25	65	73	70	71	70	20	5	8	11	19	22	81	84	85	84	81
6	1	6	11	16	21	52	54	62	54	55	21	1	8	15	17	24	70	65	61	66	66
7	2	7	12	17	22	56	51	52	57	51	22	2	9	11	18	25	84	82	86	85	86
8	3	8	13	18	23	55	60	59	60	61	23	3	10	12	19	21	72	85	77	82	79
9	4	9	14	19	24	88	76	77	77	74	24	4	6	13	20	22	85	75	78	82	83
10	5	10	15	20	25	65	68	72	74	77	25	5	7	14	16	23	58	64	58	57	58
11	1	10	14	18	22	79	77	77	77	79	26	1	7	13	19	25	66	71	73	70	70
12	2	6	15	19	23	70	66	63	62	66	27	2	8	14	20	21	73	67	63	70	66
13	3	7	11	20	24	48	49	51	48	50	28	3	9	15	16	22	58	70	69	61	71
14	4	8	12	16	25	75	64	75	68	65	29	4	10	11	17	23	95	84	88	88	87
15	5	9	13	17	21	79	77	81	79	83	30	5	6	12	18	24	47	47	51	49	56

Analyze these data to determine if graders differ, and if so, how. Be sure to

describe the design.

    Thirty consumers are asked to rate the softness of clothes washed by ten different detergents, but each consumer rates only four different detergents. The design and responses are given below:

Rater	Trts	Softness	Rater	Trts	Softness
1	A B C D	37 23 37 41	16	A B C D	52 41 45 48
2	A B E F	35 32 39 37	17	A B E F	46 42 45 42
3	A C G H	39 45 39 41	18	A C G H	44 43 41 36
4	A D I J	44 42 46 44	19	A D I J	32 42 36 29
5	A E G I	44 44 45 50	20	A E G I	43 42 44 44
6	A F H J	55 45 53 49	21	A F H J	46 41 43 45
7	B C F I	47 50 48 52	22	B C F I	43 51 40 42
8	B D G J	37 42 40 37	23	B D G J	38 37 36 34
9	B E H J	32 34 39 29	24	B E H J	40 49 43 44
10	B G H I	36 41 39 43	25	B G H I	23 20 27 29
11	C E I J	45 44 40 36	26	C E I J	46 49 48 43
12	C F G J	42 38 39 39	27	C F G J	48 43 48 41
13	C D E H	47 48 46 47	28	C D E H	35 35 31 26
14	D E F G	43 47 48 41	29	D E F G	45 47 47 42
15	D F H I	39 32 32 31	30	D F H I	43 39 38 39

Analyze these data for treatment effects and report your findings.

When recovering interblock information in a BIBD, we take the weighted average of intra- and interblock estimates

Question 13.1

$$\bar{\zeta} = \lambda\hat{\zeta} + (1 - \lambda)\tilde{\zeta} \ .$$

Suppose that $\sigma^2 = \sigma_\beta^2 = 1$, $g = 8$, $k = 7$, and $b = 8$. Find the mean and standard deviation of $1/\lambda$. Do you feel that λ is well determined?

Chapter 14

Optimal Design

Optimal design is a way of thinking about designing experiments rather than a specific type of design or design setting. In brief, one has a model for the data, an optimality criterion, a set of design constraints, and an algorithm. The model for the data is a statement of the distribution of the data, typically involving parameters such as means, variances, and so on. We have seen many such models and will see more as we go on. The optimality criterion is a mathematical function that defines how good an experiment is. Typical examples relate to how well the model parameters can be estimated, or how well the model can predict future data. The feasible experiment with the best optimality criterion is the optimal design.

The design constraints may take many forms. The most obvious and common constraint is sample size, but other design constraints could take the form of a factor having only discrete levels (versus continuous possible values) or a continuous factor being constrained between minimum and maximum values.

Finally, the algorithm determines the actual optimal design based on the model, criterion, and constraints. In rare cases, we can determine the optimal design analytically, and thus the algorithm is mathematical proof. Usually, we rely on computer programs to search for the optimal design, but even here different programs can select different designs (and might even select a different design if you ran the program again).

With some types of non-linear models (e.g., logistic regression), the situation is more complex, because the optimal design depends on the values of the parameters being estimated. This puts us in the ironic situation of needing to know the values of the parameters in order to design an optimal experiment to estimate the values of those parameters. In practice, it is sufficient to have a range of plausible values for the parameters, but we always need to bear in mind that with non-linear models a design that is quite good for one set of parameters may be quite poor for another set of parameters.

Many standard designs in standard settings are optimal designs. The utility of optimal design arises in non-standard settings. For example,

- What if the block size is not a multiple of the number of treatments?

- What if different blocks have different numbers of units?

- What if the sum of the levels of factors one and two must be less than 7 for safety reasons?

- What if your primary constraint is not the number of runs but the amount of factor one that is available for use?

- What if you cannot afford to run all of the factor/level combinations?

These and many other considerations kick you out of standard settings and into a situation where optimal design is preferred.

Optimal design sounds great, but it cannot be used blindly. The simplest example of this is suppose that you believe that you have a linear (that is, first order) model over one variable. The optimal design will have half of the design points at the lower limit of the design space and half at the upper limit. This is the best that you can do if your first order model is, in fact, correct. But if the true model is actually quadratic (second order), your first order design will not be able to estimate the parameters of that quadratic model, and you will not even be able to detect that you need something more complicated than the first order model.

> Incorrect assumptions combined with optimal design can lead one far astray.

14.1   Notation and Preliminaries

Optimal design criteria can often be expressed in non-mathematical terms, but it is generally more compact to express them in mathematical terms. Thus parts of this chapter will be more mathematical than most of this text in that we will be using matrix algebra.

Let's begin by thinking about the simple one-way model with g treatments. The are many ways to parameterize that model. We often express that model as

$$y_{ij} = \mu_i + \epsilon_{ij}$$

together with the statement that the ϵ_{ij} are independent normal random variables with mean 0 and variance σ^2. An alternative way to write this is as

$$y_{ij} = w_{1ij}\mu_1 + w_{2ij}\mu_2 + \cdots + w_{gij}\mu_g + \epsilon_{ij}$$

where w_{kij} is 1 when $k = i$ and is 0 when $k \neq i$. The w_{kij} terms allow us to pick out which mean to use in the model, as only one of the group means will have a coefficient of 1 for any data value.

We have N data points. Stack all of the data into one vector $\boldsymbol{y}$ (of length N), and stack all of the ϵ_{ij} values (in the same order) into a vector $\boldsymbol{\epsilon}$. Now make an $N \times g$ matrix X, where each row of X contains w_{1ij} through w_{gij}

Rearrange into matrix form

for the ij pair corresponding to the data in that row. Finally, stack μ_1 through μ_g into a vector $\boldsymbol{\beta}$. Then in matrix notation we have

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

If $k = 3$; the treatment sample sizes were 2, 3, and 5; and the units were in treatment order; then X could be written

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

with

$$\boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

We have mentioned many times that the parameterization in fixed effects is not unique, and many choices are available. Consider the parameterization where $\mu_i = \mu + \alpha_i$ with $\alpha_1 = 0$. For this parameterization we have

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

Suppose instead we want the parameterization where $\sum_i \alpha_i = 0$; this is equivalent to $\alpha_g = -\alpha_1 - \alpha_2 - \cdots - \alpha_{g-1}$. That is, adding the last treatment effect is the same as subtracting all the other treatment effects. Then we

would have

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}$$

We also considered polynomial models in cases where we have a continuous predictor. Suppose that level i of the factor has quantitative level z_i. Then we can fit models of the form

$$y_{ij} = \eta_0 + \eta_1 z_{ij} + \eta_2 z_{ij}^2 + \cdots + \eta_{g-1} z_{ij}^{g-1} + \epsilon_{ij}.$$

For this kind of parameterization, we have

$$X = \begin{bmatrix} 1 & z_1 & z_1^2 \\ 1 & z_1 & z_1^2 \\ 1 & z_2 & z_2^2 \\ 1 & z_2 & z_2^2 \\ 1 & z_2 & z_2^2 \\ 1 & z_3 & z_3^2 \\ 1 & z_3 & z_3^2 \\ 1 & z_3 & z_3^2 \\ 1 & z_3 & z_3^2 \\ 1 & z_3 & z_3^2 \end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \eta_0 \\ \eta_1 \\ \eta_2 \end{bmatrix}$$

The point of all this is that regardless of the parameterization that we choose, we can set up the standard linear model in terms of $\boldsymbol{y}$, X, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$.

The parameterization, X, and $\boldsymbol{\beta}$ determine each other.

(Bayesian models will be more complex.)

One other important aspect of the model is $V_{\boldsymbol{\epsilon}}$, the matrix of variances and covariances for the elements of $\boldsymbol{\epsilon}$. Usually, we assume that the individual ϵ_{ij} have constant variance σ^2 and are independent; independence means that the different ϵ_{ij}s have zero covariance. With these standard assumptions, $V_{\boldsymbol{\epsilon}}$ is an $N \times N$ matrix with σ^2 on the diagonal and 0 on the off-diagonal. Alternatively,

$$V_{\boldsymbol{\epsilon}} = \sigma^2 I_N$$

where I_N is the $N \times N$ identity matrix.

Need variance structure of random errors

In linear mixed effects, there are additional random terms beyond $\boldsymbol{\epsilon}$. We usually write these models as

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

Need variance structure of mixed terms

In these models, $\boldsymbol{\gamma}$ is a vector of normally distributed random elements each with mean zero. The matrix Z has N rows, and the ith row of Z determines how the random effects contribute to the ith element of $\boldsymbol{y}$. In many situations, Z is a matrix of ones and zeros, meaning that for each element of $\boldsymbol{y}$, some elements of $\boldsymbol{\gamma}$ are added in and some are not.

We denote the matrix of variances and covariances of $\boldsymbol{\gamma}$ by $V_{\boldsymbol{\gamma}}$. The variances are usually unknown, although we usually know that certain elements have the same variance. For example, elements corresponding to the effects of the levels of a random factor B would all have the same variance σ_β^2, and the elements corresponding to the effects of the levels of a random interaction BC would all have the same variance $\sigma_{\beta\gamma}^2$, even though we do not know the values of σ_β^2 and $\sigma_{\beta\gamma}^2$. In the models we have seen so far we usually assume that the covariances between the elements of $\boldsymbol{\gamma}$ are zero, but that is not a requirement.

The matrix of variances and covariances for $\boldsymbol{y}$ is $V_{\boldsymbol{y}}$; it is determined by $V_{\boldsymbol{\epsilon}}$ and, if present, Z and $V_{\boldsymbol{\gamma}}$:

$$V_{\boldsymbol{y}} = V_{\boldsymbol{\epsilon}}$$

or

$$V_{\boldsymbol{y}} = Z V_{\boldsymbol{\gamma}} Z^\mathsf{T} + V_{\boldsymbol{\epsilon}}$$

For linear models, the (generalized) least squares estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = (X^\mathsf{T} V_{\boldsymbol{y}}^{-1} X)^{-1} X^\mathsf{T} V_{\boldsymbol{y}}^{-1} \boldsymbol{y}$$

When there are no random effects, this reduces to

$$\widehat{\boldsymbol{\beta}} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} \boldsymbol{y}$$

The variance/covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$V_{\widehat{\boldsymbol{\beta}}} = (X^\mathsf{T} V_{\boldsymbol{y}}^{-1} X)^{-1}$$

Variance of estimated parameters

or just

$$V_{\widehat{\boldsymbol{\beta}}} = \sigma^2 (X^{\mathsf{T}} X)^{-1}$$

when there are no random effects and the observations are independent.

Note that we we usually must use an estimate of $V_{\boldsymbol{y}}$ when we have random effects. For example, with REML we first estimate the variances to get an estimate of $V_{\boldsymbol{\gamma}}$, and then we plug that estimated $V_{\boldsymbol{\gamma}}$ into our formulae for $\widehat{\boldsymbol{\beta}}$ and $V_{\widehat{\boldsymbol{\beta}}}$. One result of treating the estimated $V_{\boldsymbol{\gamma}}$ as if it were the truth is that we wind up with a $V_{\widehat{\boldsymbol{\beta}}}$ that is a little bit too small.

To make a prediction at some point, we first determine $\mathbf{x}_0^{\mathsf{T}}$, which would be that point's row in the X matrix if we had observed data there. The prediction at $\mathbf{x}_0$ is then

$$\widehat{y}_0 = \mathbf{x}_0^{\mathsf{T}} \widehat{\boldsymbol{\beta}}$$

with variance

$$V_{\widehat{y}_0} = \mathbf{x}_0^{\mathsf{T}} V_{\widehat{\boldsymbol{\beta}}} \mathbf{x}_0$$

Variance of prediction

In the situation of independent data, that variance reduces to

$$V_{\widehat{y}_0} = \sigma^2 \mathbf{x}_0^{\mathsf{T}} (X^{\mathsf{T}} X)^{-1} \mathbf{x}_0$$

Clearly, the $(X^{\mathsf{T}} X)^{-1}$ matrix is key in determining how well an experiment can estimate or predict.

14.2 Optimality Criteria

Many optimality criteria are abbreviated with a letter, giving us D-optimal, A-optimal, G-optimal, and so forth. There is a veritable alphabet soup of optimality criteria, but they generally break down into those directed toward estimation of parameters and those directed toward prediction.

Optimal designs for different models can be dramatically different. This makes sense as, for example, it takes three points to fit a quadratic but only two to fit a line. On the other hand, optimal designs for different criteria are often rather similar. Thus we need to take a great deal of care when specifying the model, but specification of the optimality criterion is somewhat less critical.

Correct model is crucial

14.2.1 Estimation-based criteria

The variance of $\widehat{\boldsymbol{\beta}}$ is $V_{\widehat{\boldsymbol{\beta}}}$. Estimation-based optimality criteria try to measure the size of $V_{\widehat{\boldsymbol{\beta}}}$, with "smaller" $V_{\widehat{\boldsymbol{\beta}}}$ being better. Different criteria derive from different ways of measuring how small a matrix is. If only a subset of the parameters are of interest, the optimality criteria can be applied to the subset of $V_{\widehat{\boldsymbol{\beta}}}$ corresponding to the parameters of interest.

A-optimality The A is short for average. A-optimal designs are those that minimize the average (or sum) of the estimation variances for all parameters. This is the average (or sum) of the diagonal elements of $V_{\widehat{\boldsymbol{\beta}}}$.

The sum of the diagonal elements of a matrix is called the *trace* of a matrix, so the criterion to minimize is

$$A(V_{\widehat{\boldsymbol{\beta}}}) = \mathsf{trace}(V_{\widehat{\boldsymbol{\beta}}})$$

C-optimality Suppose that there is a certain linear combination of parameters that is of particular interest, and we wish to minimize the variance of the estimate of that linear combination. If we write the linear combination as $\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}$, then the criterion to minimize is

$$C(V_{\widehat{\boldsymbol{\beta}}}) = \boldsymbol{c}^{\mathsf{T}} V_{\widehat{\boldsymbol{\beta}}} \boldsymbol{c}$$

E-optimality More generally than C-optimality, we want to ensure that the variance of the *worst case* linear combination estimate is as small as possible. Of course, if we double the size of the coefficients the variance quadruples, so we need to fix the size of the coefficients at a certain amount; here we require that $\boldsymbol{c}^{\mathsf{T}}\boldsymbol{c} = 1$. Thus the criterion to minimize is

$$E(V_{\widehat{\boldsymbol{\beta}}}) = \max_{\boldsymbol{c}^{\mathsf{T}}\boldsymbol{c}=1} \boldsymbol{c}^{\mathsf{T}} V_{\widehat{\boldsymbol{\beta}}} \boldsymbol{c}$$

The E-criterion is equivalent to the maximum eigenvalue of the matrix $V_{\widehat{\boldsymbol{\beta}}}$, thus motivating the E in the name.

D-optimality A confidence interval for a single parameter has a length that increases with the variance of the estimate. A confidence region for a vector of parameters usually takes the form of an ellipsoidal region, and the volume of the confidence region increases with the *determinant* of the variance matrix. D-optimal designs minimize the volume of the confidence region for the parameters by minimizing the determinant of the variance matrix. Thus the criterion to minimize is

$$D(V_{\widehat{\boldsymbol{\beta}}}) = \left| V_{\widehat{\boldsymbol{\beta}}} \right| = \mathsf{det}(V_{\widehat{\boldsymbol{\beta}}})$$

14.2.2   Prediction-based criteria

The prediction at $\mathbf{x}_0$ is

$$\widehat{y}_0 = \mathbf{x}_0^{\mathsf{T}}\widehat{\boldsymbol{\beta}}$$

with variance

$$V_{\widehat{y}_0} = \mathbf{x}_0^{\mathsf{T}} V_{\widehat{\boldsymbol{\beta}}} \mathbf{x}_0$$

The optimality criteria differ by considering which values of $\mathbf{x}_0$ to consider and how to combine the different prediction variances.

G-optimality G-optimality chooses the design that minimizes the largest prediction variance when considering all of the points in the design itself (i.e., corresponding to rows of the X matrix). The criterion to minimize is thus

$$\max(\mathsf{diag}(XV_{\widehat{\boldsymbol{\beta}}}X^{\mathsf{T}}))$$

The G is said to stand for global, although "all the points in the design" is a narrow interpretation of global. Speaking generally, you can improve the G criterion by moving design points from regions of low prediction variance to regions of high prediction variance.

When $V_{\boldsymbol{y}}$ is a multiple of the identity matrix (independent, constant variance errors), the smallest that the G criterion can be is $\sigma^2 p/n$ where p is the number of parameters in the model for the mean. Thus any design that achieves that bound must be G-optimal and you can determine how far any design is from G-optimality.

I-optimality I-optimality works with the integrated (or averaged) prediction variance across a defined design space. The criterion to minimize is thus

$$\mathsf{average}(\mathbf{x}^{\mathsf{T}}V_{\widehat{\boldsymbol{\beta}}}\mathbf{x})$$

where the average is taken across the potential $\mathbf{x}$ values in the design space. This is equivalent to

$$\mathsf{trace}(V_{\widehat{\boldsymbol{\beta}}}M)$$

where

$$M = \mathsf{average}(\mathbf{x}\mathbf{x}^{\mathsf{T}})$$

with the average again taken over the defined design space.

14.2.3 Relationships

There are many relationships that can be proven about these criteria and designs, but here are three results.

Equivalence A design that is D-optimal is also G-optimal (and vice versa).

Invariance A design that is D-optimal for one parameterization is also D-optimal for any other parameterization (assuming all parameterizations are full rank).

Dominance The criteria in the mixed effects case (that is, with $V_{\boldsymbol{\gamma}}$ non-zero) will always be larger than the corresponding criterion with $V_{\boldsymbol{\gamma}} = 0$.

Equivalence is completely non-obvious, but it does give us some linkage between the estimation-type and prediction-type criteria. It is also useful in proving optimality. Invariance means that we do not need to obsess over which parameterization we use when choosing a D-optimal design. Dominance can help establish optimality. For example, if a design is optimal when

$V_\gamma = 0$ and its V_β does not depend on V_γ, then it will be optimal for any value of V_γ (this means, for example, that orthogonal blocking is a good idea).

14.3 Algorithms

Most optimal design algorithms fall into one of two classes: deterministic greedy algorithms or stochastic search. Both classes of algorithms start with a base design and then modify the design repeatedly in an attempt to improve the criterion. Greedy algorithms deterministically and repeatedly consider a set of modifications to the design, and accept modifications that improve the criterion. Stochastic search algorithms choose the modifications randomly and generally have the possibility of making a move that actually makes the criterion worse. Commercial design packages tend to use greedy algorithms.

Greedy and stochastic algorithms

If there are multiple local optima, greedy algorithms can get stuck in a local optimum and miss the global optimum. They are usually run several times with different starting designs in hopes of finding the global optimum. On the positive side, greedy algorithms tend to be fast. Stochastic search algorithms often come with a theoretical guarantee that they will (eventually) find the global optimum. However, they tend to be slow, and if you aren't clever in how you set them up they can be even slower. The theoretical guarantees aren't worth much if it takes a billion years of computer time for them to work.

Local optima

The original design algorithm was called point exchange. You set up a list of potential design points and an initial design. Then you repeatedly consider swapping each point in the design for a point in the list not in the design. If the criterion improves, make the swap. Keep doing this until no swap improves the criterion. Point exchange is a greedy algorithm.

Point exchange

A second greedy algorithm is coordinate exchange. In this algorithm, instead of swapping out entire design points you change one factor level of a design point, exchanging the current level for a potential level. The potential levels could be a discrete list, or they could be a continuous range of feasible values. Coordinate exchange is a bit more complex than point exchange, but it tends to give better designs.

Coordinate exchange

Two stochastic search methods are genetic search and simulated annealing. Both methods emulate physical/biological processes that tend to find optimum states. Simulated annealing connects the criterion to the energy in a physical system; annealing slowly cools the system, and as it cools the system tends to the low energy state. Genetic search connects the criterion to fitness, and the process emulates survival of the fittest: as the organism (that is, the design) evolves, it tends toward better fitness (better criteria).

Simulated annealing and genetic search

14.4   Examples

We end this chapter with a handful of examples. These examples indicate the optimality of some of the recommendations we have seen in earlier chapters, and they illustrate some of the power of algorithmically chosen designs.

Example 14.1 Balanced designs

We have seen that balanced designs are often less susceptible to violations of assumptions, but they are also optimal in certain situations.

Consider a completely randomized design with sample sizes n_i, $N = rg$, and the model that uses treatment means. Then $V_{\widehat{\beta}}$ is diagonal with σ^2/n_i on the diagonal. The A and D criteria are

$$A = \sum_{i=1}^{g} \frac{\sigma^2}{n_i} \qquad D = \prod_{i=1}^{g} \frac{\sigma^2}{n_i}$$

It is easy to see that if $n_i > n_j + 1$, then you can decrease (improve) the criterion by moving observations from group i to group j. Thus a balanced design will be optimal.

Thinking about the G criterion, the sample size in a treatment for a balanced design is N/g, so the prediction variance will be the variance of the treatment mean: $\sigma^2 g/N$. Because there are g parameters in the model for the mean, this design achieves the lower bound and must be G-optimal.

Example 14.2 Compare with control

Group 1 is a control group and we are only interested in the parameters $\alpha_1 - \alpha_i$ for $i = 2, ..., g$. Assume that the non-control sample sizes are equal (call them n_2), then $N = n_1 + (g-1)n_2$, or, equivalently $n_2 = (N - n_1)/(g-1)$. The variance of each contrast is

$$V_{\alpha_1 - \alpha_i} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Minimizing the A criterion means minimizing $1/n_1 + 1/n_2$ or, equivalently, $1/n_1 + (g-1)/(N - n_1)$.

Setting the derivative with respect to n_1 to zero and solving leads to

$$n_1 = N \left(\frac{\sqrt{g-1} - 1}{g - 2} \right)$$

which means that

$$\frac{n_1}{n_2} = \sqrt{g-1}$$

This is the sample size ratio recommendation given in Section 5.5.1. Of course, sample sizes must be integers, so this ratio will rarely be achieved exactly.

Example 14.3 Orthogonal blocking

We have g treatments and $N = gr$ units is r blocks of size g. We wish to allocate the treatments to blocks so as to optimally estimate the treatment effects. If we create a randomized complete block design with each treatment used once in each block, then contrasts between treatment means have the same variance regardless of the size of the random block effects. We know that the balanced design is optimal in the no-block-effects situation (see balanced designs above), and the RCB achieves the same criterion value even with non-zero block variances. Therefore, by dominance it is optimal for the blocked situation.

The same argument applies to Latin Squares and other orthogonal blocking designs.

Example 14.4 Fitting a Second Order Model

Suppose that we have two continuous factors. We wish to fit a model that includes linear, quadratic, and cross product terms from both factors. Our available set of design points is a 5 by 5 grid, with each factor having potential levels -2, -1, 0, 1, or 2. We want to choose a D-optimal design with 14 design points, but we allow multiple trials at the same design factors.

This sort of optimality cannot usually be done by hand, so we use software. In **R**, there is a package `AlgDesign` that does just what we need.

```
1 > library(AlgDesign)
2 > pointgrid <- data.frame(A=rep(-2:2,each=15),
        B=rep(-2:2,each=3,length=75))
3 > desD <- optFederov(~quad(A,B),pointgrid,14,crit="D")
4 > desA <- optFederov(~quad(A,B),pointgrid,14,crit="A")
```

Line 2 establishes the set of allowable design points. We wish to allow replication, so we include three copies of every point on the grid. The function `optFederov` uses Federov's point exchange algorithm to find the design. We must give it the model we want to fit, a table of potential design points (including repeats if we want to allow repeats), the number of points needed, and the design criterion. Lines 3 and 4 compute the D and A optimal designs.

```
5 > desD[1:2]
  $D
  [1] 2.996672

  $A
  [1] 1.054768

6 > desA[1:2]
  $D
  [1] 2.368295

  $A
  [1] 0.7189806
```

`AlgDesign` defines the D criterion as the reciprocal of how we have defined it, thus large values of D are good. Lines 5 and 6 show the D and A criteria for both designs. As would be expected, the D-optimal design is better on the D criterion than is the A-optimal design, and vice versa.

```
7 > desD$design
       A  B
  1   -2 -2
  2   -2 -2
  7   -2  0
  13 -2   2
  14 -2   2
  31  0  -2
  37  0   0
  38  0   0
  43  0   2
  61  2  -2
  62  2  -2
  67  2   0
  73  2   2
  74  2   2
```

```
8 > desA$design
       A  B
  1   -2 -2
  7   -2  0
  13 -2   2
  22 -1   0
  31  0  -2
  34  0  -1
  37  0   0
  38  0   0
  39  0   0
  43  0   2
  53  1   0
  61  2  -2
  67  2   0
  73  2   2
```

Finally, the designs themselves are rather different (lines 7 and 8). The D-optimal design has replication at the center and the corners, with a single observation at the midpoint of each edge. In contrast, the A-optimal design

only has replication at the center, with single observations at the corners and along the midlines of each factor.

What if we do not have the full space? For example, what if we can only use design points where $A + B \leq 2$? Simple! Just remove those points from the candidate list and refit.

```
9 > use <- rep(TRUE,75)
10 > use[pointgrid[,1]+pointgrid[,2]>2] <- FALSE
11 > smallgrid <- pointgrid[use,]
12 > desD <- optFederov(~quad(A,B),smallgrid,14,crit="D")
13 > desD$D
 [1] 2.538178
14 > desD$design
      A  B
 1  -2 -2
 2  -2 -2
 7  -2  0
13  -2  2
14  -2  2
33   0 -2
37   0  0
38   0  0
43   0  2
44   0  2
58   2 -2
60   2 -2
64   2  0
65   2  0
```

Lines 9–12 create the reduced candidate set and create the optimal design. Line 13 shows that the D criterion is worse than before; this is expected, because we have fewer points to choose from. Line 14 shows the design. The change is that the points two points at (2,2) (now disallowed) have become points at (0,2) and (2,0).

Changing the model can make a big difference in the optimal design.

```
15 > desD <- optFederov(~quad(A)+B,pointgrid,14,crit="D")
16 > desD
      A  B
 1  -2 -2
 2  -2 -2
 3  -2 -2
13  -2  2
14  -2  2
31   0 -2
32   0 -2
33   0 -2
43   0  2
44   0  2
61   2 -2
63   2 -2
73   2  2
75   2  2
```

In line 15, we ask for an optimal design for a model that is quadratic in A and linear in B with no cross-product terms. Line 16 shows the design: B is

only at –2 and 2, and A is roughly balanced across –2, 0, and 2. This design can't even estimate the full quadratic model.

Example 14.5 Balanced Incomplete Blocks

The BIBD is optimal when it exists for the given combination of g, r, b, and k (Kiefer 1958). The proof of this is well beyond the level of this text, but we can illustrate this fact algorithmically. Suppose that we want an incomplete block design for a factor with five levels, and we have available ten blocks of size two. A BIBD is available using all ten pairs of treatments.

```
1 > dspace <- data.frame(trt=factor(1:5))
2 > desD <- optBlock(~trt,dspace,rep(2,10))
3 > desD$rows
 [1] 2 5 3 4 1 3 1 5 4 5 1 2 2 3 1 4 2 4 3 5
4 > table(desD$rows,rep(1:10,each=2))

     1 2 3 4 5 6 7 8 9 10
  1  0 0 1 1 0 1 0 1 0  0
  2  1 0 0 0 0 1 1 0 1  0
  3  0 1 1 0 0 0 1 0 0  1
  4  0 1 0 0 1 0 0 1 1  0
  5  1 0 0 1 1 0 0 0 0  1
> tcrossprod(table(desD$rows,rep(1:10,each=2)))
     1 2 3 4 5
  1  4 1 1 1 1
  2  1 4 1 1 1
  3  1 1 4 1 1
  4  1 1 1 4 1
  5  1 1 1 1 4
```

Line 1 sets up the design space as a single factor, and line 2 asks for ten blocks of size two. Line 3 shows the rows in the design, with each pair of rows indicating a block. Line 4 shows a matrix showing the number of times each treatment occurs in each block, and finally line 5 shows the number of times each treatment co-occurs with the other treatments. We have replication of 4 and each pair occurs together once.

Example 14.6 Other Incomplete Blocks

Suppose now that we still have a factor with five levels, but now we have five blocks of size 2 and five blocks of size 3. This is not any of our standard designs, because our standard designs have all had equal block sizes. What should we do?

```
1 > desD <- optBlock(~trt,dspace,rep(c(2,3),each=5))
2 > table(desD$rows,rep(1:10,rep(2:3,each=5)))
     1 2 3 4 5 6 7 8 9 10
   1 0 1 0 0 1 1 0 0 1  1
   2 1 0 0 1 0 1 1 0 1  0
   3 1 0 0 0 1 0 1 1 0  1
   4 0 1 1 0 0 0 1 1 1  0
   5 0 0 1 1 0 1 0 1 0  1
3 > tcrossprod(table(desD$rows,rep(1:10,rep(2:3,each=5))))
     1 2 3 4 5
   1 5 2 2 2 2
   2 2 5 2 2 2
   3 2 2 5 2 2
   4 2 2 2 5 2
   5 2 2 2 2 5
```

Line 1 shows that all we need to do is change the size of the blocks to indicate different block sizes. Line 2 gives the incidence matrix of each treatment in each block, and Line 3 shows the number of times each pair of treatments co-occurs. We see that each pair occurs together twice, as would occur in a BIBD. Here, some pairs only occur together in blocks of size three, while others occur in one block of each size.

14.5   Bayesian Optimal Design

Bayesian analysis is based on a prior distribution for all unknowns. Data are collected, and then the prior distribution is updated (via Bayes Rule) to include the information from the data, creating the posterior distribution for the unknowns. We have a measure of the quality/quantity of the information called the *utility*, and we can compute the utility for both the prior and the posterior. Bayesian optimal design chooses the design in such a way that the expected increase in utility is as large as possible. See Chaloner and Verdinelli (1995) for a survey of Bayesian design results.

<div style="text-align:right">Maximize increase in utility</div>

Different definitions of utility (as with different optimality criteria in the non-Bayesian case) lead to different designs. In the simple (if somewhat unrealistic) case of linear models with normally distributed errors with known variance, several closed forms have been derived. If you use Shannon information as the utility, then the optimal design minimizes the determinant of the posterior variance matrix of the parameters (this posterior variance matrix includes information from the prior as well as the data). If you use mean squared error between the estimate and the true parameter as the utility, then the optimal design minimizes the trace of the posterior variance matrix. Thus there are analogs of D and A-optimality. Note that if the variance of the errors is unknown (as it almost always is), then there are no simple formulae and clear linkages.

<div style="text-align:right">Analogs of D and A optimality</div>

One can create predictive analogs of the G and I-criteria for the Bayesian setting, but these are not known to arise from maximizing utility gain.

We have mentioned that in non-linear models, for example, logistic regression, you need to have at least an approximate idea of the parameter

values in order to find a good design. Bayesian analysis always incorporates any prior information, so Bayesian optimal design adapts to these kinds of situations very naturally via the prior distribution for the parameters.

While many scientists dislike the use of a prior distribution in analysis, experimental design is a setting where you *cannot avoid using prior information*. The experimenter thinks that the treatment effect will be about a certain size when doing sample size analysis. The experimenter thinks that certain factors are likely to influence the result. The experimenter thinks that the maximum response is probably close to this combination of factor levels. The experimenter thinks that certain ranges are feasible for factor levels. The experimenter thinks that the LD50 will be in a certain range.

> You cannot, and should not, avoid prior information when you design.

You can choose not to do Bayesian analysis, but if you want a good design, you use prior information. Everyone is Bayesian in that sense during design.

Chapter 15

Factorials in Incomplete Blocks—Confounding

We may use the complete or incomplete block techniques of the last two chapters when treatments have factorial structure; just consider that there are $g = abc$ treatments and proceed as usual. However, there are some incomplete block techniques that are specialized for factorial treatment structure. We consider these factorial-specific methods in this chapter and the next.

This chapter describes *confounding* as a design technique. A design with confounding is unable to distinguish between some treatment comparisons and other sources of variation. For example, if the experimental drug is only given to patients with advanced symptoms, and the standard therapy is given to other patients, then the treatments are confounded with patient condition. We usually go to great lengths to avoid confounding, so why would we deliberately introduce confounding into an experiment?

Incomplete blocks are less efficient than complete blocks; we always less power or greater estimation variance when we use incomplete blocks instead of complete blocks. Thus the issue with incomplete blocks is not whether we lose information, but how much information we lose, and which particular comparisons lose information. Incomplete block designs like the BIBD and PBIBD spread the inefficiency around every comparison. Confounded factorials allow us to isolate the inefficiency of incomplete blocks in particular contrasts that we specify at design time and retain full efficiency for all other contrasts.

Let's restate that. With factorial treatment structure we are usually more interested in main effects and low-order interactions than we are in multi-factor interactions. Confounding designs will allow us to isolate the inefficiency of incomplete blocks in the multi-factor interactions and have full efficiency for main effects and low-order interactions. We can have complete loss of information on the confounded effects (the price of incomplete blocks) while retaining full information on the unconfounded effects.

Table 15.1: All contrasts and grand mean for a 2^3 design.

	I	A	B	C	AB	AC	BC	ABC
(1)	+	−	−	−	+	+	+	−
a	+	+	−	−	−	−	+	+
b	+	−	+	−	−	+	−	+
ab	+	+	+	−	+	−	−	−
c	+	−	−	+	+	−	−	+
ac	+	+	−	+	−	+	−	−
bc	+	−	+	+	−	−	+	−
abc	+	+	+	+	+	+	+	+

15.1 Confounding the Two-Series Factorial

Let's begin with a review of some notation and facts from Chapter 9. The 2^k factorial has k factors, each at two levels for a total of $g = 2^k$ treatments. There are two common ways to denote factor-level combinations. First is a lettering method. Let (1) denote all factors at their low level. Otherwise, denote a factor-level combination by including (lower-case) letters for all factors at their high levels. Thus bc denotes factors B and C at their high levels and all other factors are their low levels. Second, there is a numbering method. Each factor-level combination is denoted by a k-tuple, with a 1 for each factor at the high level and a 0 for each factor at the low level. For example, in a 2^3, bc corresponds to 110. To refer to individual factors, let x_A be the level of A, and so on, so that $x_A = 0$, $x_B = 1$, and $x_C = 1$ in 110.

Letter or digit labels for factor-level combinations

Standard order for a two-series design arranges the factor-level combinations in a specific order. Begin with (1). Then proceed through the remainder of the factor-level combinations with factor A varying fastest, then factor B, and so on. In a 2^3, the standard order is (1), a, b, ab, c, ac, bc, abc. Standard order is numerical order when using the binary digit method of indicating factor levels.

Standard order

Each main effect and interaction in a two-series factorial is a single degree of freedom and can be described with a single contrast. It is customary to use contrast coefficients of $+1$ and -1, and the contrast is often represented as a set of plus and minus signs, one for each factor-level combination. The full table of contrasts for a 2^3 is shown in Table 15.1, which also includes a column of all $+$ signs corresponding to the grand mean.

Table of $+$ and $-$

The 2^k factorial can be confounded into two blocks of size 2^{k-1} or four blocks of 2^{k-2}, and so on, to 2^q blocks of size 2^{k-q} in general. Let's begin with just one replication of the experiment confounded in two blocks of size 2^{k-1}; we look at smaller blocks and additional replication later.

2^q blocks of size 2^{k-q}

15.1.1 Two blocks

Confounding a 2^k design into two blocks of size 2^{k-1} is simple; the steps are given in Display 15.1. Every factorial effect corresponds to a contrast with

1. Choose a factorial effect to confound with blocks and get its contrast.

2. Put all factor-level combinations with a plus sign in the contrast in one block and all the factor-level combinations with a minus sign in the other block.

Display 15.1: Steps to confound a 2^k design into two blocks.

2^{k-1} plus signs and 2^{k-1} minus signs. Choose a factorial effect to confound with blocks; this is the *defining contrast*. Put all factor-level combinations with a plus sign on the defining contrast in one block and all the factor-level combinations with a minus sign in the other block. This confounds the block difference with the defining contrast effect, so we have zero information on that effect. However, all factorial effects are orthogonal, so block differences are orthogonal to the unconfounded factorial effects, and we have complete information and full efficiency for all unconfounded factorial effects.

<div style="text-align:right">Confound defining contrast with blocks</div>

It makes sense to choose as defining contrast a multifactor interaction, because multi-factor interactions are generally of less interest, and we will lose all information about whatever contrast is used as defining contrast. For the 2^k factorial in two blocks of size 2^{k-1}, the obvious defining contrast is the k-factor interaction.

<div style="text-align:right">Use k-factor interaction as defining contrast</div>

Example 15.1 2^3 in two blocks of size four

Suppose that we wish to confound a 2^3 into two blocks of size four. We use the ABC interaction as the defining contrast, because it is the highest-order interaction. The pattern of plus and minus signs is the last column of Table 15.1. The four factor-level effects with minus signs are (1), ab, ac, and bc; the four factor-level effects with plus signs are a, b, c, and abc. Thus the two blocks are

(1)		a
ab		b
ac		c
bc		abc

This idea of finding the contrast pattern for a defining contrast to confound into two blocks works for any two-series design, but finding the pattern becomes tedious for large designs. For example, dividing a 2^6 into two blocks of 32 with ABCDEF as defining contrast requires finding the ABCDEF contrast, which is the product of the six main-effects contrasts. Here are two equivalent procedures that you may find easier, though which method you like best is entirely a personal matter.

<div style="text-align:right">Alternative methods for finding blocks</div>

First is the "even/odd" rule. Examine the letter designation for every factor-level combination. Divide the factor-level combinations into two groups

<div style="text-align:right">Even/odd rule and 0/1 rule</div>

depending on whether the letters of a factor-level combination contain an even or odd number of letters from the defining contrast. The second approach is the "0/1" rule. Now we work with the numerical 0/1 designations for the factor-level combinations. What we do is compute for each factor-level combination the sum of the 0/1 level indicators for the factors that appear in the defining contrast, and then reduce this modulo 2. (Reduction modulo 2 subtracts any multiples of 2; 0 stays 0, 1 stays 1, 2 becomes 0, 3 becomes 1, and so on.) For the defining contrast ABC, we compute

$$L = x_A + x_B + x_C \bmod 2 \;\;;$$

those factor-level combinations that yield an L value of 0 go in one block, and those that yield a 1 go in the second block. It is not too hard to see that this 0/1 rule is just the even/odd rule in numerical form.

Example 15.2 2^4 in two blocks of eight

Suppose that we have a 2^4 that we wish to block into two blocks using BCD as the defining contrast. To choose blocks using the even/odd rule, we first find the letters from each factor-level combination that appear in the defining contrast, as shown in Table 15.2. We then count whether there is an even or odd number of these letters and put the factor-level combinations with an even number of letters matching in one block and those with an odd number matching in a second block. For example, the combination ac has one letter in BCD, so ac goes in the odd group; and the combination bc has two letters in BCD, so it goes in the even group. Note that we would not ordinarily use BCD as the defining contrast; we use it here for illustration to show that even and odd is not simply the number of letters in a factor-level combination, but the number in that combination that occur in the defining contrast.

To use the 0/1 rule, we start by computing $x_B + x_C + x_D$. We then reduce the sum modulo 2, and assign the zeroes to one block and the ones to a second block. For 0111 (bcd), this sum is $1 + 1 + 1 = 3$, and $3 \bmod 2 = 1$; for 1110 (abc), the sum is $1 + 1 + 0 = 2$, and $2 \bmod 2 = 0$. Table 15.3 shows the results of the 0/1 rule for our example.

The block containing (1) or 0000 is called the *principal block*. The other block is called the *alternate block*. These blocks have some nice mathematical properties that we will find useful in more complicated confounding situations. Consider the following modified multiplication which we will denote by $\odot$. Let (1) act as an identity—anything multiplied by (1) is just itself. So $a \odot (1) = a$ and $bcd \odot (1) = bcd$. For any other pair of factor-level combinations, multiply as usual but then reduce exponents modulo 2. Thus $a \odot ab = a^2 b = a^0 b = b$, and $a \odot a = a^2 = a^0 = (1)$.

There is an analogous operation we can perform with the 0/1 representation of the factor-level combinations. Think of the zeroes and ones as exponents; for example, 1011 corresponds to $d^1 c^0 b^1 a^1 = abd$. Exponents add when we multiply, so the corresponding operation is to add the zeroes and ones componentwise and then reduce them mod 2. Thus $abd \odot acd =$

Principal block and alternate block

Multiply and reduce exponents mod 2

Table 15.2: Confounding a 2^4 with defining contrast BCD using the even/odd rule.

	Matches	Even/odd	Block 1	Block 2
(1)	none	even	(1)	b
a	none	even	a	ab
b	B	odd	bc	c
ab	B	odd	abc	ac
c	C	odd	bd	d
ac	C	odd	abd	ad
bc	BC	even	cd	bcd
abc	BC	even	acd	$abcd$
d	D	odd		
ad	D	odd		
bd	BD	even		
abd	BD	even		
cd	CD	even		
acd	CD	even		
bcd	BCD	odd		
$abcd$	BCD	odd		

$a^2bcd^2 = bc$ corresponds to $1011 \oplus 1101 = 2112 = 0110$. Personally, I prefer the letters, but some people prefer the numbers.

Here are the useful mathematical properties. If you multiply any two elements of the principal block together reducing exponents modulo two, you get another element of the principal block. If you multiply all elements of the principal block by an element not in the principal block, you get an alternate block. What this means is that you can find alternate blocks easily once you have the principal block. This is no big deal when there are only two blocks, but can be very useful when we have four, eight, or more blocks.

Get alternate blocks from principal block

Example 15.3 2^4 in two blocks of eight, continued

In our 2^4 example with BCD as the defining contrast, ac is not in the principal block. Multiplying every element of the principal block by ac, we get the following

$$
\begin{aligned}
(1) \odot ac &= ac && = ac \\
a \odot ac &= a^2c && = c \\
bc \odot ac &= abc^2 && = ab \\
abc \odot ac &= a^2bc^2 && = b \\
bd \odot ac &= abcd && = abcd \\
abd \odot ac &= a^2bcd && = bcd \\
cd \odot ac &= ac^2d && = ad \\
acd \odot ac &= a^2c^2d && = d
\end{aligned}
$$

This is the alternate block, but in a different order than Table 15.2.

Table 15.3: Confounding a 2^4 with defining contrast BCD using the 0/1 rule.

	$x_B + x_C + x_D$	Reduced mod 2	Block 1	Block 2
0000	0	0	0000	0100
1000	0	0	1000	1100
0100	1	1	0110	0010
1100	1	1	1110	1010
0010	1	1	0101	0001
1010	1	1	1101	1001
0110	2	0	0011	0111
1110	2	0	1011	1111
0001	1	1		
1001	1	1		
0101	2	0		
1101	2	0		
0011	2	0		
1011	2	0		
0111	3	1		
1111	3	1		

15.1.2   Four or more blocks

A single replication of a 2^k design can be confounded into two blocks, four blocks, eight blocks, and so on. The last subsection showed how to confound into two blocks using one defining contrast. We can confound into four blocks using two defining contrasts, and in general we can confound into 2^q blocks using q defining contrasts. Let's begin with four blocks.

Use q defining contrasts for 2^q blocks

Start by choosing two defining contrasts for confounding a 2^4 design into four blocks of size four. It turns out that choosing these defining contrasts is very important, and bad choices lead to poor designs. We will use ABC and BCD as defining contrasts; these are good choices. Later on we will see what can happen with bad choices.

Choose defining contrasts carefully

Each defining contrast divides the factor-level combinations into evens and odds (or ones and zeroes). If we look at those factor-level combinations that are even for BCD, half of them will be even for ABC and the other half will be odd for ABC. Similarly, those combinations that are odd for BCD are evenly split between even and odd for ABC. Our blocks will be formed as those combinations that are even for both ABC and BCD, those that are odd for both ABC and BCD, those that are even for ABC and odd for BCD, and those that are odd for ABC and even for BCD. Table 15.4 shows the results of confounding on ABC and BCD. Alternatively, we compute L_1 and L_2 for the two defining contrasts, and take as blocks those combinations that are zero on both, one on both, zero on the first and one on the second, and zero on the second and one on the first.

Combinations of defining contrasts form blocks

We have confounded into four blocks, so there are 3 degrees of freedom between blocks. We know that the two defining contrasts are confounded

Table 15.4: Confounding the 2^4 into four blocks using ABC and BCD as defining contrasts.

	ABC	BCD
(1)	even	even
a	odd	even
b	odd	odd
ab	even	odd
c	odd	odd
ac	even	odd
bc	even	even
abc	odd	even
d	even	odd
ad	odd	odd
bd	odd	even
abd	even	even
cd	odd	even
acd	even	even
bcd	even	odd
$abcd$	odd	odd

	BCD even	BCD odd
ABC even	(1) bc abd acd	ab ac d bcd
ABC odd	a abc bd cd	b c ad $abcd$

with block differences, but what is the third degree of freedom that is confounded with block differences? The ABC contrast is constant (plus or minus 1) within each block, and the BCD contrast is also constant within each block. Therefore, their product is constant within each block. Recall that each contrast is formed as the product of the corresponding main-effect contrasts, so the product of the ABC and BCD contrasts must be the contrast for $AB^2C^2D = AD$. Squared terms disappear because their elements are all ones. The term AD is called the *generalized interaction* of ABC and BCD. When we confound into four blocks using two defining contrasts, we not only confound the defining contrasts with blocks, we also confound their generalized interaction. If you examine the blocks in Table 15.4, you will see that two of them always have exactly one of a or d (odd), and the other two always have both or neither (even).

Generalized interactions of defining contrasts are confounded

Note that if we had chosen AD and ABC as our defining contrasts, we would get the same four blocks, and the generalized interaction BCD would also be confounded with blocks.

This fact that we also confound the generalized interaction explains why we need to be careful when choosing defining contrasts. It is very tempting to use the intuition that we want to confound interactions with as high an order as possible, so we choose, say, ABCD and BCD as defining contrasts. This intuition leads to disaster, because the generalized interaction of ABCD and BCD is A, and we would thus confound a main effect with blocks.

Check generalized interactions when choosing defining contrasts

When choosing defining contrasts, we need to look at the full set of effects that are confounded with blocks. We want first to find a set such that the lowest-order term confounded with blocks is as high an order as possible. Among all the sets that meet the first criterion, we want sets that have

as few low-order terms as possible. For example, consider the sets (A, BCD, ABCD), (ABC, BCD, AD), and (AB, CD, ABCD). We prefer the second and third sets to the first, because the first confounds a main effect, and the second and third confound two-factor interactions. We prefer the second set to the third, because the second set confounds only one two-factor interaction, while the third set confounds two two-factor interactions.

Section B.5 suggests defining contrasts and their generalized interactions for two-series designs with up to eight factors.

Use three defining contrasts to get eight blocks. These defining contrasts must be independent of each other, in the sense that none of them is the generalized interaction the other two. Thus we cannot use ABC, BCD, and AD as three defining contrasts to get eight blocks, because AD is the generalized interaction of ABC and BCD. Divide the factor-level combinations into eight groups using the even/odd patterns of the three defining contrasts: (even, even, even), (even, even, odd), (even, odd, even), (even, odd, odd), (odd, even, even), (odd, even, odd), (odd, odd, even), and (odd, odd, odd). There are eight blocks, so there must be 7 degrees of freedom between them. The three defining contrasts are confounded with blocks, as are their three two-way generalized interactions and their three-way generalized interaction, for a total of 7 degrees of freedom.

We again note that once you have the principal block, you can find the other blocks by choosing an element not in the principal block and multiplying all the elements of the principal block by the new element and reducing exponents mod 2.

Example 15.4 2^5 in eight blocks of four

Suppose that we wish to block a 2^5 design into eight blocks of four. Section B.5 suggests ABC, BD, and AE for the defining contrasts. The principal block is that block containing (1), or equivalently those factor-level combinations that are even for ABC, BD, and AE. The principal block is (1), bcd, ace, and $abde$. This principal block was found by inspection, meaning working through the factor-level combinations finding those that are even for all three defining contrasts.

The remaining blocks can be found by multiplying the elements of the principal block by a factor-level combination not already accounted for. For example, a is not in the principal block, so we multiply and get a, $abcd$, ce, and bde for a second block. Next, b has not been listed, so we multiply by b and get b, cd, $abce$, and ade for the third block. Table 15.5 gives the remaining blocks.

For 2^q blocks, we use q defining contrasts. These q defining contrasts must be independent; no defining contrast can be a generalized interaction of two or more of the others. Form blocks by grouping the factor-level combinations according to the 2^q different even-odd combinations for the q defining contrasts. There will be 2^{k-q} factor-level combinations in each block. There are 2^q blocks, so there are $2^q - 1$ degrees of freedom confounded with blocks. These are the q defining contrasts, their two-way, three-way, and up to q-way

Table 15.5: 2^5 in eight blocks of four using ABC, BD, and AE as defining contrasts, found by products with principal block.

				Multiply by			
P.B.	a	b	c	d	e	ab	ad
(1)	a	b	c	d	e	ab	ad
bcd	$abcd$	cd	bd	bc	$bcde$	acd	abc
ace	ce	$abce$	ae	$acde$	ac	bce	cde
$abde$	bde	ade	$abcde$	abe	abd	de	be

generalized interactions.

Doing the actual blocking is rather tedious in large designs, so it is helpful to have software that will do confounding. The usual even/odd or 0/1 methods are available if you must do the confounding by hand, but a little thinking first can save a lot of calculation.

Example 15.5 2^7 in 16 blocks of eight

Suppose that we are going to confound a 2^7 design into 16 blocks of size eight using the defining contrasts ABCD, BCE, ACF, and ABG. The effects that are confounded with blocks will be

ABCD	ACEG = (BCE)(ABG)
BCE	BCFG = (ACF)(ABG)
ACF	CDEF = (ABCD)(BCE)(ACF)
ABG	BDEG = (ABCD)(BCE)(ABG)
ADE = (ABCD)(BCE)	ADFG = (ABCD)(ACF)(ABG)
BDF = (ABCD)(ACF)	EFG = (BCE)(ACF)(ABG)
CDG = (ABCD)(ABG)	ABCDEFG = (ABCD)(BCE)(ACF)(ABG)
ABEF = (BCE)(ACF)	

We get exactly the same blocks using BCE, ACF, ABG, and ABCDEFG as defining contrasts. Combinations in the principal block always have an even number of letters from every defining contrast. Because the full seven-way interaction including all the letters is one of the defining contrasts, all elements in the principal block must have an even number of letters. Next, no pair of letters occurs an even number of times in BCE, ACF, and ABG, so no two-letter combinations can be in the principal block. Similarly, no six-letter combinations can be in the principal block. This indicates that the principal block will contain (1) and combinations with four letters.

Start going through groups of four letters. We find $abcd$ is a match right at the start. We next find $abef$. We can either get this with a direct search, or by reasoning that if we have a and b, then we can't have g, so we must have two of c, d, e, and f. The combinations with c or d don't work, but $abef$ does work. Similarly, if we start with bc, then we can't have e, and we must have two of a, d, f, and g. The combinations with a and d don't work, but $bcfg$ does work.

We now have (1), $abcd$, $abef$, and $bcfg$ in the principal block. We know that in the principal group we can multiply any two elements together, reduce the exponent mod 2, and get another element of the block. Thus we find that $abcd \odot abef = cdef$, $abcd \odot bcfg = adfg$, $abef \odot bcfg = aceg$, and $abcd \odot abef \odot bcfg = bdeg$ are also in the principal block.

Now that we have the principal block, we can find alternate blocks by finding a factor-level combination not already accounted for and multiplying the elements of the principal block by this new element. For example, a is not in the principal block, so we can find a second block as $a = (1) \odot a$, $bcd = abcd \odot a$, $bef = abef \odot a$, $abcfg = bcfg \odot a$, $acdef = cdef \odot a$, $dfg = adfg \odot a$, $ceg = aceg \odot a$, and $abdeg = bdeg \odot a$. Next, b is not in these first two blocks, so $b = (1) \odot b$, $acd = abcd \odot b$, $aef = abef \odot b$, $cfg = bcfg \odot b$, $bcdef = cdef \odot b$, $abdfg = adfg \odot b$, $abceg = aceg \odot b$, and $deg = bdeg \odot b$ are the next block.

The approach given above is faster than the brute force approach of finding the even/odd pattern for all 128 factor-level combinations on the four defining contrasts, but it is still tedious. Fortunately, the `conf.design` package can relieve the tedium.

The functions in `conf.design` put the defining contrasts in a matrix, with one row for each defining contrast and one column for each factor. Here, we will need four rows and seven columns. Each contrast is designated by zeroes and ones, with a one meaning that the respective factor appears in the contrast.

```
1 > library(conf.design)
2 > ABCD <- c(A=1,B=1,C=1,D=1,E=0,F=0,G=0)
3 > BCE   <- c(0,1,1,0,1,0,0)
4 > ACF   <- c(1,0,1,0,0,1,0)
5 > ABG   <- c(1,1,0,0,0,0,1)
6 > all.generators <- rbind(ABCD,BCE,ACF,ABG)
7 > all.generators
        A B C D E F G
   ABCD 1 1 1 1 0 0 0
   BCE  0 1 1 0 1 0 0
   ACF  1 0 1 0 0 1 0
   ABG  1 1 0 0 0 0 1
```

Lines 2–5 create the defining contrasts as patterns of zeroes and ones. Line 6 combines the defining contrasts into a matrix, with each contrast as its own row. Line 7 prints the matrix.

```
8 > conf.set(all.generators,2)
        A B C D E F G
 [1,] 1 1 1 1 0 0 0
 [2,] 0 1 1 0 1 0 0
 [3,] 1 0 0 1 1 0 0
 [4,] 1 0 1 0 0 1 0
 [5,] 0 1 0 1 0 1 0
 [6,] 1 1 0 0 1 1 0
 [7,] 0 0 1 1 1 1 0
 [8,] 1 1 0 0 0 0 1
 [9,] 0 0 1 1 0 0 1
[10,] 1 0 1 0 1 0 1
[11,] 0 1 0 1 1 0 1
[12,] 0 1 1 0 0 1 1
[13,] 1 0 0 1 0 1 1
[14,] 0 0 0 0 1 1 1
[15,] 1 1 1 1 1 1 1
```

The `conf.set()` function takes a matrix of defining contrasts (and the number of levels of each factor) and computes all of the generalized interactions. We do this in line 8.

```
1 > conf.design(all.generators,2)
     Blocks A B C D E F G
1      0000 0 0 0 0 0 0 0
2      0000 1 1 1 1 0 0 0
3      0000 1 1 0 0 1 1 0
4      0000 0 0 1 1 1 1 0
5      0000 1 0 1 0 1 0 1
6      0000 0 1 0 1 1 0 1
7      0000 0 1 1 0 0 1 1
8      0000 1 0 0 1 0 1 1
9      0001 1 0 1 0 1 0 0
10     0001 0 1 0 1 1 0 0
 ...
127    1111 1 1 1 0 1 1 1
128    1111 0 0 0 1 1 1 1
```

The `conf.design()` function takes a matrix of defining contrasts and allocates the factor-level combinations to the blocks. We do this in line 9, although we only show a portion of the lengthy output. You can verify that the principal block derived earlier matches the output here.

15.1.3 Analysis of a single-replication confounded two-series

With a single replication of a two-series factorial, you can analyze using one of the specialized techniques for that situation (PSE or Basso-Salmaso), or you can use the generic technique of pooling some high-order interactions into error. You have the same choices with a single replication of a confounded two-series design, but you must take care with the effects confounded with blocks.

Use standard methods with nonblock effects

For example, if you use PSE or Basso-Salmaso, recognize that that apparently significant interactions could actually be block effects. If you analyze a single replication of a 2^4 in two blocks via Basso-Salmaso and the C and

Table 15.6: Fraction of images identified in vision experiment. Data in standard order reading down columns. Data set `ImageID`.

.27	.47	.20	.73	.40	.73	.20	.33
.40	.87	.20	.33	.33	.53	.27	.60
.40	.60	.53	.47	.27	.60	.53	.67
.40	.87	.20	.67	.27	.40	.80	.93
.47	.53	.53	.53	.47	.73	.47	.47
.47	.60	.13	.73	.27	.87	.47	.47
.40	.33	.47	.80	.53	.73	.33	.80
.33	.60	.47	.47	.33	.73	.33	.60
.20	.67	.20	.67	.27	.53	.40	.73
.27	.33	.60	.73	.33	.87	.40	.53
.60	.60	.20	.53	.33	.47	.27	.67
.40	.67	.47	.73	.60	.40	.20	.33
.60	.27	.13	.67	.07	.47	.47	.73
.27	.60	.73	.60	.47	.60	.33	.73
.27	.67	.27	.47	.33	.67	.27	.60
.53	.80	.20	.60	.27	.93	.20	.47

ABCD effects look big, you are likely seeing a main effect and the block effect (ABCD).

On the other hand, if you use the pooling approach, you must ensure that the block degrees of freedom are not pooled into error. In **R**, the best way to do that is to construct a block factor (with 2^q levels) and enter it into the model before any of the treatment factors or interactions. Then you can pool as you normally would.

Example 15.6 Visual perception

We wish to study how image properties affect visual perception. In this experiment we will have a subject look at a white computer screen. At random intervals averaging about 5 seconds, we will put a small image on the screen for a very short time. The subject is supposed to click the mouse button when she sees an image on the screen. The experiment takes place in sixteen ten-minute sessions to prevent tiring; during each session we present 120 images. In fact, these are eight images repeated fifteen times each and presented in random order. We record as the response the fraction of times that the mouse is clicked for a given image type.

We wish to study 128 different images, the factorial combinations of seven factors each at two levels: size of image, shape of image, color of image, orientation of image, duration of image, vertical location of image, and horizontal location of image. Because we anticipate session to session variability, we should design the experiment to account for that. A confounded factorial with sixteen blocks of size eight will work. We use the defining contrasts of Example 15.5, and Table 15.6 gives the responses in standard order.

Half Normal Plot for fracID, alpha=0.05
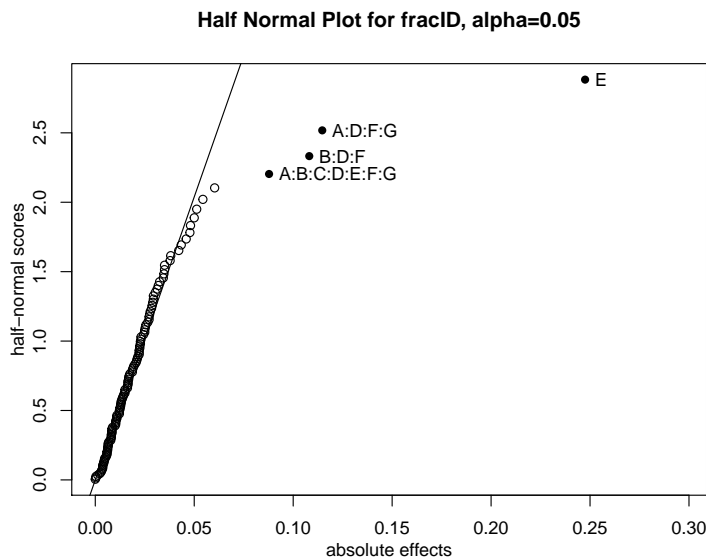
Figure 15.1: Halfnormal plot of factorial effects for transformed vision data, including those confounded with blocks.

There are fifteen factorial effects confounded with blocks, seven three-way interactions, seven four-way interactions, and the seven-way interaction. The remaining $127 - 15 = 112$ are not confounded with blocks. We could pool the five- and six-way interaction degrees of freedom for a 28-degree-of-freedom estimate of error, and then use this surrogate error in testing the lower-order terms that are not confounded with blocks. Alternatively, we could make a half normal plot of the total effects and interpret them with Basso-Salmaso. It would be best to make these plots using only the 112 nonconfounded terms, but it is usually tedious to remove the confounded terms. Outliers in a plot of all terms will need to be interpreted with blocks in mind.

We begin the analysis by noting that the responses are binomial proportions ranging from .07 to .93; for such data we anticipate nonconstant variance, so we transform using arcsine-square roots at the start. Next we make the half-normal plot of effects shown in Figure 15.1. This plot has all 127, including those confounded with blocks, and interprets the significance using Basso-Salmaso. The E main effect is a clear outlier. ADFG, BDF, and ABCDEFG are also identified, but all three are confounded with blocks, so we regard this as block rather than treatment effects.

We conclude that of the treatments we chose, only factor E (duration) has an effect; images that are on the screen longer are easier to see.

15.1.4   Replicating a confounded two-series

We replicate confounded two-series designs for the same reasons that we replicate any design—replication gives us more power, shorter confidence intervals, and better estimates of error. We must choose defining contrasts for the confounding in each replication, and here we have an option. We can confound the same defining contrasts in all replications, or we can confound different contrasts in each replication. Contrasts confounded in all replications are called *completely confounded*, and contrasts confounded in some, but not all, replications are called *partially confounded*.

As before, completely confounded effects cannot be estimated, because they are confounded with block differences in all replications. However, we can get estimates of a partially confounded effect by using data from the replications where the effect is not confounded. This is not all of the replications, so in that sense we have a smaller sample size and thus less information for these partially confounded effects that we do for the non-confounded effects. But we still have some information. Partial confounding generally seems like the better choice, because we will have at least some information on every effect.

Suppose that we have four replications of a 2^3 factorial with two blocks of size four per replication, for a total of eight blocks. One partial confounding scheme would use a different defining contrast in each replication, say ABC in the first replication, AB in the second replication, AC in the third, and BC in the fourth. What can we estimate? First, we can estimate the variation between blocks. There are eight blocks, so there are 7 degrees of freedom between blocks, and the sum of squares for blocks is the sum of squares between the eight groups formed by the blocks. Second, the effects and sums of squares for A, B, and C can be computed in the usual way. This is true for any effect that is never confounded. Next, we can compute the sums of squares and estimated effects for AB, AC, BC, and ABC. Here we must be careful, because all these effects are partially confounded.

Consider first ABC, which is confounded with blocks in the first replication but not in the other replications. The degree of freedom that the ABC effect would estimate in the first replication has already been accounted for as block variation (it is one of the 7 block degrees of freedom), so the first replication tells us nothing about ABC. The ABC effect is not confounded with blocks in replications two through four, so compute the ABC sum of squares and estimated effects from replications two through four. Similarly, we compute the AB effect from replications one, three, and four. In general, estimate an effect and compute its sum of squares from those replication where the effect is not confounded. All that remains after blocks and treatments is error or residual variation. In summary, there are 7 degrees of freedom between blocks, 1 degree of freedom each for A, B, C, AB, AC, BC, and ABC, and $31 - 14 = 17$ degrees of freedom for error.

Let's repeat the pattern one more time. First remove block to block variation. Compute sums of squares and estimated effects for any main effect or interaction by using the standard formulae applied to those replications

Table 15.7: Milk chiller sensory ratings, by blocks; data set
`MilkChiller`.

(1)	86	a	88	(1)	82	b	93
ab	87	b	97	a	74	ab	91
ac	84	c	82	bc	84	c	79
bc	91	abc	85	abc	83	ac	81

in which the main effect or interaction is not confounded. Any effect confounded in every replication cannot be estimated. Error variation is the remainder. This pattern works for complete or partial confounding, and when using statistical software for analysis is most easily expressed as treatments adjusted for blocks.

We can estimate all effects in a partially confounded factorial, but we do not have full information on the partially confounded effects. The effective sample size for any effect is the number of replications in which the effect is not confounded. In the example, the effective sample size is four for A, B, and C, but only three for AB, AC, BC, and ABC. Each of these loses one replication due to confounding. The fraction of information available for an effect is the effective sample size divided by the number of replications. Thus in the example we have full or 100% information for the main effects and 3/4 information for the interactions.

Partial information on partially confounded effects

Example 15.7 Milk chiller

Milk is chilled immediately after Pasteurization, and we need to design a chiller. The goal is to get high flow at low capital and operating costs while still chilling the milk quickly enough to maintain sensory qualities. Basic chiller design is a set of refrigerated plates over which the hot milk is pumped. We are investigating the effect of the spacing between the plates (two levels), the temperature of the plates (two levels), and the flow rate of the milk (two levels) on the perceived quality of the resulting milk. There is a fresh batch of raw milk each day, and we expect batch to batch differences in quality. Because of the time involved in modifying the chiller, we can use at most four factor-level combinations in a day.

This constraint of at most four observations a day suggests a confounded design. We use two replicates, confounding ABC and BC in the two replicates. The processed milk is judged daily by a trained expert who is blinded to the treatments used; the design and results are in Table 15.7. Here is an ANOVA for these data (from Minitab).

```
Source              DF      Seq SS      Adj SS      Seq MS         F       P
block                3      125.19      106.19       41.73      4.07   0.083
space                1       27.56       27.56       27.56      2.69   0.162
temp                 1      189.06      189.06      189.06     18.42   0.008
rate                 1       52.56       52.56       52.56      5.12   0.073
space*temp           1       18.06       18.06       18.06      1.76   0.242
space*rate           1       14.06       14.06       14.06      1.37   0.295
temp*rate            1        0.00        0.00        0.00      0.00   1.000
space*temp*rate      1       10.12       10.12       10.12      0.99   0.366
Error                5       51.31       51.31       10.26
Total               15      487.94


Term               Coef       StDev          T       P
space
1                1.3125      0.8009       1.64   0.162
temp
1               -3.4375      0.8009      -4.29   0.008
rate
1                1.8125      0.8009       2.26   0.073
```

All effects can be estimated because of the partial confounding. There is
evidence for an effect of plate temperature, with lower temperatures giving
better sensory results. There is very slight evidence for a rate effect.

By way of illustration, the sum of squares for the three-factor interaction
in the second replicate is 10.12, what the listing above shows for the three-
factor interaction after adjusting for blocks. The block sum of squares is the
sum of the between replicates, ABC in replicate one, and BC in replicate two
sums of squares (68.06, 2.00, and 55.13 respectively).

15.1.5   Double confounding

Latin Squares, Youden Squares, and related designs allow us to block on
two sources of variation at once; *double confounding* allows us to block on
two sources of variation in a confounding design. Suppose that we have a
2^k treatment structure and that we have two sources of variation on which
to block; there are 2^q levels of blocking on one source and 2^{k-q} levels of
blocking on the other source. Arrange the treatments in a rectangle with 2^q
rows and 2^{k-q} columns. The rows and columns form the blocks for the two
sources of variation.

Double
confounding
blocks on two
sources of
variation

In double confounding, we choose q defining contrasts to generate row
blocking, and $k - q$ defining contrasts to generate column blocking. To pro-
duce the design, we find the principal blocks for rows and columns and put
these in the first row and column of the rectangular arrangement. The remain-
der of the arrangement is filled by taking products and reducing exponents
modulo 2.

Products of
principal blocks

For example, in a 2^4 factorial we could block on two sources of variation
with four levels each. Put the treatments in a four by four arrangement, using
AB and BCD to generate the row blocking, and ABC and CD to generate
the column blocking. The generalized interactions ACD and ABD are also

Confound rows
and columns
separately

confounded. The column principal block is (1), ab, bcd, and acd; the row principal block is (1), abc, cd, and abd; and the full design is

$$
\begin{array}{llll}
(1) & ab & acd & bcd \\
abd & d & bc & ac \\
cd & abcd & a & b \\
abc & c & bd & ad
\end{array}
$$

For example, we take the third row element cd times the fourth column element bcd to get b for the $3, 4$ element of the table. Each row of the treatment arrangement contains a block from the row-defining contrasts, and each column of the arrangement contains a block from the column-defining contrasts.

15.2 Confounding the Three-Series Factorial

Confounding in the three-series factorial is analogous to confounding in the two-series, but threes keep popping up instead of twos. The 2^k is confounded into 2^q blocks each with 2^{k-q} units. The 3^k is confounded into 3^q blocks, each with 3^{k-q} units. When we replicate a three-series design with confounding, we can use complete or partial confounding, just as for the two-series design.

3^q blocks of 3^{k-q} units; partial or complete confounding

The levels of a factor in a three-series design are denoted 0, 1, or 2; for example, the factor-level combinations of a 3^2 design are 00, 10, 20, 01, 11, 21, 02, 12, and 22. The level for factor A is denoted by x_A, just as for the two-series design.

Main effects in a three-series design have 2 degrees of freedom, two-factor interactions have 4 degrees of freedom, and q-factor interactions have 2^q degrees of freedom. We can partition all three-series effects into two-degree-of-freedom bundles. Each main effect contains one of these bundles, each two-factor interaction contains two of these bundles, each three-factor interaction contains four of these bundles, and so on. Each two-degree-of-freedom bundle arises by, in effect, splitting the factor-level combinations into three groups and assessing the variation in the 2 degrees of freedom between these three groups. These two-degree-of-freedom splits provide the basis for confounding the three series, just as one-degree-of-freedom contrasts are the basis for confounding the two series.

Partition three-series effects into two-degree-of-freedom bundles

Each two-degree-of-freedom split has a label, and the labels can be confused with the ordinary interactions, so let's explain them carefully at the beginning. The label for an interaction effect is the letters in the interaction, for example, BCD. The label for a two-degree-of-freedom split is the letters from the factors, each with an exponent of either 0, 1, or 2. By convention, we drop the letters with exponent 0, and by further convention, the first nonzero exponent is always a 1. Thus $A^1 C^2$ and $B^1 C^1 D^2$ are examples of two-degree-of-freedom splits. The two-degree-of-freedom splits that make up an interaction are those splits that have nonzero exponents for the same set of factors as the interaction. Thus the splits in BCD are $B^1 C^1 D^1$, $B^1 C^1 D^2$, $B^1 C^2 D^1$, and $B^1 C^2 D^2$.

Label two-degree-of-freedom splits with exponents

We use these two-degree-of-freedom splits to generate confounding in the three-series in the same way that defining contrasts generate confounding in a two-series, so these splits are often called *defining contrasts*, even though they are not really contrasts (which have just 1 degree of freedom).

15.2.1 Building the design

Each two-degree-of-freedom portion corresponds to a different way to split the factor-level combinations into three groups. For concreteness, consider the $B^1 C^2 D^1$ split in a 3^4 design. Compute for each factor-level combination

$$L = x_B + 2x_C + x_D \bmod 3 \ .$$

Sums of factor levels mod 3 determine splits

The L values will be 0, 1, or 2, and we split the factor-level combinations into three groups according to their values of L. In general, for the split $A^{r_A} B^{r_B} C^{r_C} D^{r_D}$, we compute for each factor-level combination

$$L = r_A x_A + r_B x_B + r_C x_C + r_D x_D \bmod 3 \ .$$

These L values will again be 0, 1, or 2, determining three groups. The block containing the combination with all factors low is the principal block.

Principal block

Example 15.8 A 3^2 with $A^1 B^2$ confounded

Suppose that we want to confound a 3^2 design into three blocks of size three using $A^1 B^2$ as the defining split. We need to compute the defining split L values, and then group the factor-level combinations into blocks, as shown here:

$x_A x_B$	$x_A + 2x_B$	L
00	0	0
10	1	1
20	2	2
01	2	2
11	3	0
21	4	1
02	4	1
12	5	2
22	6	0

$L = 0$	$L = 1$	$L = 2$
00	10	20
11	21	01
22	02	12

This particular arrangement into blocks forms a Latin Square, as can be seen when the block numbers are superimposed on the three by three pattern below:

		x_B		
		0	1	2
	0	0	2	1
x_A	1	1	0	2
	2	2	1	0

If we had used A^1B^1 as the defining split, we would again get a Latin Square arrangement, but that Latin Square would be orthogonal to this one.

To block a three-series into nine blocks, we must use two defining splits P_1 and P_2 with corresponding L values L_1 and L_2. Each L can take the values 0, 1, or 2, so there are nine combinations of L_1 and L_2 values, and these form the nine blocks. To get 27 blocks, we use three defining splits and look at all combinations of 0, 1, or 2 from the L_1, L_2, and L_3 values, and so on for more blocks.

For 3^q blocks, we follow the same pattern but use q defining splits. The only restriction on these splits is that none can be a generalized interaction of any of the others (see the next section). Thus we cannot use A^1C^2, B^1D^1, and $A^1B^1C^2D^1$ as our defining splits. As with two-series confounded designs, we try to find defining splits that confound interactions of as high an order as possible.

Example 15.9 Confounding a 3^3 in nine blocks

Suppose that we wish to confound a 3^3 design into nine blocks using defining splits A^1B^1 and A^1C^2. The L equations are

$$L_1 \quad = \quad x_A + x_B \bmod 3$$

and

$$L_2 \quad = \quad x_A + 2x_C \bmod 3$$

We need to go through all 27 factor-level combinations and compute the L_1 and L_2 values. Once we have the L-values, we can make the split into nine blocks. For example, the 110 treatment has an L_1 value of $1 + 1 = 2$ and an L_2 value of $1 + 2 \times 0 = 1$, so it belongs in the 2/1 block; the 102 treatment has an L_1 value of $1 + 0 = 1$ and an L_2 value of $1 + 2 \times 2 \bmod 3 = 2$, so it belongs in the 1/2 block. The full design follows:

Treatment	L_1	L_2
000	0	0
100	1	1
200	2	2
010	1	0
110	2	1
210	0	2
020	2	0
120	0	1
220	1	2
001	0	2
101	1	0
201	2	1
011	1	2
111	2	0
211	0	1
021	2	2
121	0	0
221	1	1
002	0	1
102	1	2
202	2	0
012	1	1
112	2	2
212	0	0
022	2	1
122	0	2
222	1	0

0/0	0/1	0/2
000	120	210
121	211	001
212	022	122

1/0	1/1	1/2
010	100	220
101	221	011
222	012	102

2/0	2/1	2/2
020	110	200
111	201	021
202	022	112

In the two-series using the 0/1 labels, any two elements of the principal block could be combined using the operation $\oplus$ with the result being an element of the principal block. Furthermore, if you combine the principal block with any element not in the principal block, you get another block. These properties also hold for the three-series design, provided you interpret the operation $\oplus$ as "add the factor levels individually and reduce modulo three."

Combine factor levels mod 3

For example, the principal block in Example 15.9 was 000, 121, and 212. We see that $121 \oplus 121 = 242 = 212$, which is in the principal block. Also, the combination 210 is not in the principal block, so $000 \oplus 210 = 210$, $121 \oplus 210 = 331 = 001$, and $212 \oplus 210 = 422 = 122$ form a block (the one labeled 0/2).

15.2.2 Confounded effects

Confounding a three-series design into three blocks uses one defining split with 2 degrees of freedom. There are 2 degrees of freedom between the three blocks, and these 2 degrees of freedom are exactly those of the defining split.

Confounding a three-series design into nine blocks uses two defining splits, each with 2 degrees of freedom. The 4 degrees of freedom for these

two defining splits are confounded with block differences. There are 8 degrees of freedom between the nine blocks, so 4 more degrees of freedom must be confounded along with the two defining splits. These additional degrees of freedom are from the generalized interactions of the defining splits. If P_1 and P_2 are the defining splits, then the generalized interactions are $P_1 P_2$ and $P_1 P_2^2$.

Recall that we always write these two-degree-of-freedom splits in a three series with exponents of 0, 1, or 2, with the first nonzero exponent always being a 1. Products like $P_1 P_2$ won't always be in that form, so how can we convert? First, reduce exponents modulo three. Second, if the leading nonzero exponent is not a 1, then square the term and reduce exponents modulo three again. The net effect of this second step is to leave zero exponents as zero and swap ones and twos.

Example 15.10 Confounding a 3^3 in nine blocks, continued

The defining splits in Example 15.9 were $A^1 B^1$ and $A^1 C^2$, so the generalized interactions are

$$
\begin{aligned}
P_1 P_2 &= A^1 B^1 \times A^1 C^2 \\
&= A^2 B^1 C^2 \\
&= (A^2 B^1 C^2)^2 \quad \text{leading exponent was 2, so square} \\
&= A^4 B^2 C^4 \\
&= A^1 B^2 C^1 \quad \text{reduce exponents modulo 3}
\end{aligned}
$$

$$
\begin{aligned}
P_1 P_2^2 &= A^1 B^1 (A^1 C^2)^2 \\
&= A^3 B^1 C^4 \\
&= B^1 C^1 \quad \text{reduce exponents modulo 3}
\end{aligned}
$$

Thus the full set of confounded effects is $A^1 B^1$, $A^1 C^2$, $A^1 B^2 C^1$, $B^1 C^1$.

When we confound into 27 blocks using defining splits P_1, P_2, and P_3, there are 26 degrees of freedom between blocks, comprising thirteen two-degree-of-freedom splits. Now it makes sense to give the general rule. Suppose that there are q defining contrasts, P_1, P_2, ... P_q. The confounded degrees of freedom will be $P_1^{v_1} P_2^{v_2} \cdots, P_q^{v_q}$, for all exponent sets that use exponents 0, 1, or 2, and with the leading nonzero exponent being a 1. Applying this to $q = 3$, we get the following confounded terms: P_1, P_2, P_3, $P_1 P_2$, $P_1 P_2^2$, $P_1 P_3$, $P_1 P_3^2$, $P_2 P_3$, $P_1 P_3^2$, $P_1 P_2 P_3$, $P_1 P_2 P_3^2$, $P_1 P_2^2 P_3$, and $P_1 P_2^2 P_3^2$.

Example 15.11 Confounding a 3^5 in 27 blocks

Suppose that we wish to confound a 3^5 into 27 blocks using $A^1 C^1$, $A^1 B^1 D^1$, and $A^1 B^2 E^2$ as defining splits. The complete list of confounded

effects will be

$$
\begin{aligned}
P_1 = A^1 C^1 &= A^1 C^1 \\
P_2 = A^1 B^1 D^1 &= A^1 B^1 D^1 \\
P_3 = A^1 B^2 E^2 &= A^1 B^2 E^2 \\
P_1 P_2 = A^2 B^1 C^1 D^1 &= A^1 B^2 C^2 D^2 \\
P_1 P_2^2 = A^3 B^2 C^1 D^2 = B^2 C^1 D^2 &= B^1 C^2 D^1 \\
P_1 P_3 = A^2 B^2 C^1 E^2 &= A^1 B^1 C^2 E^1 \\
P_1 P_3^2 = A^3 B^4 C^1 E^4 &= B^1 C^1 E^1 \\
P_2 P_3 = A^2 B^3 D^1 E^2 = A^2 D^1 E^2 &= A^1 D^2 E^1 \\
P_2 P_3^2 = A^3 B^5 D^1 E^4 = B^2 D^1 E^1 &= B^1 D^2 E^2 \\
P_1 P_2 P_3 = A^3 B^3 C^1 D^1 E^2 &= C^1 D^1 E^2 \\
P_1 P_2 P_3^2 = A^4 B^5 C^1 D^1 E^4 &= A^1 B^2 C^1 D^1 E^1 \\
P_1 P_2^2 P_3 = A^4 B^4 C^1 D^2 E^2 &= A^1 B^1 C^1 D^2 E^2 \\
P_1 P_2^2 P_3^2 = A^5 B^6 C^1 D^2 E^4 = A^2 C^1 D^2 E^1 &= A^1 C^2 D^1 E^2
\end{aligned}
$$

This design confounds 2 degrees of freedom in the AC interaction, but otherwise confounds three-way interactions and higher.

15.2.3 Analysis of confounded three-series

Analysis of a confounded three-series is analogous to analysis of a confounded two-series. First remove variation between blocks, then remove any treatment variation that can be estimated; any remaining variation is used as error. When there is only one replication, the highest-order interaction is typically used as an estimate of error. With most statistical software, you can get this analysis by requesting an ANOVA with treatment sums of squares adjusted for blocks.

<div style="float:right">Treatments adjusted for blocks</div>

The accounting is a little more complicated in a confounded three-series than it was in the two-series, because confounding is done via two-degree-of-freedom splits, whereas the ANOVA is usually tabulated by interaction terms. For example, consider two replications of a 3^2 with $A^1 B^1$ completely confounded. There are eighteen experimental units, with 17 degrees of freedom between them. There are 5 degrees of freedom between the blocks, 2 degrees of freedom for each main effect, 2 degrees of freedom for the AB interaction, and 6 degrees of freedom for error. The 2 degrees of freedom for AB are the $A^1 B^2$ degrees of freedom, which are not confounded with blocks.

<div style="float:right">Interactions containing completely confounded splits have fewer than nominal degrees of freedom</div>

When we use partial confounding, we can estimate all treatment effects, but we will only have partial information on those effects that are partially confounded. Again consider two replications of a 3^2, but confound $A^1 B^1$ in the first replication and $A^1 B^2$ in the second. We can estimate $A^1 B^1$ in the second replication and $A^1 B^2$ in the first, so we have 4 degrees of freedom for interaction. However, the effective sample size for each of these interaction effects is nine, rather than eighteen.

15.3 Further Reading and Extensions

Two- and three-series are the easiest factorials to confound, but we can use confounding for other factorials too. John (1971) is a good place to get started with these other designs. Kempthorne (1952) also has a good discussion. Derivation and methods for some of these other designs takes some (abstract) algebra. In fact, this algebra is present in the two- and three-series designs; we've just been ignoring it. For example, we have stated that multiplying two elements of the principal block together gives another element in the principal block, and that multiplying the principal block by any element not in the principal block yields an alternate block. These are a consequence of the facts that the factor-level combinations form an (algebraic) group, the principal block is a subgroup, and the alternate blocks are cosets.

Confounding s^k designs when s is prime is the straightforward generalization of the 0/1 and 0/1/2 methods we used for 2^k and 3^k designs. For example, when $s = 5$ and $k = 4$, represent the factor levels by 0, 1, 2, 3, and 4. Block into five blocks of size 125 using the defining split $A^{r_A} B^{r_B} C^{r_C} D^{r_D}$ by computing

$$L = r_A x_A + r_B x_B + r_C x_C + r_D x_D \bmod 5$$

and splitting into groups based on L. If you have two defining splits P_1 and P_2, the confounded effects are P_1, P_2, $P_1 P_2$, $P_1 P_2^2$, $P_1 P_2^3$, and $P_1 P_2^4$. More generally, use powers up to $s - 1$.

To confound s^k designs when s is the mth power of a prime, reexpress the design as a p^{mk} design, where p is the prime factor of s. Now use standard methods for confounding a p^{mk}, but take care that none of the generalized interactions that get confounded are actually main effects. For example, confound a 4^2 design into four blocks of four. A 4^2 design can be reexpressed as a 2^4 design, with the AB combinations indexing the first four-level factor, and the BC combinations indexing the second four-level factor. We could confound ABC and AD (and their generalized interaction BCD). All three of these degrees of freedom are in the 9-degree-of-freedom interaction for the four-series design. We would not want to confound AB, BCD, and ACD, because AB is a degree of freedom in the main effect of the first four-level factor.

Mixed-base factorials are more limited. Suppose we have a $s_1^{k_1} s_2^{k_2}$ factorial, where s_1 and s_2 are different primes. It is straightforward to choose s_1^q blocks of size $s_1^{k_1-q} s_2^{k_2}$ or s_2^q blocks of size $s_1^{k_1} s_2^{k_2-q}$. Just use methods for the factors in play and carry the other factors along. Getting $s_1 s_2$ blocks of size $s_1^{k_1-1} s_2^{k_2-1}$ is considerably more difficult.

15.4 Problems

Confound a 2^5 factorial into four blocks of eight, confounding BCD and **Exercise 15.1**

ACD with blocks. Write out the factor-level combinations that go into each block.

We want to confound a 2^4 factorial into four blocks of size four using ACD and ABD as defining contrasts. Find the factor-level combinations that go into each block.

Exercise 15.2

Suppose that we confound a 2^8 into sixteen blocks of size 16 using ABCF, ABDE, ACDE, and BCDH as defining contrasts. Find the all the confounded effects.

Exercise 15.3

Divide the factor-level combinations in a 3^3 factorial into three groups of nine according to the $A^1B^1C^2$ interaction term.

Exercise 15.4

Suppose that we have a partially confounded 3^3 factorial design run in four replicates, with $A^1B^1C^1$, $A^1B^1C^2$, $A^1B^2C^1$, and $A^1B^2C^2$ confounded in the four replicates. Give a skeletal ANOVA for such an experiment (sources and degrees of freedom only).

Exercise 15.5

A 2^{4-1} fractional factorial is created by the aliasing $I = ABD$, and is then blocked into two blocks of size four using $AC = BCD$. Find the factor-level combinations in the two blocks.

Exercise 15.6

Confound a 2^4 design into eight blocks of two each using the generators AB, BC, and CD. Give the factor/level combinations in each block; what effects are confounded with blocks?

Exercise 15.7

Briefly describe the experimental design you would choose for each of the following situations, and why. Describe treatments, blocks, etc.

Problem 15.1

(a) Nitrification is a bacterial process that plays an important role in the nitrogen cycle of an ecosystem. We wish to assess the variability of nitrification in grasslands at different spatial scales. Specifically, we are interested in the variability in nitrification when measurements are taken within about a meter of each other, and the variability when measurements are taken several meters apart. We have an experimental grassland of one hectare (100 m by 100 m), and we can make 50 measurements of nitrification.

(b) We are trying to understand how procedural changes affect the yield of a bioreactor. The process takes about two hours to run, so we can do four runs per day. There may be day to day variation, we're just not sure. We can set the temperature to high or low, we can set the agitator to fast or slow, and we can either include or omit an additive. We have funds for 16 runs. What kind of design should we use, and why?

(c) Animals are generally used to test medical devices prior to experimenting with the devices in humans. We wish to compare four different designs of arterial stints. (Stints are spring or mesh-like objects that are placed in a narrowed artery and then expand, holding the artery open.) The current experiment concerns the use of stints in the three major coronary arteries.

Each experimental animal is fed a high fat diet that leads to narrowing of the coronary arteries. We can then place stints in the three coronary

arteries. The response that we measure is the increase in diameter of the artery after placing the stint. We have 12 animals, and we expect large animal to animal differences.

(d) Some chemicals may migrate from polystyrene ("styrofoam") cups used to serve hot coffee into the coffee. We will measure the concentration of toluene in hot water solutions after they have sat in polystyrene cups. Three factors are of interest: temperature of water (80 and 90 degrees C), length of time the water sits in the cup before measuring the response (15 minutes and 30 minutes), and pH of the water solution (6 and 6.5). We have resources for 48 measurements. There are no restrictions on the order in which the measurements should be done, but we can only make eight measurements before the equipment needs to be cleaned and recalibrated.

(e) A French press device is one possible choice for brewing coffee. However, one drawback of this device is that the coffee produced can contain some coffee grounds that affect the flavor of the coffee. We would like to run an experiment to investigate the effects of three factors on the flavor of French press-made coffee. Factor A is the amount of grounds put into the press (.1 pound or .12 pound); factor B is the type of roast (French roast of Full City roast); factor C is whether the coffee is stirred once while it is brewing (yes or no).

I am the one who will rate the flavor, and I have time in the morning to make and taste four different brews of coffee. I am not a trained coffee rater, and I think it is likely that my ratings will not be consistent from day to day. I can taste sixteen cups, because I can only do this Monday through Thursday (needing to submit my analysis on Friday).

(f) Untrained consumer judges cannot reliably rate their liking of more than about fifteen to twenty similar foods at one sitting. However, you have been asked to design an experiment to compare the liking of cookies made with 64 recipes, which are the factorial combinations of six recipe factors, each at two levels. The judges are paid, and you are allowed to use up to 50 judges.

(g) Seed germination is sensitive to environmental conditions, so many experiments are performed in laboratory growth chambers that seek to provide a uniform environment. Even so, we know that the environment is not constant: temperatures vary from the front to the back with the front being a bit cooler. We wish to determine if there is any effect on germination due to soil type. We have resources for 64 units (pots with a given soil type). There are eight soil types of interest, and the growth chamber is big enough for 64 pots in an eight by eight arrangement.

(h) Acid rain seems to kill fish in lakes, and we would like to study the mechanism more closely. We would like to know about effects due to the kind of acid (nitric versus sulfuric), amount of acid exposure (as measured by two levels of pH in the water), amount of aluminum present (two levels of aluminum; acids leach aluminum from soils, so it could be the aluminum that is killing the fish instead of the acid), and time of exposure (that is,

a single peak acute exposure versus a chronic exposure over 3 months). We have 32 aquariums to use, and a large supply of homogeneous brook trout.

(i) "Habitat improvement" (HI) is the term used to describe the modification of a segment of a stream to increase the numbers of trout in the stream. HI has been used for decades, but there is little experimental evidence on whether it works. We have eight streams in southeastern Minnesota to work with, and we can make up to eight habitat improvements (that is, modify eight stream segments). Each stream flows through both agricultural and forested landscapes, and for each stream we have identified two segments for potential HI, one in the forested area and one in the agricultural area. We anticipate large differences between streams in trout numbers; there may be differences between forested and agricultural areas. We can count the trout in all sixteen segments.

(j) We wish to study how the fracturability of potato chips is affected by the recipe for the chip. (Fracturability is related to crispness.) We are going to study five factors, each at two levels. Thus there are 32 recipes to consider. We can only bake and measure eight recipes a day, and we expect considerable day to day variation due to environmental conditions (primarily temperature and humidity). We have resources for eight days.

(k) One of the issues in understanding the effects of increasing atmospheric CO_2 is the degree to which trees will increase their uptake of CO_2 as the atmospheric concentration of CO_2 increases. We can manipulate the CO_2 concentration in a forest by using Free-Air CO_2 Enrichment (FACE) rings. Each ring is a collection of sixteen towers (and other equipment) 14 m tall and 30 m in diameter that can be placed around a plot in a forest. A ring can be set to enrich CO_2 inside the ring by 0, 100, or 200 ppm. We have money for six rings and can work at two research stations, one in North Carolina and one in South Carolina. Both research stations have plantations of 10-year-old loblolly pine. The response we measure will be the growth of the trees over 3 years.

(l) We wish to study the effects of soil density, pH, and moisture on snapdragon seed germination, with each factor at two levels. Twenty-four pots are prepared with appropriate combinations of the factors, and then seeds are added to each pot. The 24 pots are put on trays that are scattered around the greenhouse, but only 4 pots fit on a tray.

Briefly describe the experimental design used in each of the following and give a skeleton ANOVA.

Problem 15.2

(a) We wish to study the effects of funding attribution and rigor of reported methods on the perceived authoritativeness of medical research abstracts. There are two fake drugs that we will call A and B. There are two levels of rigor: high and low. There are two levels of funding attribution: NIH and pharmaceutical company. Sixty-four doctors will each be given two fake abstracts to rate for authoritativeness. Each doctor will receive one abstract for drug A and one for drug B. Half of the doctors will receive a

low rigor/pharma abstract and a high rigor/NIH abstract; the other half of the doctors will receive a low rigor/NIH abstract and a high rigor/pharma abstract. Abstracts are arranged so that "drugs" A and B each get an equal number of each kind of abstract.

(b) Our dog's feet get cold and packed with ice when they go for walks in the winter. We can buy little dog booties, but they always come off. What we would like to do is find the brand of booties that stay on the longest. We have purchased booties of six different brands. As our dog only has four feet, we can only test four of them at one time. There are 15 different sets of four brands that can be taken from the six brands. We randomly assign the 15 sets to 15 consecutive days. On each day, we randomly assign the four brands from that day's set to the four feet on the dog. When we go on the walk, we time how long it takes until each bootie comes off, and that is the response.

(c) Raising butterflies at "industrial" scale is done by rearing the caterpillars on an artificial diet rather than a natural diet of leaves. The feed we have available was optimized for a different species of butterfly, and we are unsure whether it has enough protein or choline for our species of butterfly. We want the caterpillars to grow up big and strong, so we are trying to find a combination of the two factors that leads to greater mass of the pupa that the caterpillar forms.

We create a design with a 2x2 factorial treatment structure (no added choline vs added choline; no added protein vs added protein). We have eight petri dishes (one treatment per dish), add five caterpillars per petri dish, and then put the petri dishes in a warmer while the caterpillars grow. Unfortunately, the warmer is not uniformly warm and differs slightly in temperature from back to front. To account for this we create four rows (back to front) of two dishes each. Dishes in row 1 contain treatments (1) and ab; the same is true for row 4. Dishes in row 2 contain treatments a and b; the same is true for row 3.

(d) Psilocybin is the active ingredient in "magic mushrooms" and is reputed to induce religious or spiritual experiences. This experiment tests that claim. Thirty healthy adults were recruited. All subjects participated in two 8-hour sessions two months apart. In each session, the subject was given a strong dose of a drug (either psilocybin or ritalin). A psychologist was present with the subjects during the drug sessions and interviewed the subjects during their experiences (and offered support if needed). The response of interest is the presence and/or intensity of any religious or spiritual experiences during the drug session.

This experiment was randomized, controlled, and double blind with informed consent. The drugs were randomized to the sessions such that each subject received both drugs and each drug was used an equal number of times in the first or second session.

(e) We are studying the Kraft pulping process for making paper. In this experiment, we look at the charge level (705, 853, or 1000), and which additive is used (control, DQ2016 at .1, DQ2016 at .2, AQ at .1, or DTPA at

.2). We can make 10 batches of pulp per day and do the experiment over three days, producing two batches of pulp for each of the 15 combinations of charge level and additive. The fifteen factor/level combinations are assigned so that the treatments with charge at 705 or 853 are in the first day, those where the charge is 705 or 1000 are in the second day, and those where the charge is 853 or 1000 treatments are in the third day.

(f) Two common fish species in cold water streams are slimy sculpin and brown trout. These species tend to inhabit "riffles", which are shallow running stretches of the stream, sort of like miniature rapids. We are interested in whether the presence of the two species together inhibits or enhances total fish growth (combined across species). To study this, we place small cages called enclosures in riffles (each riffle is large enough for multiple enclosures). Into each cage we can place either equal weights x of slimy sculpin and brown trout, or a weight $2x$ of brown trout, or a weight $2x$ of slimy sculpin. After a month, we weigh the fish in each cage to assess total growth.

In our experiment there are five riffles. In each riffle we place three enclosures. The three treatments are randomized to the enclosures subject to the restriction that each treatment occurs once in each riffle.

(g) Neurologists use functional Magnetic Resonance Imaging (fMRI) to determine the amount of the brain that is "activated" (in use) during certain activities. We have twelve right-handed subjects. Each subject will lie in the magnet. On a visual signal, the subject will perform an action (tapping of fingers in a certain order) using either the left or the right hand (depending on the signal). The measured response is the number of "pixels" on the left side of the brain that are activated. We expect substantial subject to subject variation in the response, and there may be a consistent difference between the first trial and the second trial. Six subjects are chosen at random for the left-right order, and the other six get right-left. We obtain responses for each subject under both right- and left-hand tapping.

(h) We wish to study the winter hardiness of four new varieties of rosebushes compared with the standard variety. An experimental unit will consist of a plot of land suitable for 4 bushes, and we have 25 plots available in a five by five arrangement (a total of 100 bushes). The plots are located on the side of a hill, so the rows have different drainage. Furthermore, one side of the garden is sheltered by a clump of trees, so that we expect differences in wind exposure from column to column. The five varieties are randomly arranged subject to the constraint that each variety occurs once in each row and each column. The response of interest is the number of blooms produced after the first winter.

(i) Nisin is a naturally occurring antimicrobial substance, and *Listeria* is a microbe we'd like to control. Consider an experiment where we examine the effects of the two factors "amount of nisin" (factor A, three levels, 0, 100, and 200 IU) and "heat" (factor B, three levels, 0, 5, and 10 second scalds) on the number of live *Listeria* bacteria on poultry skin. We use six

chicken thighs. The skin of each thigh is divided into three sections, and each section receives a different A-B combination. We expect large thigh to thigh variability in bacteria counts. The factor-level combinations used for each skin section follow (using 0,1,2 type notation for the three levels of each factor):

| Section | Thigh | | | | | |
	1	2	3	4	5	6
1	00	10	20	00	10	02
2	11	21	01	21	01	20
3	22	02	12	12	22	11

(j) Semen potency is measured by counting the number of fertilized eggs produced when the semen is used. Consider a study on the influence of four treatments on the potency of thawed boar semen. The factors are cryoprotector used (factor A, two levels) and temperature regime (factor B, two levels). We expect large sow to sow differences in fertility, so we block on sow by using one factor-level combination in each of the two horns (halves) of the uterus. Eight sows were used, with the following treatment assignment.

| Sow | | | | | | | |
1	2	3	4	5	6	7	8
a	ab	(1)	b	b	(1)	(1)	a
b	(1)	ab	a	a	ab	ab	b

Problem 15.3

Individuals perceive odors at different intensities. We have a procedure that allows us to determine the concentration of a solution at which an individual first senses the odor (the threshold concentration). We would like to determine how the threshold concentrations vary over sixteen solutions. However, the threshold-determining procedure is time consuming and any individual judge can only be used to find threshold concentrations for four solutions.

Each solution is a combination of five compounds in various ratios. The sixteen solutions are formed by manipulating four factors, each at two levels. Factor 1 is the ratio of the concentration of compound 1 to the concentration of compound 5. Factors 2 through 4 are similar.

We have eight judges. Two judges are assigned at random to each of the solution sets $[(1), bc, abd, acd]$, $[a, abc, bd, cd]$, $[ab, ac, d, bcd]$, and $[b, c, ad, abcd]$. We then determine the threshold concentration for the solutions for each judge. The threshold concentrations are normalized by dividing by a reference concentration. The ratios are given below (data set `OdorIntensity`):

Judge

	1		2		3		4
(1)	8389	a	4351	ab	6	b	375
bc	816	abc	78	ac	262	c	33551
abd	4	bd	5941	d	1230	ad	246
acd	46	cd	27138	bcd	98	$abcd$	10

	5		6		7		8
(1)	56034	a	2346	ab	67	b	40581
bc	25046	abc	35	ac	3081	c	90293
abd	109	bd	228	d	50991	ad	19103
acd	490	cd	6842	bcd	784	$abcd$	61

Analyze these data to determine how the compounds affect the threshold concentration. Are there any deficiencies in the design?

Problem 15.4

Eurasian water milfoil is a nonnative plant that is taking over many lakes in Minnesota and driving out the native northern milfoil. However, there is a native weevil (an insect) that eats milfoil and may be useful as a control. We wish to investigate how eight treatments affect the damage the weevils do to Eurasian milfoil. The treatments are the combinations of whether a weevil's parents were raised on Eurasian or northern, whether the weevil was hatched on Eurasian or northern, and whether the weevil grew to maturity on Eurasian or northern.

We have eight tanks (big aquariums), each of which is subdivided into four sections. The subdivision is accomplished with a fine mesh that lets water through, but not weevils. The tanks are planted with equal amounts of Eurasian milfoil. We try to maintain uniformity between tanks, but there will be some tank to tank variation due to differences in light and temperature. The tanks are planted in May, then weevils are introduced. In September, milfoil biomass is measured as response and is shown here (data set `Milfoil`):

Tank

	1		2		3		4
(1)	10.4	a	4.8	(1)	16.8	a	12.3
ab	17.5	b	8.9	ab	19.6	b	17.1
ac	22.2	c	6.8	c	16.4	ac	13.3
bc	27.7	abc	17.6	abc	35.6	bc	19.5

	5		6		7		8
(1)	7.7	a	6.3	(1)	14.9	b	7.1
ac	13.3	c	7.3	bc	34.0	c	8.3
b	12.4	ab	11.2	a	16.9	ab	15.3
abc	17.7	bc	25.0	abc	36.8	ac	7.0

Analyze these data to determine how the treatments affect milfoil biomass.

Scientists wish to understand how the amount of sugar (two levels), cul-

Problem 15.5

ture strain (two levels), type of fruit (blueberry or strawberry), and pH (two levels) influence shelf life of refrigerated yogurt. In a preliminary experiment, they produce one batch of each of the sixteen kinds of yogurt. The yogurt is then placed in two coolers, eight batches in each cooler. The response is the number of days till an off odor is detected from the batch (data set `YogurtCooler`).

	Cooler		
1		2	
(1)	34	a	35
ab	34	b	36
ac	32	c	39
ad	34	d	41
bc	34	abc	39
bd	39	abd	44
cd	38	acd	44
$abcd$	37	bcd	42

Analyze these data to determine how the treatments affect time till off odor.

Question 15.1   Consider a defining split in a three-series design, say $A^{r_A} B^{r_B} C^{r_C} D^{r_D}$. Now double the exponents and reduce them modulo 3 to generate a new defining split. Show that the two splits lead to the same three sets of factor-level combinations.

Question 15.2   Show that in a three-series design, any defining split with leading nonzero exponent 2 is equivalent to a a defining split with leading nonzero exponent 1.

Question 15.3   Show that in a three-series design with defining splits P_1 and P_2, the generalized interactions $P_1 P_2^2$ and $P_1^2 P_2$ are equivalent.

Chapter 16

Split-Plot Designs

Split plots are another class of experimental designs for factorial treatment structure. We generally choose a split-plot design when some of the factors are more difficult or expensive to vary than the others, but split plots can arise for other reasons. Split plots can be described in several ways, including incomplete blocks and restrictions on the randomization, but the key features to recognize are that split plots have more than one randomization and more than one idea of experimental unit.

Use split plots when some factors more difficult to vary

16.1   What Is a Split Plot?

The terminology of split plots comes from agricultural experimentation, so let's begin with an agricultural example. Suppose that we wish to determine the effects of four corn varieties and three levels of irrigation on yield. Irrigation is accomplished by using sprinklers, and these sprinklers irrigate a large area. Thus it is logistically difficult to use a design with smallish experimental units, with adjacent units having different levels of irrigation. At the same time, we might want to have small units, because there may be a limit on the total amount of land available for the experiment, or there may be variation in the soils leading us to desire small units grouped in blocks. Split plots give us something of a compromise.

Divide the land into six *whole plots*. These whole plots should be sized so that we can set the irrigation on one whole plot without affecting its neighbors. Randomly assign each irrigation level to two of the whole plots. Irrigation is the *whole-plot factor*, sometimes called the *whole-plot treatment*. Divide each whole plot into four *split plots*. Randomly assign the four corn varieties to the four split plots, with a separate, independent randomization in each whole plot. Variety is the *split-plot factor*. One possible arrangement is as follows, with the six columns representing whole plots with four split plots within each:

Whole plots and whole-plot factor

Split plots and split-plot factor

I2 V1	I3 V4	I3 V1	I1 V3	I2 V3	I1 V2
I2 V3	I3 V3	I3 V3	I1 V2	I2 V1	I1 V1
I2 V2	I3 V1	I3 V4	I1 V1	I2 V2	I1 V4
I2 V4	I3 V2	I3 V2	I1 V4	I2 V4	I1 V3

What makes a split-plot design different from other designs with factorial treatment structure? Here are three ways to think about what makes the split plot different. First, the split plot has two sizes of units and two separate randomizations. Whole plots act as experimental units for one randomization, which assigns levels of the whole-plot factor irrigation to the whole plots. The other randomization assigns levels of the split-plot factor variety to split plots. In this randomization, split plots act as experimental units, and whole plots act as blocks for the split plots. There are two separate randomizations, with two different kinds of units that can be identified before randomization starts. This is the way I usually think about split plots.

Split plots have two sizes of units and two randomizations

Second, a split-plot randomization can be done in one stage, assigning factor-level combinations to split plots, provided that we restrict the randomization so that all split plots in any whole plot get the same level of the whole-plot factor and no two split plots in the same whole plot get the same level of the split-plot factor. Thus a split-plot design is a restricted randomization. We have seen other restrictions on randomization; for example, RCB designs can be considered a restriction on randomization.

Split plots restrict randomization

Third, a split plot is a factorial design in incomplete blocks with one main effect confounded with blocks. The whole plots are the incomplete blocks, and the whole-plot factor is confounded with blocks. We will still be able to make inference about the whole-plot factor, because we have randomized the assignment of whole plots to levels of the whole-plot factor. This is analogous to recovering interblock information in a BIBD, but is fortunately much simpler.

Split plots confound whole-plot factor with incomplete blocks

Here is another split-plot example to help fix ideas. A statistically oriented music student performs the following experiment. Eight pianos are obtained, a baby grand and a concert grand from each of four manufacturers. Forty music majors are divided at random into eight panels of five students each. Two panels are assigned at random to each manufacturer, and will hear and rate the sound of the baby and concert grand pianos from that manufacturer. Logistically, each panel goes to the concert hall for a 30-minute time period. The panelists are seated and blindfolded. The curtain opens to reveal the two pianos of the appropriate brand, and the same piece of music is played on the two pianos in random order (the pianos are randomized, not the music!). Each panelist rates the sound on a 1–100 scale after each piece.

The whole plots are the eight panels, and the whole-plot factor is manufacturer. The split plots are the two listening sessions for each panel, and the split-plot factor is baby versus concert grand. How can we tell? We have to follow the randomization and see how treatments were assigned to units. Manufacturer was randomized to panel, and piano type was randomized to session within each panel. The randomization was restricted in such a way that both sessions for a panel had to have the same level of manufacturer.

Follow the randomization to identify a split plot

Thus panel was the unit for manufacturer, and session was the unit for type. Individual panelist is a measurement unit in this experiment, not an experimental unit. The response for any session must be some summary of the five panelist ratings.

You cannot distinguish a split-plot design from some other design simply by looking at a table of factor levels and responses. You *must* know how the randomization was done. We also have been speaking as if the whole plot randomization was done first; this is often true, but is not required.

Before moving on, we should state that the flexibility that split plots provide for dealing with factors that are difficult to vary comes at a price: comparisons involving the split-plot factor are more precise than those involving the whole-plot factor. This will be more explicit in the Hasse diagrams below, where we will see two separate error terms, the one for whole plots having a larger expectation.

<div align="right">Split-plot
comparisons
more precise than
whole-plot
comparisons</div>

16.2   Fancier Split Plots

The two examples given in the last section were the simplest possible split-plot design: the treatments have a factorial structure with two factors, levels of the whole-plot factor are assigned to whole plots in a completely randomized fashion; and levels of the split-plot factor are assigned to split plots in randomized complete block fashion with whole plots as blocks. The key to a split plot is two sizes of units and two randomizations; we can increase the number of factors and/or change the whole-plot randomization and still have a split plot.

Begin with the number of factors. The treatments assigned to whole plots need not be just the levels of a single factor: they can be the factor-level combinations of two or more factors. For example, the four piano manufacturers could actually be the two by two factorial combinations of the factors source (levels domestic and imported) and cost (levels expensive and very expensive). Here there would be two whole-plot factors. Other experiments could have more.

<div align="right">Can have more
than one
whole-plot factor</div>

Similarly, the treatments assigned to split plots at the split-plot level can be the factor-level combinations of two or more factors. The four varieties of corn could be from the combinations of the two factors insect resistant/not insect resistant, and fungus resistant/not fungus resistant. This would have two split-plot factors, and more are possible.

<div align="right">Can have more
than one split plot
factor</div>

Of course, these can be combined to have two or more factors at the whole-plot level and two or more factors at the split-plot level. The key feature of the split plot is not the number of factors, but the kind of randomization.

<div align="right">Randomization is
key</div>

Next consider the way that whole-plot treatments are assigned to whole plots. Our first examples used completely randomized design; this is not necessary. It is very common to have the whole plots grouped together into blocks, and assign whole-plot treatments to whole plots in RCB design. For

<div align="right">Whole plots
blocked in RCB</div>

example, the six whole plots in the irrigation experiment could be grouped into two blocks of three whole plots each. Then we randomly assign the three levels of irrigation to the whole plots in the first block, and then perform an independent randomization in the second block of whole plots. In this kind of design, there are two kinds of blocks: blocks of whole plots for the whole-plot treatment randomization, and whole plots acting as blocks for split plots in the split-plot treatment randomization.

We can use other designs at the whole-plot level, arranging the whole plots in Balanced Incomplete Blocks, Latin Squares, or other blocking designs. These are not common, but there is no reason that they cannot be used if the experimental situation requires it.

<div style="text-align: right">*Other block designs for whole plots*</div>

Whole plots always act as blocks for split plots. Additional blocking at the split-plot level is possible, but fairly rare. For example, we might expect a consistent difference between the first and second pianos rated by a panel. The two panels for a given manufacturer could then be run as a Latin Square, with panel as column-blocking factor and first or second session as the row-blocking factor. This would block on the additional factor time.

<div style="text-align: right">*Additional split plot blocking*</div>

16.3   Analysis of a Split Plot

Analysis of a split-plot design is fairly straightforward, once we figure out what the model should be. We assume that there is a random effect for every randomization. Thus we get a random value for each whole plot; if we ignore the split plots, we have a design with whole plot as experimental unit, and this random value is the experimental error. We also get a random value for each split plot to go with the split-plot randomization; this is experimental error at the split-plot level. Here are several examples of split plots and models for them.

<div style="text-align: right">*Random effect for every randomization*</div>

Example 16.1 Split plot with one whole-plot factor, one split-plot factor, and CRD at the whole-plot level

Suppose that there is one whole-plot factor A, with a levels, one split-plot factor B, with b levels, and n whole plots for each level of A. The model is

$$
\begin{aligned}
y_{ijk} &= \mu + \alpha_i + \eta_{k(i)} \\
&\quad + \beta_j + \alpha\beta_{ij} + \epsilon_{k(ij)} \;,
\end{aligned}
$$

with $\eta_{k(i)}$ as the whole-plot level random error, and $\epsilon_{k(ij)}$ as the split-plot level random error. Note that there is an $\eta_{k(i)}$ value for each whole plot (some whole plots have bigger responses than others), and an $\epsilon_{k(ij)}$ for each split plot. The whole-plot error term nests within whole-plot treatments in the same way that an ordinary error term nests within treatments in a CRD. In fact, if you just look at whole-plot effects (those not involving j) and ignore the split-plot effects in the second line, this model is a simple CRD on the whole plots with the whole-plot factor as treatment. Similarly, if you lump

together all the whole-plot effects in the first line and think of them as blocks, then we have a model for an RCB with the first line as block, some treatment effects, and an error.

Below are two Hasse diagrams. The first is generic and the second is for a split plot with $an = 10$ whole plots, whole-plot factor A with $a = 2$ levels, and split-plot factor B with $b = 3$ levels. The denominator for the whole-plot factor A is whole-plot error (WPE); the denominator for the split-plot factor B and the AB interaction is split-plot error (SPE).
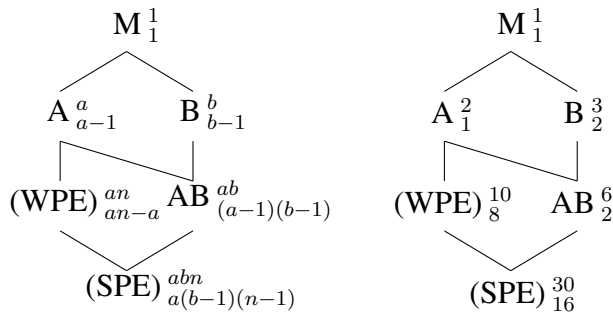
Example 16.2 Split plot with two whole-plot factors, one split-plot factor, and CRD at the whole-plot level

Now consider a split-plot design with three factors, two at the whole-plot level and one at the split-plot level. We still assume a completely randomized design for whole plots. An appropriate model for this design would be

$$
\begin{aligned}
y_{ijkl} = \ & \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \eta_{l(ij)} \\
& + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon_{l(ijk)} \ ,
\end{aligned}
$$

where we have again arranged the model into a first line with whole-plot effects (those without k) and a second line with split-plot effects. The indices i, j, and k run up to a, b, and c, the number of levels of factors A, B, and C; and the index l runs up to n, the replication at the whole-plot level.

Here are two Hasse diagrams. The first is generic for this setup, and the second is for such a split plot with $n = 5$ and whole-plot factors A and B with $a = 2$ and $b = 3$ levels, and split-plot factor C with $c = 5$ levels. The denominator for the whole-plot effects A, B, and AB is whole-plot error; the denominator for the split-plot effects C, AC, BC, and ABC is split-plot error.

$$M\,{}^{1}_{1}$$

$$A\,{}^{a}_{a-1} \qquad B\,{}^{b}_{b-1} \qquad C\,{}^{c}_{c-1}$$

$$AB\,{}^{ab}_{(a-1)(b-1)} \qquad AC\,{}^{ac}_{(a-1)(c-1)} \qquad BC\,{}^{bc}_{(b-1)(c-1)}$$

$$(WPE)\,{}^{abn}_{ab(n-1)} \qquad ABC\,{}^{abc}_{(a-1)(b-1)(c-1)}$$

$$(SPE)\,{}^{abcn}_{ab(n-1)(c-1)}$$

$$M\,{}^{1}_{1}$$

$$A\,{}^{2}_{1} \qquad B\,{}^{3}_{2} \qquad C\,{}^{5}_{4}$$

$$AB\,{}^{6}_{2} \qquad AC\,{}^{10}_{4} \qquad BC\,{}^{15}_{8}$$

$$(WPE)\,{}^{30}_{24} \qquad ABC\,{}^{30}_{8}$$

$$(SPE)\,{}^{150}_{96}$$
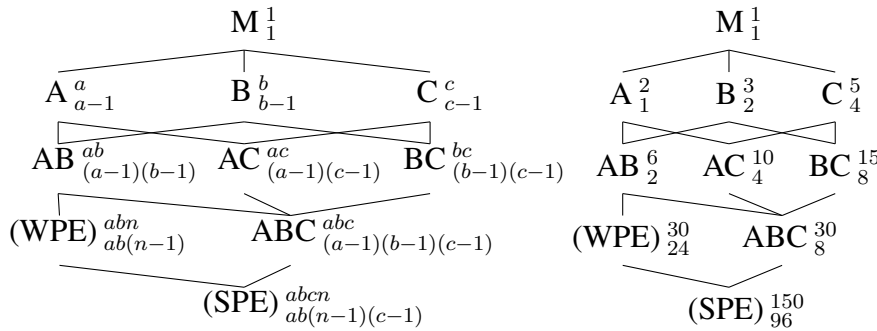
Example 16.3 Split plot with one whole-plot factor, two split-plot factors, and CRD at the whole-plot level

This split plot again has three factors, but now only one is at the whole-plot level and two are at the split-plot level. We keep a completely randomized design for whole plots. An appropriate model for this design would be

$$
\begin{aligned}
y_{ijkl} \;=\; & \mu + \alpha_i + \eta_{l(i)} \\
& + \beta_j + \alpha\beta_{ij} + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon_{l(ijk)} \;\;,
\end{aligned}
$$

where we have arranged the model into a first line with whole-plot effects (those without j or k) and a second line with split-plot effects. The indices i, j, and k run up to a, b, and c, the number of levels of factors A, B, and C; and the index l runs up to n, the amount of replication at the whole-plot level.

Below is the generic Hasse diagram for such a split plot. The denominator for the whole-plot effect A is whole-plot error; the denominator for the split plot effects B, AB, C, AC, BC, and ABC is split-plot error.

$$M\,{}^{1}_{1}$$

$$A\,{}^{a}_{a-1} \qquad B\,{}^{b}_{b-1} \qquad C\,{}^{c}_{c-1}$$

$$(WPE)\,{}^{an}_{a(n-1)} \qquad AB\,{}^{ab}_{(a-1)(b-1)} \qquad AC\,{}^{ac}_{(a-1)(c-1)} \qquad BC\,{}^{bc}_{(b-1)(c-1)}$$

$$ABC\,{}^{abc}_{(a-1)(b-1)(c-1)}$$

$$(SPE)\,{}^{abcn}_{a(n-1)(bc-1)}$$

Draft of March 4, 2021

Example 16.4 Split plot with one whole-plot factor, one split-plot factor, and RCB at the whole-plot level
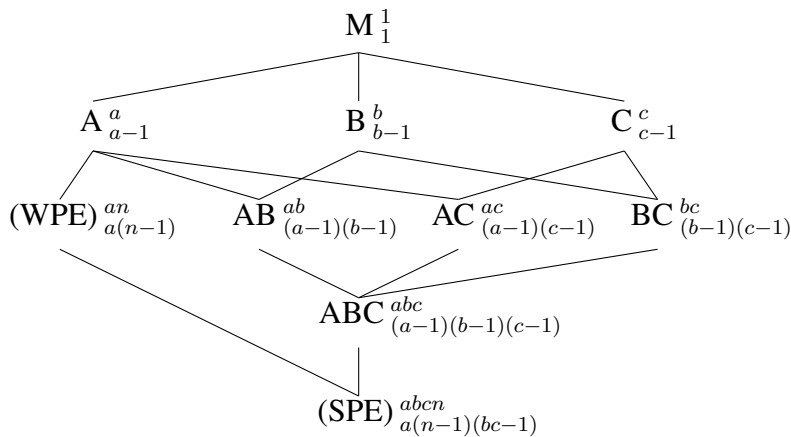
Now consider a split-plot design with two factors, one at the whole-plot level and one at the split-plot level, but use a block design for the whole plots. An appropriate model for this design would be

$$
\begin{aligned}
y_{ijkl} \;=\; & \mu + \alpha_i + \gamma_k + \eta_{l(ik)} \\
& + \beta_j + \alpha\beta_{ij} + \epsilon_{l(ijk)} \;,
\end{aligned}
$$

where we have again arranged the model into a first line with whole-plot effects (those without j) and a second line with split-plot effects. The indices i and j run up to a and b, the number of levels of factors A and B; the index k runs up to n, the number of blocks at the whole-plot level; and the index l runs up to 1, the number of whole plots in each block getting a given whole-plot treatment or the number of split plots in each whole plot getting a given split-plot treatment. Thus the model assumes that block effects are fixed and additive with whole-plot treatments, and there is a random error for each whole plot. This is just the standard RCB model applied to the whole plots.

Below is a generic Hasse diagram for a blocked split plot and a sample Hasse diagram for a split plot with $n = 5$ blocks and whole-plot factor A with $a = 2$ levels, and split-plot factor B with $b = 3$ levels. The denominator for the whole-plot effect A is whole-plot error; the denominator for the split-plot effects B and AB is split-plot error.
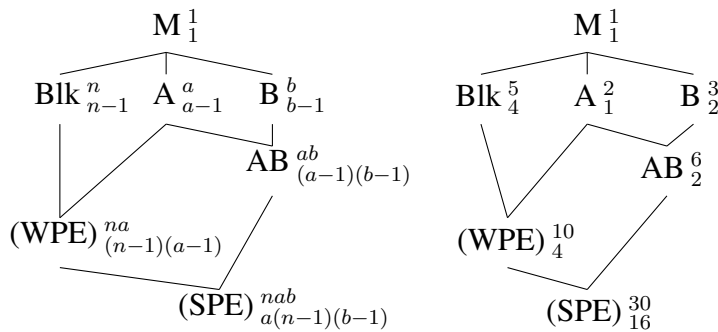
This model assumes that blocks are additive. If we allow a block by whole-plot factor interaction, then there will be no degrees of freedom for whole-plot error, and we will need to use the block by whole-plot factor interaction as surrogate error for whole-plot factor.

We can use our standard methods for mixed-effects factorials from Chapter 11 to analyze split-plot designs using these split-plot models. Alternatively, we can achieve the same results using the following heuristic approach. A split plot has two sizes of units and two randomizations, so first split the variation in the data into two bundles, the variation between whole plots and the variation within whole plots (between split plots). Using a simple split-plot design with just two factors, there are an whole plots and

Partition variation into between and within whole plots

$N - 1 = abn - 1$ degrees of freedom between all the responses. We can get the variation between whole plots by considering the whole plots to be an "treatment groups" of b units each and doing an ordinary one-way ANOVA. There are thus $an - 1$ degrees of freedom between the whole plots and $(abn - 1) - (an - 1) = an(b - 1)$ degrees of freedom within whole plots, between split plots. Visualize this decomposition as:

$$\text{Total}_{(N-1)}$$
$$\text{Between WP}_{an-1} \qquad \text{Within WP}_{an(b-1)}$$
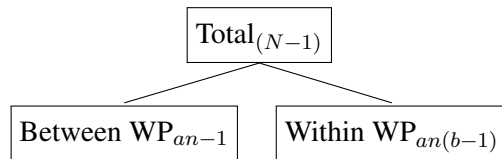
The between whole plots variation is made up of effects that affect complete whole plots. These include the whole-plot treatment factor(s), whole-plot error, and any blocking that might have been done at the whole-plot level. This variation yields the following decomposition, assuming the whole plots were blocked.

Whole-plot variation includes blocks, whole-plot factor, and whole-plot error

$$\text{Between WP}_{an-1}$$
$$\text{Blocks}_{n-1} \quad \text{A}_{a-1} \quad \text{WPE}_{(a-1)(n-1)}$$
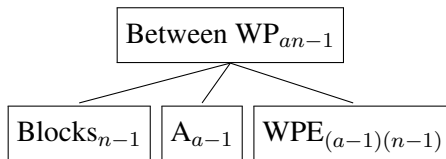
The variation between split plots (within whole plots) is variation in the responses that depends on effects that affect individual split plots, including the split-plot treatment factor(s), interaction between whole-plot and split-plot treatment factors, and split-plot error. The variation is decomposed as

Split-plot variation includes split-plot factor, whole-by-split-factor interaction, and split-plot error

$$\text{Within WP}_{an(b-1)}$$
$$\text{B}_{b-1} \quad \text{AB}_{(a-1)(b-1)} \quad \text{SPE}_{a(b-1)(n-1)}$$
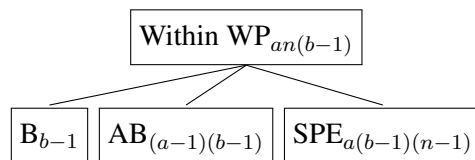
The easiest way to get the degrees of freedom for split-plot error is by subtraction. There are $an(b - 1)$ degrees of freedom between split plots within whole plots; $b - 1$ of these go to B, $(a - 1)(b - 1)$ go to AB, and the remainder must be split-plot error.

Get df by subtraction

It may not be obvious why the interaction between the whole- and split-plot factors should be a split-plot level effect. Recall that one way to describe this interaction is how the split-plot treatment effects change as we vary the whole-plot treatment. Because this is dealing with changing split-plot treatment levels, this effect cannot be at the whole-plot level; it must be lower.

Interaction at split-plot level

Assembling the pieces, we get the overall decomposition:

Table 16.1: Number of memory errors by type, tension, and anxiety level; subjects are columns.

	Anxiety/Tension											
	1	1	1	1	1	1	2	2	2	2	2	2
Type	1	1	1	2	2	2	1	1	1	2	2	2
1	18	19	14	16	12	18	16	18	16	19	16	16
2	14	12	10	12	8	10	10	8	12	16	14	12
3	12	8	6	10	6	5	8	4	6	10	10	8
4	6	4	2	4	2	1	4	1	2	8	9	8

$$\text{Total}_{(N-1)}$$

$$\text{Between WP}_{an-1} \qquad \text{Within WP}_{an(b-1)}$$

$$\text{Blk}_{n-1} \quad \text{A}_{a-1} \quad \text{WPE}_{(a-1)(n-1)} \quad \text{B}_{b-1} \quad \text{AB}_{(a-1)(b-1)} \quad \text{SPE}_{a(b-1)(n-1)}$$
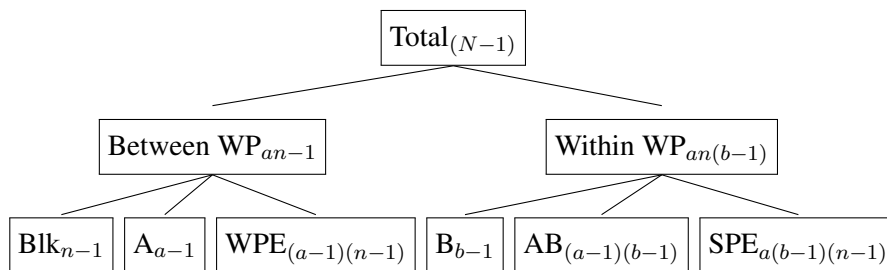
I find that this decomposition gives me a little more understanding about what is going on in the split-plot analysis than just looking at the Hasse diagram.

We compute sums of squares and estimates of treatment effects in the usual way. When it is time for testing or computing standard errors for contrasts, effects at the split-plot level use the split-plot error with its degrees of freedom, and effects at the whole-plot level use the whole-plot error with its degrees of freedom.

Example 16.5 Anxiety, tension, and memory

We wish to study the effects of anxiety and muscular tension on four different types of memory. Twelve subjects are assigned to one of four anxiety-tension combinations at random. The low-anxiety group is told that they will be awarded \$5 for participation and \$10 if they remember sufficiently accurately, and the high-anxiety group is told that they will be awarded \$5 for participation and \$100 if they remember sufficiently accurately. Everyone must squeeze a spring-loaded grip to keep a buzzer from sounding during the testing period. The high-tension group must squeeze against a stronger spring than the low-tension group. All subjects then perform four memory trials in random order, testing four different types of memory. The response is the number of errors on each memory trial, as shown in Table 16.1 (data set `Anxiety`).

This is a split-plot design. There are two separate randomizations. We first randomly assign the anxiety-tension combinations to each subject. Even though we will have four responses from each subject, the randomization is restricted so that all four of those responses will be at the same anxiety-tension combination. Anxiety and tension are thus whole-plot treatment fac-

tors. Each subject will do four memory trials. The trial type is randomized to the four trials for a given subject. Thus the four trials for a subject are the split plots, and the trial type is the split-plot treatment. At the whole-plot level, the anxiety-tension combinations are assigned according to a CRD, so there is no blocking.

Here is some Minitab output from an analysis of these data.

```
ource                      DF     Seq SS     Adj SS     Adj MS        F      P
nxiety                      1     10.083     10.083     10.083     0.98  0.352
ension                      1      8.333      8.333      8.333     0.81  0.395
nxiety*tension              1     80.083     80.083     80.083     7.77  0.024
ubject(anxiety tension)     8     82.500     82.500     10.312     4.74  0.001
ype                         3    991.500    991.500    330.500   152.05  0.000
nxiety*type                 3      8.417      8.417      2.806     1.29  0.300
ension*type                 3     12.167     12.167      4.056     1.87  0.162
nxiety*tension*type         3     12.750     12.750      4.250     1.96  0.148
rror                       24     52.167     52.167      2.174
```

The ANOVA table has been arranged so that the whole-plot analysis is on top and the split-plot analysis below, as is customary. The whole-plot error is shown as *subject* nested in *anxiety* and *tension,* and the split-plot error is just denoted *Error.* Note that the split-plot error is smaller than the whole-plot error by a factor of nearly 5. Subject to subject variation is not negligible, and split-plot comparisons, which are made with subjects as blocks, are much more precise than whole-plot comparisons, where subjects are units.

At the split-plot level, the effect of type is highly significant. All the type effects γ_k differ from each other by more than 3, and the standard error of the difference of two type means is $\sqrt{2.174(1/12 + 1/12)} = .602$. Thus all type means are at least 5 standard errors apart and can be distinguished from each other. No interactions with type appear to be significant.

Analysis at the whole-plot level is more ambiguous. The main effects of anxiety and tension are both nonsignificant, but their interaction is moderately significant. Figure 16.1 shows an interaction plot for anxiety and tension. We see that more errors occur when anxiety and tension are both low or both high. With such strong interaction, it makes sense to examine the treatment means themselves. The greatest difference between the four whole plot treatment means is 3.5, and the standard error for a difference of two means is $\sqrt{10.312(1/12 + 1/12)} = 1.311$. This is only a bit more than 2.5 standard errors and is not significant after adjusting for multiple comparisons; for example, the Bonferroni p-value is .17. This is in accordance with the result we obtain by considering the four whole-plot treatments to be a single factor with four levels. Pooling sums of squares and degrees of freedom for anxiety, tension, and their interaction, we get a mean square of 32.83 with 3 degrees of freedom and a p-value of .08.

The residuals-versus-predicted plot shows slight nonconstant variance; no transformation makes much improvement, so the data have been analyzed on the original scale.

In conclusion, there is strong evidence that the number of errors differs between memory type. There is no evidence that this difference depends on
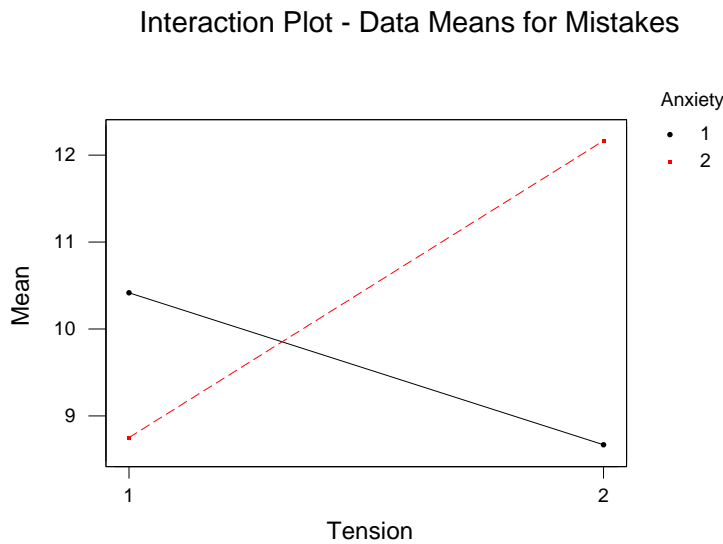
Figure 16.1: Anxiety by tension interaction plot for memory errors data, using Minitab.

anxiety or tension individually. There is mild evidence that there are more errors when anxiety and tension are both high or both low, but none of the actual anxiety-tension combinations can be distinguished.

Let me note here that some authors prefer an alternate model for the split plot with one whole-plot factor, one split-plot factor, and RCB structure on the whole plots. This model assumes that blocks are a random effect that interact with all other factors; effectively this is a three-way factorial model with one random factor.

Alternate model
has blocks
random and
interacting

16.4   Split-Split Plots

What we have split once, we can split again. Consider an experiment with three factors. The levels of factor A are assigned at random to n whole plots each (total of an whole plots). Each whole plot is split into b split plots. The levels of factor B are assigned at random to split plots, using whole plots as blocks. So far, this is just like a split-plot design. Now each split plot is divided into c split-split plots, and the levels of factor C are randomly assigned to split-split plots using split plots as blocks. Obviously, once we get used to splitting, we can split again for a fourth factor, and keep on going.

Split the split plots

Split-split plots arise for the same reasons as ordinary split plots: some factors are easier to vary than others. For example, consider a chemical experiment where we study the effects of the type of feedstock, the temperature

of the reaction, and the duration of the reaction on yield. Some experimental setups require extensive cleaning between different feedstocks, so we might wish to vary the feedstock as infrequently as possible. Similarly, there may be some delay that must occur when the temperature is changed to allow the equipment to equilibrate at the new temperature. In such a situation, we might choose type of feedstock as the whole-plot factor, temperature of reaction as the split-plot factor, and duration of reaction as the split-split-plot factor. This makes our experiment more feasible logistically, because we have fewer cleanups and temperature delays; comparisons involving time will be more precise than those for temperature, which are themselves more precise than those for feedstock.

Split-split plots have three sizes of units. Whole plots act as unit for the whole-plot treatments. Whole plots act as blocks for split plots, and split plots act as unit for the split-plot treatments. Split plots act as blocks for split-split plots, and split-split plots act as unit for the split-split-plot treatments. The whole plots can be blocked, just as in the split plot.
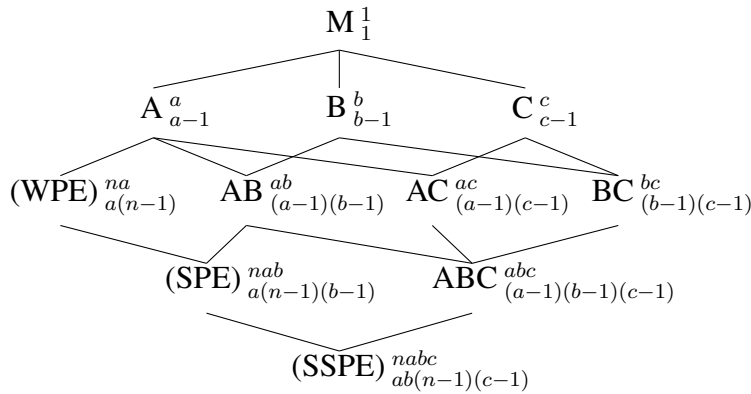
Example 16.6 Split-split plot with one whole-plot factor, one split-plot factor, one split-split-plot factor and CRD at the whole plot level

Now consider a split-split-plot design with three factors, one at the whole-plot level, one at the split-plot level, and one at the split-split-plot level, with a completely randomized design for whole plots. An appropriate model for this design would be

$$
\begin{aligned}
y_{ijkl} \;=\;& \mu + \alpha_i + \eta_{l(i)} \\
& + \beta_j + \alpha\beta_{ij} + \zeta_{l(ij)} \\
& + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon_{l(ijk)} \;\;,
\end{aligned}
$$

where we have arranged the model into a first line with whole-plot effects (those without j or k), a second line with split-plot effects (those with j but not k), and the last line with split-split-plot effects. The indices i, j, and k run up to a, b, and c, the number of levels of factors A, B, and C; and the index l runs up to n, the amount of replication at the whole plot level.

Below is a Hasse diagram for this generic split-split plot with three factors and a CRD at the whole-plot level. The denominator for the whole-plot effect A is whole-plot error; the denominator for the split-plot effects B and AB is the split-plot error; and the denominator for the split-split-plot effects C, AC, BC, and ABC is split-split-plot error (SSPE).

$$\text{M}\,^1_1$$

$$\text{A}\,^a_{a-1} \qquad \text{B}\,^b_{b-1} \qquad \text{C}\,^c_{c-1}$$

$$(\text{WPE})\,^{na}_{a(n-1)} \quad \text{AB}\,^{ab}_{(a-1)(b-1)} \quad \text{AC}\,^{ac}_{(a-1)(c-1)} \quad \text{BC}\,^{bc}_{(b-1)(c-1)}$$

$$(\text{SPE})\,^{nab}_{a(n-1)(b-1)} \qquad \text{ABC}\,^{abc}_{(a-1)(b-1)(c-1)}$$

$$(\text{SSPE})\,^{nabc}_{ab(n-1)(c-1)}$$

A split-split plot has at least three treatment factors, but it can have more than three. Any of whole-, split-, or split-split-plot treatments can have factorial structure. Thus you cannot distinguish a split plot from a split-split plot or other design solely on the basis of the number of factors; the units and randomization determine the design.

*Randomization, not number of factors, determines design*

Analysis of a split-split plot can be conducted using standard methods for mixed-effects factorials, but I find that a graphical partitioning of degrees of freedom and their associated sums of squares helps me understand what is going on. Consider three factors with a, b, and c levels, in a split-split-plot design with n replications. Begin the decomposition just as for a split plot:

Partition variation between levels of the design

$$\boxed{\text{Total}_{(abcn-1)}}$$

$$\boxed{\text{Between WP}_{an-1}} \qquad \boxed{\text{Within WP}_{an(bc-1)}}$$
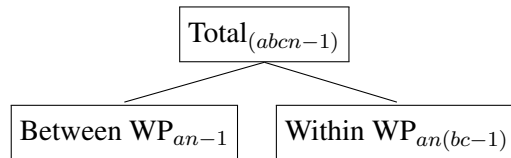
The only difference between this and a split-plot design is that we have $bc-1$ degrees of freedom within each whole plot, because each whole plot is a bundle of bc split-split-plot values instead of just b split-plot values.

The between whole plots variation partitions in the same way as for a split-plot design. For example, with blocking we get:

$$\boxed{\text{Between WP}_{an-1}}$$

$$\boxed{\text{Blocks}_{n-1}} \quad \boxed{\text{A}_{a-1}} \quad \boxed{\text{WPE}_{(a-1)(n-1)}}$$
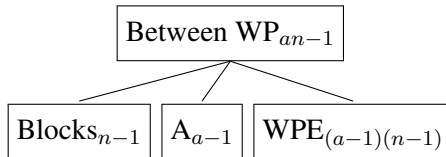
Variation within whole plots can be divided into variation between split plots and variation between split-split plots within the split plots. This is like split plots as block variation, and split-split plots as unit to unit within block variation. This partition is:

Between and within split plots

```
          ┌─────────────────────┐
          │ Within WP_{an(bc-1)} │
          └─────────────────────┘
             ╱              ╲
┌───────────────────────┐ ┌──────────────────────┐
│ Between SP_{an(b-1)}  │ │ Within SP_{abn(c-1)} │
└───────────────────────┘ └──────────────────────┘
```
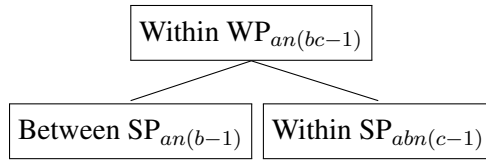
There are b split plots in each whole plot, so $b-1$ degrees of freedom between split plots in a single whole plot, and $an(b-1)$ total degrees of freedom between split plots within whole plots. There are c split-split plots in each split plot, so $c-1$ degrees of freedom between split-split plots in a single split plot, and $abn(c-1)$ total degrees of freedom between split-split plots within a split plot.

The variation between split plots within whole plots is partitioned just as for a split-plot design:
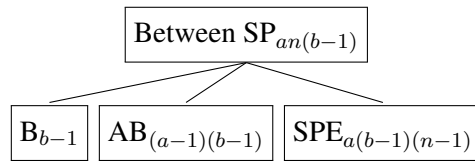
```
                ┌──────────────────────┐
                │ Between SP_{an(b-1)}  │
                └──────────────────────┘
              ╱            │            ╲
    ┌─────────┐ ┌──────────────────┐ ┌───────────────────┐
    │ B_{b-1} │ │ AB_{(a-1)(b-1)}  │ │ SPE_{a(b-1)(n-1)} │
    └─────────┘ └──────────────────┘ └───────────────────┘
```

Finally, we come to the variation between split-split plots within split plots. This is variation due to factor C and its interactions, and split-split-plot error:

```
                      ┌──────────────────────┐
                      │ Within SP_{abn(c-1)} │
                      └──────────────────────┘
          ╱         ╱          │          ╲          ╲
┌─────────┐ ┌──────────────┐ ┌──────────────┐ ┌──────────────────────┐ ┌─────────────────────┐
│ C_{c-1} │ │ AC_{(a-1)(c-1)} │ │ BC_{(b-1)(c-1)} │ │ ABC_{(a-1)(b-1)(c-1)} │ │ SSPE_{ab(c-1)(n-1)} │
└─────────┘ └──────────────┘ └──────────────┘ └──────────────────────┘ └─────────────────────┘
```
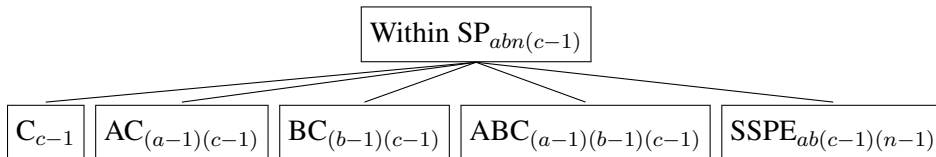
Table 16.2: Percent of wetland biomass that is nonweed, by table (T), nitrogen (N), weed (W), and clipping (C).

T	N	W 1 C 1	W 1 C 2	W 2 C 1	W 2 C 2	W 3 C 1	W 3 C 2
1	1	87.2	88.8	70.4	75.7	75.9	80.6
	2	80.5	83.8	59.2	61.5	59.5	62.5
	3	76.8	80.8	47.8	49.5	48.4	52.9
	4	77.7	81.5	35.7	37.3	38.3	42.4
2	1	78.2	80.5	65.1	68.3	65.3	66.6
	2	79.8	85.2	57.6	61.4	58.5	61.6
	3	82.4	83.1	50.5	54.0	51.6	54.7
	4	75.5	78.7	39.0	43.9	41.9	45.1

Example 16.7 Weed biomass in wetlands

An experiment studies the effect of nitrogen and weeds on plant growth in wetlands. We investigate four levels of nitrogen, three weed treatments (no additional weeds, addition of weed species 1, addition of weed species 2), and two herbivory treatments (clipping and no clipping). We have eight trays; each tray holds three artificial wetlands consisting of rectangular wire baskets containing wetland soil. The trays are full of water, so the artificial wetlands stay wet. All of the artificial wetlands receive a standard set of seeds to start growth.

Four of the trays are placed on a table near the door of the greenhouse, and the other four trays are placed on a table in the center of the greenhouse. On each table, we randomly assign one of the trays to each of the four nitrogen treatments. Within each tray, we randomly assign the wetlands to the three weed treatments. Each wetland is split in half. One half is chosen at random and will be clipped after 4 weeks, with the clippings removed; the other half is not clipped. After 8 weeks, we measure the fraction of biomass in each wetland that is nonweed as our response. Responses are given in Table 16.2, data set `WeedBiomass`.

This is a split-split-plot design. Everything in a given tray has the same level of nitrogen, so the trays are whole plots, and nitrogen is the whole-plot factor. The whole plots are arranged in two blocks, with table as block accounting for any differences between the door and center of the greenhouse. Both measurements for a given wetland have the same weed treatment, so the wetlands are split plots, and weed is the split-plot factor. Finally each wetland half gets its own clipping treatment, so wetland halves are split-split plots, and clipping is the split-split-plot factor.

Here is some SAS output for these data.

ource	DF	Seq SS	Adj SS	Adj MS	F	P
nxiety	1	10.083	10.083	10.083	0.98	0.352
ension	1	8.333	8.333	8.333	0.81	0.395
nxiety*tension	1	80.083	80.083	80.083	7.77	0.024
ubject(anxiety tension)	8	82.500	82.500	10.312	4.74	0.001
ype	3	991.500	991.500	330.500	152.05	0.000
nxiety*type	3	8.417	8.417	2.806	1.29	0.300
ension*type	3	12.167	12.167	4.056	1.87	0.162
nxiety*tension*type	3	12.750	12.750	4.250	1.96	0.148
rror	24	52.167	52.167	2.174		

Notice that F-ratios and p-values in the ANOVA table use the 12-degree-of-freedom error term as denominator. This is correct for split-split-plot terms (those including clipping), but is incorrect for whole-plot and split-plot terms. Those must be tested separately in SAS by specifying the appropriate denominators. This is important, because the whole-plot error mean square is about 15 times as big as the split-plot error mean square, which is about 6 times as big as the split-split-plot mean square.

All main effects and the nitrogen by weed interaction are significant. An interaction plot for nitrogen and weed shows the nature of the interaction, Figure 16.2. Weeds do better as nitrogen is introduced, but the effect is much larger when the weeds have been seeded. Clipping slightly increases the fraction of nonweed biomass.

Residual plots show that the variance increases somewhat with the mean, but no reasonable transformation fixes the problem.

16.5   Other Generalizations of Split Plots

One way to think about split plots is that the units have a structure somewhat like that of nested factorial treatments. In a split plot, the split plots are nested in whole plots; in a split-split plot, the split-split plots are nested in split plots, which are themselves nested in whole plots. In the split-plot design, levels of different factors are assigned to the different kinds of units. This section deals with some other unit structures that are possible.

Other unit
structures
besides nesting
are possible

Example 16.8 Machine shop

Consider a machine shop that is producing parts cut from metal blanks. The quality of the parts is determined by their strength and fidelity to the desired shape. The shop wishes to determine how brand of cutting tool and supplier of metal blank affect the quality. An experiment will be performed one week, and then repeated the next week. Four brands of cutting tools will be obtained, and brand of tool will be randomly assigned to four lathes. A different supplier of metal blank will be randomly selected for each of the 5 work days during the week. That way, all brand-supplier combinations are observed.

A schematic for the design might look like this:

Interaction Plot - Data Means for y
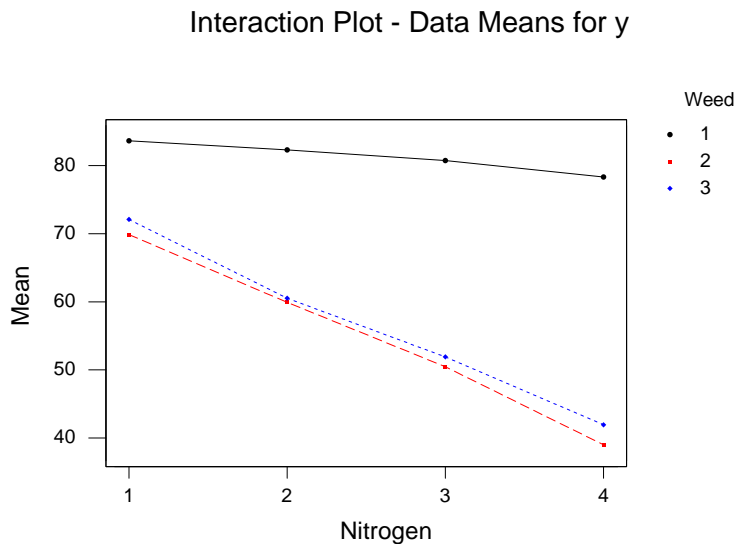
Figure 16.2: Nitrogen by weed interaction plot for for wetland weeds data, using Minitab.

	Day 1	Day 2	Day 3	Day 4	Day 5
Lathe 1	Br 3 Sp 5	Br 3 Sp 1	Br 3 Sp 2	Br 3 Sp 4	Br 3 Sp 3
Lathe 2	Br 2 Sp 5	Br 2 Sp 1	Br 2 Sp 2	Br 2 Sp 4	Br 2 Sp 3
Lathe 3	Br 1 Sp 5	Br 1 Sp 1	Br 1 Sp 2	Br 1 Sp 4	Br 1 Sp 3
Lathe 4	Br 4 Sp 5	Br 4 Sp 1	Br 4 Sp 2	Br 4 Sp 4	Br 4 Sp 3

The table shows the combinations of the four lathes and 5 days. Brand is assigned to lathe, or row of the table. Thus the unit for brand is lathe. Supplier of blanks is assigned to day, or column of the table. Thus the unit for supplier is day. There are two separate randomizations done in this design to two different kinds of units, but this is not a split plot, because here the units do not nest as they would in a split plot.

The design used in the machine shop example has been given a couple of different names, including *strip plot* and *split block*. What we have in a strip plot is two different kinds of units, with levels of factors assigned to each unit, but the units *cross* each other. This is in contrast to the split plot, where the units nest.

Strip plot or split block, with units that cross

Like the split plot, the strip plot arises through ease-of-use considerations. It is easier to use one brand of tool on each lathe than it is to change. Similarly, it is easier to use one supplier all day than to change suppliers during the day. When units are large and treatments difficult to change, but the units and treatments can cross, a strip plot can be the design of choice.

Strip plot easy to use

The usual assumptions in model building for split plots and related designs such as strip plots are that there is a random term for each kind of unit, or kind of randomization if you prefer, and there is a random term whenever two units cross. For the split plot, there is a random term for whole plots that we call whole-plot error, and a random term for split plots that we call split-plot error. There are no further random terms because the unit structure in a whole plot does not cross; it nests.

For the strip plot, there is a random term for rows and a random term for columns, because these are the two basic units. There is also a random term for each row-column combination, because this is where two units cross. For the machine tool example, we have the model

$$\begin{aligned} y_{ijkl} \quad = \quad & \mu + \gamma_k + \alpha_i + \eta_{l(ik)} + \\ & \beta_j + \zeta_{l(jk)} + \\ & \alpha\beta_{ij} + \epsilon_{l(ijk)} \quad , \end{aligned}$$

where i and j index the levels of brand and supplier, k indexes the week (weeks are acting as blocks), and l is always 1 and indicates a particular unit for a block-treatment-unit size combination. The term $\eta_{l(ik)}$ is the random effect for machine to machine (row to row) differences within a week; the term $\zeta_{l(jk)}$ is the random effect for day to day (column to column) differences within a week; $\epsilon_{l(ijk)}$ is unit experimental error.

Here is a Hasse diagram for the machine shop example. We denote brand and supplier by B and S; R and C denote the row and column random effects.

$$M\,^1_1$$

$$\mathrm{Blk}\,^2_1 \qquad B\,^4_3 \qquad S\,^5_4$$

$$(R)\,^8_3 \qquad (C)\,^{10}_4 \qquad BS\,^{20}_{12}$$

$$(RC)\,^{40}_{12}$$
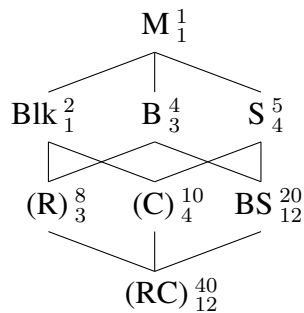
We can see from the Hasse diagram that row and column mean squares tend to be larger than the error for individual cells. This means that a strip plot experiment has less precise comparisons and lower power for main effects, and more precision and power for interactions.
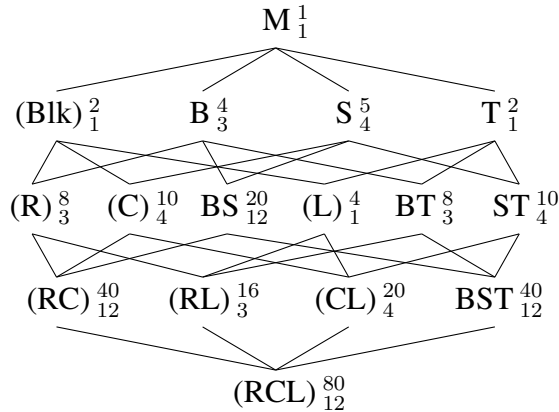
When we saw that treatment factors could cross or nest, a whole world of new treatment structures opened to us. Many combinations of crossing and nesting were useful in different situations. The same is true for unit structures—we can construct more diverse designs by combining nesting and crossing of units. Just as with the split plot and strip plot, these unit structures usually arise through ease-of-use requirements.

Now extend the machine tool example by supposing that in addition to four brands of tool, there are also two types. Brands of tool are assigned

to each lathe at random as before, but we now assign at random the first or second tool type to morning or afternoon use. If all the lathes use the same type of tool in the morning and the other type in the afternoon, then our units have a three-way crossing structure, with lathe, day, and hour being rows, columns, and layers in a three-way table. There will be separate random terms for each unit type (lathe, day, and hour) and for each crossing of unit types (lathe by day, lathe by hour, day by hour, and lathe by day by hour).

$$M\,{}^1_1$$

$$(Blk)\,{}^2_1 \qquad B\,{}^4_3 \qquad S\,{}^5_4 \qquad T\,{}^2_1$$

$$(R)\,{}^8_3 \quad (C)\,{}^{10}_4 \quad BS\,{}^{20}_{12} \quad (L)\,{}^4_1 \quad BT\,{}^8_3 \quad ST\,{}^{10}_4$$

$$(RC)\,{}^{40}_{12} \quad (RL)\,{}^{16}_3 \quad (CL)\,{}^{20}_4 \quad BST\,{}^{40}_{12}$$

$$(RCL)\,{}^{80}_{12}$$

In the Hasse diagram, R, C, and L are the random effects for rows, columns, and layers (lathes, days, and hours). The interaction RCL cannot be distinguished from the usual experimental error E. The appropriate test denominators are
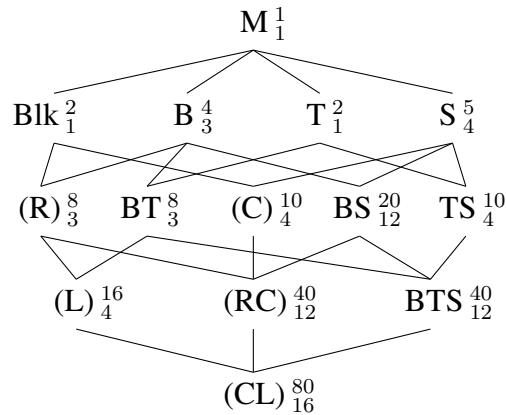
Term	B	S	T	BS	BT	ST	BST
Denominator	R	C	L	RC	RL	CL	RCL

Alternatively, suppose that instead of using the same type of tool for all lathes in the mornings and afternoons, we instead randomize types to morning or afternoon separately for each lathe. Then ignoring supplier and day, we have hour units nested in lathe units, so that the experiment is a split plot in brand and type. Overall we have three treatment factors, all crossed, and unit structure hour nested in lathe and crossed with day. This is a split plot (in brand and type, with lathe as whole plot, time as split plot, and week as block) crossed with an RCB (in supplier, with day as unit and week as block).

The Hasse diagram for this setup is on the next page. In the Hasse diagram, R, C, and L are the random effects for rows, columns, and layers (lathes, days, and hours). The layer effects L (hours) are nested in rows (lathes). Again, the interaction CL cannot be distinguished from the usual experimental error E. The appropriate test denominators are

Term	B	T	BT	S	BS	TS	BTS
Denominator	R	L	L	C	RC	CL	CL

$$M\,{}^{1}_{1}$$

$$Blk\,{}^{2}_{1} \qquad B\,{}^{4}_{3} \qquad T\,{}^{2}_{1} \qquad S\,{}^{5}_{4}$$

$$(R)\,{}^{8}_{3} \quad BT\,{}^{8}_{3} \quad (C)\,{}^{10}_{4} \quad BS\,{}^{20}_{12} \quad TS\,{}^{10}_{4}$$

$$(L)\,{}^{16}_{4} \qquad (RC)\,{}^{40}_{12} \qquad BTS\,{}^{40}_{12}$$

$$(CL)\,{}^{80}_{16}$$

16.6 Repeated Measures

Consider the following experiment, which looks similar to a split-plot design but lacks an important ingredient. We wish to study the effects of different infant formulas and time on infant growth. Thirty newborns are assigned at random to three different infant formulas. (All the formulas are believed to provide adequate nutrition, and informed consent of the parents is obtained.) The weights of the infants are measured at birth, 1 week, 4 weeks, 2 months, and 6 months. The main effect of time is expected; the research questions relate to the main effect of formula and interaction between time and formula.

This looks a little like a split-plot design, with infant as whole plot and formula as whole-plot treatment, and infant time periods as split plot and age as split-plot treatment. However, this is not a split-plot design, because age was not randomized; indeed, age cannot be randomized. A split-plot design has two sizes of units and two randomizations. This experiment has two sizes of units, but only one randomization.

Split plot needs two randomizations

This is the prototypical *repeated-measures* design. The jargon used in repeated measures is a bit different from split plots. Whole plots are usually called "subjects," whole-plot treatment factors are called "grouping factors" or "between subjects factors," and split-plot treatment factors are called "repeated measures" or "within subjects factors" or "trial factors." In a repeated-measures design, the grouping factors are randomized to the subjects, but the repeated measures are not randomized. The example has a single grouping factor applied to subjects in a completely randomized fashion, but there could be multiple grouping factors, and the subject level design could include blocking.

Repeated measures have only one randomization

What we really have with a repeated-measures design is that subjects are units, and every unit has a *multivariate* response. That is, instead of a single response, every subject has a whole vector of responses, with one element for each repeated measure. Thus, each infant in the example above has a response that is a vector of length 5, giving weights at the five ages.

Repeated measures have multivariate response

The challenge presented by repeated measures is that the components in a vector of responses tend to be correlated, not independent, and every pair of repeated measures could have a different correlation. This correlation is both a blessing and a curse. It is a blessing because within-subject correlation makes comparisons between repeated measures more precise, in the same way that blocking makes treatment comparisons more precise. It is a curse because correlation complicates the analysis.

Correlated responses can improve precision but complicate analysis

There are three basic choices for the analysis of repeated-measures designs. First, you can do a full multivariate analysis, though such an analysis is beyond the scope of this text. Second, you can make a suitable univariate summary of the data for each subject, and then use these summaries as the response in a standard analysis. For the infant formula example, we could calculate the average growth rate for each infant and then analyze these as responses in a CRD with three treatments, or we could simply use the 6 month weight as response to see if the formulas have any effect on weight after 6 months. In fact, most experiments have more than one response, which we usually analyze separately; the trick comes in analyzing more than one response at a time.

Multivariate analysis

Univariate summaries

The third method is to analyze the data with a suitable ANOVA model. The applicability of the third method depends on whether nature has been kind to us: if the correlation structure of the responses meets certain requirements, then we can ignore the correlation and get a proper analysis using univariate mixed-effects models and ANOVA. For example, if all the repeated measures have the same variance, and all pairs of repeated measures have the same correlation (a condition called *compound symmetry*), then we can get an appropriate analysis by treating the repeated-measures design as if it were a split-plot design. Another important case is when there are only two repeated measures; then the requirements are always met. Thus you can always use the standard split-plot type analysis when there are only two repeated measures. When the ANOVA model is appropriate, it provides more powerful tests than the multivariate procedures.

Univariate ANOVA works in some cases, such as compound symmetry, or two repeated measures

The mysterious "certain requirements" mentioned above are called the Huynh-Feldt condition or circularity, and it states that all differences of repeated measures have the same variance. For example, compound symmetry implies the Huynh-Feldt condition. There is a test for the Huynh-Feldt condition, called the Mauchly test for sphericity, but it is very dependent on normality in the same way that most classical tests of equal variance are dependent on normality.

Huynh-Feldt condition and Mauchly test

The standard model in a univariate analysis of repeated measures assumes that there is a random effect for each subject, and that this random effect interacts with all repeated-measures effects and their interactions, but not with the grouping by repeated interactions. For example, consider a model
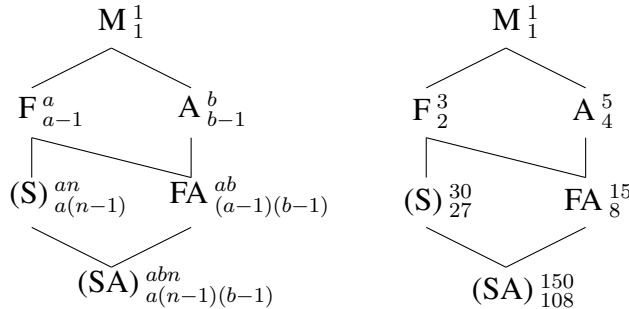
Random subject effect interacts with trial factors

Draft of March 4, 2021

for the infant weights:

$$
\begin{aligned}
y_{ijk} &= \mu + \alpha_i + \epsilon_{k(i)} + \\
&\quad \beta_j + \alpha\beta_{ij} + \epsilon\beta_{jk(i)} \;.
\end{aligned}
$$

The term α_i is the formula effect (F), and $\epsilon_{k(i)}$ is the subject random effect (S); effect β_j is age (A), and $\epsilon\beta_{jk(i)}$ is the interaction of age and subject.

$$
\begin{array}{cc}
\mathrm{M}\,^{1}_{1} & \mathrm{M}\,^{1}_{1} \\[4pt]
\mathrm{F}\,^{a}_{a-1} \qquad \mathrm{A}\,^{b}_{b-1} & \mathrm{F}\,^{3}_{2} \qquad \mathrm{A}\,^{5}_{4} \\[4pt]
(\mathrm{S})\,^{an}_{a(n-1)} \quad \mathrm{FA}\,^{ab}_{(a-1)(b-1)} & (\mathrm{S})\,^{30}_{27} \quad \mathrm{FA}\,^{15}_{8} \\[4pt]
(\mathrm{SA})\,^{abn}_{a(n-1)(b-1)} & (\mathrm{SA})\,^{150}_{108}
\end{array}
$$

We see that formula is tested against subject, and age and the formula by age interaction are tested against the subject by age interaction. This analysis is just like a split-plot design.
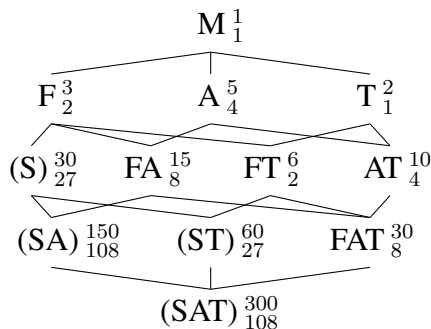
One trial factor is like split plot

Suppose now that the infants are weighed twice at each age, using two different techniques. Now the model looks like

$$
\begin{aligned}
y_{ijkl} &= \mu + \alpha_i + \epsilon_{l(i)} + \\
&\quad \beta_j + \alpha\beta_{ij} + \epsilon\beta_{jl(i)} + \\
&\quad \gamma_k + \alpha\gamma_{ik} + \epsilon\gamma_{kl(i)} + \\
&\quad \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon\beta\gamma_{jkl(i)} \;.
\end{aligned}
$$

The repeated measures effects are β_j for age, γ_k for measurement technique (T), and $\beta\gamma_{jk}$ for their interaction. Each of these is assumed to interact with the subject effect $\epsilon_{l(i)}$. This leads to the error structure shown in the Hasse diagram below, which is unlike either a split-plot design with two factors at the split-plot level or a split-split plot.

Two trial factors unlike split plot

$$
\begin{array}{c}
\mathrm{M}\,^{1}_{1} \\[4pt]
\mathrm{F}\,^{3}_{2} \qquad \mathrm{A}\,^{5}_{4} \qquad \mathrm{T}\,^{2}_{1} \\[4pt]
(\mathrm{S})\,^{30}_{27} \quad \mathrm{FA}\,^{15}_{8} \quad \mathrm{FT}\,^{6}_{2} \quad \mathrm{AT}\,^{10}_{4} \\[4pt]
(\mathrm{SA})\,^{150}_{108} \quad (\mathrm{ST})\,^{60}_{27} \quad \mathrm{FAT}\,^{30}_{8} \\[4pt]
(\mathrm{SAT})\,^{300}_{108}
\end{array}
$$

The test denominators are

Term	F	A	FA	T	FT	AT	FAT
Denominator	S	SA	SA	ST	ST	SAT	SAT

16.7 Crossover Designs

In this section we make a brief return to crossover designs, which in Chapter 12 we described as replicated Latin Squares with blocking on subjects and periods. For concreteness suppose that we have three treatments, three periods, and twelve subjects.

Crossover as Latin Square

The three treatments can be given to the subjects in any of six orders. Assign the orders at random to the subject, two subjects per order, and observe the responses to the treatments in the three periods. From this point of view, the crossover design is a repeated measures design. Order is the grouping factor, *period* is the trial factor, and treatment lies in the order by period interaction. Any carryover effects are also in the order by period interaction. It is customary not to fit the entire order by period interaction, but instead to fit only treatment and carryover effects as needed. With this reduced model, the only difference between the repeated measures and Latin Square approaches to a crossover design is that the Latin Square pools all between subjects variation into a single block term, and the repeated measure splits this into between orders and between subjects within order, allowing the estimation and testing of the overall order effect.

Crossover as repeated measure

Fit order effects

16.8 Further Reading and Extensions

Unbalanced mixed-effects designs are generally difficult to analyze, and split plots are no different. Software that can compute Type I and III mean squares and their expectations for unbalanced data helps find reasonable test statistics. Mathew and Sinha (1992) describe exact and optimal tests for unbalanced split plots.

Nature is not always so kind as to provide us with repeated-measures data that meet the Huynh-Feldt condition (Huynh and Feldt 1970), and as noted above, the Mauchly (1940) test is sensitive to nonnormality. The result of nonconforming correlations is to make the within subjects procedures liberal; that is, confidence intervals are too short and tests reject the null hypothesis more often than they should. This tendency for tests to be liberal can be reduced by modifying the degrees of freedom used when assessing p-values. For example, the within subjects tests for B and AB have $b-1, a(b-1)(n-1)$ and $(a-1)(b-1), a(b-1)(n-1)$ degrees of freedom; these degrees of freedom are adjusted by rescaling to $\lambda(b-1), \lambda a(b-1)(n-1)$ and $\lambda(a-1)(b-1), \lambda a(b-1)(n-1)$, where $1/(b-1) \leq \lambda \leq 1$.

There are two fairly common methods for computing this adjustment λ. The first is from Greenhouse and Geisser (1959); Huynh and Feldt (1976)

provide a slightly less conservative correction. Both adjustments are too tedious for hand computation but are available in many software packages. Greenhouse and Geisser (1959) also provide a simple conservative test that uses the minimum possible value of λ, namely $1/(b-1)$. For this conservative approach, the tests for B and AB have $1, a(n-1)$ and $(a-1), a(n-1)$ degrees of freedom.

16.9   Problems

Briefly describe the experimental design you would choose for each of the following situations, and explain why. Describe treatments, blocks, etc.

Problem 16.1

(a) A substantial fraction of the cholesterol in beef is in the residual blood in the meat. Vascular rinsing attempts to flush out the blood from the circulatory system of the carcass with a weak sugar solution immediately after the animal is slaughtered, thereby lowering the cholesterol. This treatment may change the sensory attributes of the meat, and any change may depend on how long the meat has been aged (4, 6, or 8 days).

We have 10 animals to work with in this experiment, and we will only consider the "freshness" attribute of the sirloin cut. Each animal must be entirely rinsed or not (you cannot rinse part of an animal), but we can produce multiple pieces of sirloin from each animal. We want to study both the effect of rinsing and the effect of the three different aging times.

(b) Underneath the pavement of a road is the subsurface soil or base; the base must be adequately strong or else the road surface will not be durable; it could even be so weak that the construction vehicles get mired—then we get no road at all! When the soil is too weak it must be modified to increase its strength up to something reasonable, say 80 psi. We are in that situation and must modify the soil; the question is which modification to use.

We need to experiment to determine how to set six factors (moisture during mixing, percent Portland cement, percent kiln dust, percent fly ash, compaction delay, drying time) to get the strongest base. Each factor is at two levels. We can get eight truckloads of soil from around the project for experimentation, and each truckload is sufficient to test 16 factor/level combinations.

(c) You work at a consumer agency that is investigating how shoppers get advised when making computer purchases. In particular, you wish to look at how the appearance of the purchaser and the disclosed use for the computer affect the total cost of the recommended package. You choose three appearances (shoppers): a twenty-something female, a twenty-something "geeky" male, and a retired couple. The shoppers will either say that they are going to use their computer for email and web surfing, or for working with multimedia. You expect considerable store to store variation in the kinds of equipment that will be recommended, and you'd

like to study eight different stores (you may send more than one shopper to a store).

(d) We wish to study the tensile strength of a non-woven cotton fabric. The 24 treatments are the factor/level combinations of calendaring temperature (4 levels), binder fibers (2 levels), and binder content (3 levels). We have resources to create and test 48 separate fabrics, but it is expensive to change the calendaring temperature, so we like to avoid changing it often.

(e) A plant breeder wishes to study the effects of soil drainage and variety of tulip bulbs on flower production. Twelve 3 m by 10 m experimental sites are available in a garden. Each site is a .5 m–deep trench. Soil drainage is changed by adding varying amounts of sand to a clay soil (more sand improves drainage), mixing the two well, and placing the mixture in the trench. The bulbs are then planted in the soils, and flower production is measured the following spring. It is felt that four different levels of soil drainage would suffice, and there are fifteen tulip varieties that need to be studied.

(f) It's Girl Scout cookie time, and the Girl Scout leaders want to find out how to sell even more cookies (make more dough?) in the future. The variables they have to work with are type of sales (two levels: door-to-door sales or table sales at grocery stores, malls, etc.) and cookie selection (four levels comprising four different "menus" of cookies offered to customers). Administratively, the Girl Scouts are organized into "councils" consisting of many "troops" of 30-or-so girls each. Each Troop in the experiment will be assigned a menu and a sales type for the year, and for logistical reasons, all the troops in a given council should have the same cookie selection. Sixteen councils have agreed to participate in the experiment.

(g) Rodent activity may be affected by photoperiod patterns. We wish to test this possibility by treating newly-weaned mouse pups with three different treatments. Treatment 1 is a control with the mice getting 14 hours of light and 10 hours of dark per day. Treatment 2 also has 14 hours of light, but the 10 hours of dark are replaced by 10 hours of a low light level. Treatment 3 has 24 hours of full light.

Mice will be housed in individual cages, and motion detectors connected to computers will record activity. We can use 24 cages, but the computer equipment must be shared and is only available to us for 1 month.

Mice should be on a treatment for 3 days—one day to adjust and then 2 days to take measurements. We may use each mouse for more than one treatment, but if we do, there should be 7 days of standard photoperiod between treatments. We expect large subject-to-subject variation. There may or may not be a change in activity as the rat pups age; we don't know.

For each of the following, describe the experimental design used and give **Problem 16.2**

a skeleton ANOVA (sources and degrees of freedom only).

(a) Proteins can be stored frozen for periods of months to years, but they may undergo degradation if subjected to repeated freeze/thaw cycles. This experiment seeks to understand the effects of freezing temperature and repeated freeze/thaw cycles.

On day one, a carton of fresh eggs is purchased at the market, and the albumin is then extracted from the eggs, composited, and homogenized. The albumin is then divided into 20 samples, which are randomly assigned to the combinations of freezing temperature (–20 C or –80 C) and number of freeze/thaw cycles (1 through 10). All subsamples are then frozen at their assigned temperatures. On day two, the samples are all thawed, and the protein concentrations of the samples assigned to one cycle are determined. Then the remaining 18 samples are refrozen. On day three, all samples are thawed, and we determine the protein concentration in the samples assigned to two cycles. This pattern of thawing, measuring, and refreezing is repeated until all samples have been measured (that will be on day 11).

On day 15, we purchase another carton of eggs and repeat the freeze/thaw/measure process for the following 10 days. And, again, on day 29 we purchase a third carton of eggs and repeat the process again.

(b) Brittle bones due to calcium loss are a problem for post-menopausal women. A study in British Columbia examined how one form of exercise for young girls can build their bone mass from ages 10 through 12. Twelve middle schools participated in the study; all are more or less generic suburban schools with no obvious ethnic or socio-economic differences. The schools were randomly assigned to three treatments (four schools per treatment). The treatments are control, make five jumps three times each school day, and jump for 10 minutes three days a week. Bone mass of each girl was measured at age 10 and again at age 12 after two years of their jumping regimen; the response for each girl is the increase in bone mass. In all, there are 1215 girls who complete the two years of the study (approximately 100 per school).

(c) Judges in taste tests use a rinse agent to clear their palates between samples. The rinse agent is supposed to remove the previous sample and restore tastes and odors to a neutral condition. In this experiment we compare sparkling water and still water as rinse agents when having people judge the spiciness of three sauces. Thirty judges are randomly divided into two groups of fifteen. One group will use sparkling water as a rinse and the other will use still water. Each judge will rinse with their type of water, taste a sauce, and then rate the sauce. They then rinse again, taste the second sauce, and so on. Each judge is presented with the sauces in a random order.

(d) An electronics retail chain wishes to compare the effects of print and radio advertising. For their experiment they choose one print ad and one radio ad. They wish to compare sales during weeks when they run neither ad, just the print ad, just the radio ad, and both ads. They run

their experiment in four widely separated cities so that the ads in one city will not be seen or heard by people in another city. They also run their experiment for four consecutive weeks, because you need at least one week to measure sales from each type of advertising. They arrange the experiment so that each ad combination is run once at each city, and once during each week.

(e) Can information affect how we perceive tastes? In this experiment, subjects will take two tastes of beer. One of the products is a premium lager beer, and the other is the same beer with a few drops of balsamic vinegar added. We have 40 male subjects. Each subject will taste both beers in random order and rate the flavor. Twenty of the subjects are chosen at random; these subjects will be told before tasting which beer is standard and which has been spiked. The remaining subjects will be told after tasting but before rating.

(f) I enjoy wine, but not so much that I drink a whole bottle at one meal. It is thus necessary to store an opened bottle for continued enjoyment later. Unfortunately, the wine begins to deteriorate once the bottle is opened. There are several products on the market that claim to retard the deterioration, but are any of them better than simply sticking the cork back in the bottle? We want to test and compare recorking with a vacuum pump and a gas injection method.

We have 12 bottles of wine (a Malbec) from the same case. The bottles are randomly assigned to recorking, vacuum seal, and gas injection, four bottles to each method. On day zero, all 12 bottles are opened, and I taste each wine and give it a flavor score. Then each bottle is resealed using its assigned method. I reopen, sample, rate the flavor, and reseal each bottle (with its assigned method) one day later. I repeat this on days 2, 3, 4, and 5. I thus have six flavor scores from each bottle.

(g) Land use has a strong effect on water quality of streams. To determine the effect of forestry methods on stream nitrate concentration, each of six small watersheds in a forest will be managed using either method A or B. Stream water from each watershed will be analyzed for nitrate concentration weekly for five years, with the average over those five years taken as a response. The six watersheds are adjacent along a ridge line, and there is a slight elevation gradient from the first to the sixth watershed. For this reason, the methods are randomly assigned to the watersheds subject to the restriction that methods A and B occur once each in the three pairs of watersheds (1,2), (3,4), and (5,6).

(h) A grocery store chain is experimenting with its weekly advertising, trying to decide among cents-off coupons, regular merchandise sales, and special-purchase merchandise sales. There are two cities about 100 km apart in which the chain operates, and the chain will always run one advertisement in each city on Wednesday, with the offer good for 1 week. The response of interest is total sales in each city, and large city to city differences in total sales are expected due to population differences. Furthermore, week to week differences are expected. The chain runs the

experiment on 12 consecutive weeks, randomizing the assignment of advertising method to each city, subject to the restrictions that each of the three methods is used eight times, four times in each city, and each of the three pairs of methods is used an equal number of times.

(i) A forest products company conducts a study on twenty sites of 1 hectare each to determine good forestry practice. Their goal is to maximize the production of wood biomass (used for paper) on a given site over 20 years. All sites in the study have been cut recently, and the factors of interest are species to plant (alder or birch) and the thinning regime (thin once at 10 years, or twice at 10 and 15 years). The species is assigned at random to each site. The sites are then split into east-west halves. The thinning regimes are assigned at random to east-west halves independently for each site.

(j) We wish to study the acidity of orange juice available at our grocery store. We choose two national brands. We then choose 3 days at random (from the next month) for each brand; cartons of brand A will be purchased only on the days for brand A, and similarly for brand B. On a purchase day for brand A, we choose five cartons of brand A orange juice at random from the shelf, and similarly for brand B. Each carton is sampled twice and the samples are measured for acidity.

(k) We wish to determine the number of warblers that will respond to three recorded calls. We will get eighteen counts, nine from each of two forest clearings. We expect variation in the counts from early to mid to late morning, and we expect variation in the counts from early to mid to late in the breeding season. Each recorded call is used three times at each clearing, arranged in such a way that each call is used once in each phase of the breeding season and once in each morning hour.

We wish to study the effect of drought stress on height growth of red maple seedlings. The factors of interest are the amount of stress and variety of tree. Stress is at two levels: no stress (that is, always well watered) and drought-stressed after 6 weeks of being well watered. There are four varieties available, and all individuals within a given variety are clones, that is, genetically identical.

Problem 16.3

This will be a greenhouse experiment so that we can control the watering. Plants will be grown in six deep sandboxes. There is space in each sandbox for 36 plants in a 6 by 6 arrangement. However, the plants in the outer row have a dissimilar environment and are used as a "guard row," so responses are observed on only the inner 16 plants (in 4 by 4 arrangement).

The six sandboxes are in a three by two arrangement, with three boxes north to south and two boxes east to west. We anticipate considerable differences in light (and perhaps temperature and other related factors) on the north to south axis. No differences are anticipated on the east to west axis.

Only one watering level can be given to each sandbox. Variety can be varied within sandbox. The response is measured after 6 months.

(a) Describe an experimental design appropriate for this setup.

(b) Give a skeleton ANOVA (sources and df only) for this design.

(c) Suppose now that the heights of the seedlings are measured ten times over the course of the experiment. Describe how your analysis would change and any assumptions that you might need to make.

Problem 16.4

Consider the following experimental design. This design was randomized independently on each of ten fields. First, each field is split into northern and southern halves, and we randomly assign herbicide/no herbicide treatments to the two halves. Next, each field is split into eastern and western halves, and we randomly assign tillage method 1 or tillage method 2 to the two halves. Finally, each tillage half is again split into east and west halves (a quarter of the whole field), and we randomly assign two different insecticides to the two different quarters, independently in the two tillage halves. Thus, within each field we have the following setup:

1	2	3	4
5	6	7	8

Plots 1, 2, 3, and 4 all receive the same herbicide treatment, as do plots 5, 6, 7, and 8. Plots 1, 2, 5, and 6, all receive the same tillage treatment, as do plots 3, 4, 7, and 8. Insecticide A is given to plot pair (1, 5) or plot pair (2, 6); the other pair gets insecticide B. Similarly, one of the plot pairs (3, 7) and (4, 8) gets insecticide A and the other gets B.

Construct a Hasse diagram for this experiment. Indicate how you would test the null hypotheses that the various terms in the model are zero.

Problem 16.5

We plan an experiment to study the effects of eight treatments on the biomass production of a wetland. The eight treatments are the factor-level combinations of A—burning or no burning, B—tillage or no tillage, and C—herbicide or no herbicide. There are 20 square-shaped study sites available. Burning must be done over a large area, so 10 sites are randomly chosen for burning, and the other 10 are left unburned. Each site is divided into two north-south strips; one north-south strip at each site is randomly assigned to receive the herbicide treatment. Each site is also divided into two east-west strips; one east-west strip from each site is randomly assigned to receive the tillage treatment.

Construct a Hasse diagram for this design.

Problem 16.6

Consider the following situation. We have four varieties of wheat to test, and three levels of nitrogen fertilizer to use, for twelve factor-level combinations. We have chosen eight blocks of land at random on an experimental study area; each block of land will be split into twelve plots in a four by three rectangular pattern. We are considering two different experimental designs. In the first design, the twelve factor-level combinations are assigned at random to the twelve plots in each block, and this randomization is redone from block to block. In the second design, a variety of wheat is assigned at random to each row of the four by three pattern, and a level of nitrogen fertilizer is assigned at random to each column of the four by three pattern; this randomization is redone from block to block.

(a) What are the types of the two designs (for example, CRD, RCB, and so on)?

(b) Give Hasse diagrams for these designs, and indicate how you would test the null hypotheses that the various terms in the model are zero.

(c) Which design provides more power for testing main effects? Which design is easier to implement?

Problem 16.7
A food scientist is interested in the production of ice cream. He has two different recipes (A and B). Additional factors that may affect the ice cream are the temperature at which the process is run and the pressure used. We wish to investigate the effects of recipe, temperature, and pressure on ice cream viscosity. The production machinery is available for 8 days, and two batches of ice cream can be made each day. A fresh supply of milk will be used each day, and there is probably some day to day variability in the quality of the milk.

The production machinery is such that temperature and pressure have to be set at the start of each day and cannot be changed during the day. Both temperature and pressure can be set at one of two levels (low and high). Each batch of ice cream will be measured for viscosity.

(a) Describe an appropriate experiment. Give a skeleton ANOVA (source and degrees of freedom only), and describe an appropriate randomization scheme.

(b) Explain how to construct simultaneous 95% confidence intervals for the differences in mean viscosity between the various combinations of temperature and pressure.

Problem 16.8
An experiment was conducted to study the effects of irrigation, crop variety, and aerially sprayed pesticide on grain yield. There were two replicates. Within each replicate, three fields were chosen and randomly assigned to be sprayed with one of the pesticides. Each field was then divided into two east-west strips; one of these strips was chosen at random to be irrigated, and the other was left unirrigated. Each east-west strip was split into north-south plots, and the two varieties were randomly assigned to plots. Data set `IVP`.

| Rep 1 | | | Rep 2 | | | | |
P1	P2	P3	P1	P2	P3	Irrig	Var
53.4	54.3	55.9	46.5	57.2	57.4	yes	1
53.8	56.3	58.6	51.1	56.9	60.2	yes	2
58.2	60.4	62.4	49.2	61.6	57.2	no	1
59.5	64.5	64.5	51.3	66.8	62.7	no	2

What is the design of this experiment? Analyze the data and report your conclusions. What is the standard error of the estimated difference in average yield between pesticide 1 and pesticide 2? irrigation and no irrigation? variety 1 and variety 2?

Problem 16.9
Most universities teach many sections of introductory calculus, and fac-

ulty are constantly looking for a method to evaluate students consistently across sections. Generally, all sections of intro-calculus take the final exam at the same time, so a single exam is used for all sections. An exam service claims that it can supply different exams that consistently evaluate students. Some faculty doubt this claim, in part because they believe that there may be an interaction between the text used and the exam used.

Three math departments (one each at Minnesota, Washington, and UC Berkeley) propose the following experiment. Three random final exams are obtained from the service: E1, E2, and E3. At Minnesota, the three exams will be used in random order in the fall, winter, and spring quarters. Randomization will also be done at Washington and Berkeley. The three schools all use the same two intro calculus texts. Sections of intro calculus at each school will be divided at random into two groups, with half of the sections using text A and the other half using text B. At the end of the year, the mean test scores are tallied with the following results (data set `CalcExams`).

School	Exam	Text A	B
UW	1	81	87
	2	79	85
	3	70	78
UM	1	84	82
	2	81	81
	3	83	84
UCB	1	87	98
	2	82	93
	3	86	90

Analyze these data to determine if there is any evidence of variation between exams, text effect, or exam by text interaction. Be sure to include an explicit description of the model you used.

Problem 16.10 Artificial insemination is widely used in the beef industry, but there are still many questions about how fresh semen should be frozen for later use. The motility of the thawed semen is the usual laboratory measure of semen quality, and this varies from bull to bull and ejaculate to ejaculate even without the freeze/thaw cycle. We wish to evaluate five freeze/thaw methods for their effects on motility.

Four bulls are selected at random from a population of potential donors; three ejaculates are collected from each of the four bulls (these may be considered a random sample). Each ejaculate is split into five parts, with the parts being randomly assigned to the five freeze/thaw methods. After each part is frozen and thawed, two small subsamples are taken and observed under the microscope for motility.

Give a skeleton ANOVA for this design and indicate how you would test the various effects. (Hint: is this a split plot or not?)

Problem 16.11 Traffic engineers are experimenting with two ideas. The first is that erect-

ing signs that say "Accident Reduction Project Area" along freeways will raise awareness and thus reduce accidents. Such signs may have an effect on traffic speed. The second idea is that metering the flow of vehicles onto on-ramps will spread out the entering traffic and lead to an average increase in speed on the freeway. The engineers conduct an experiment to determine how these two ideas affect average traffic speed.

First, twenty more-or-less equivalent freeway interchanges are chosen, spread well around a single metropolitan area and not too close to each other. Ten of these interchanges are chosen at random to get "Accident Reduction Project Area" signs (in both directions); the other ten receive no signs. Traffic lights are installed on all on-ramps to meter traffic. The traffic lights can be turned off (that is, no minimum spacing between entering vehicles) or be adjusted to require 3 or 6 seconds between entering vehicles. Average traffic speed 6:30–8:30 A.M. and 4:30–6:30 P.M. will be measured at each interchange on three consecutive Tuesdays, with our response being the average of morning and evening speeds. At each interchange, the three settings of the traffic lights are assigned at random to the three Tuesdays.

The results of the experiment follow (data set `Interchanges`). Analyze the results and report your conclusions.

| | | Timing | | |
Interchange	Sign	0	3	6
1	n	13	25	26
2	n	24	35	37
3	n	22	38	41
4	n	24	32	37
5	n	23	35	38
6	n	23	33	35
7	n	24	35	41
8	n	19	34	35
9	n	21	33	37
10	n	15	30	30
11	y	19	31	33
12	y	12	28	27
13	y	10	24	29
14	y	12	23	28
15	y	26	41	41
16	y	17	31	30
17	y	17	27	31
18	y	18	32	33
19	y	16	29	30
20	y	24	37	37

Problem 16.12

A consumer testing agency wishes to test the ability of laundry detergents, bleaches, and prewash treatments to remove soils and stains from fabric. Three detergents are selected (a liquid, an all-temperature powder, and a hot-water powder). The two bleach treatments are no bleach or chlorine bleach. The three prewash treatments are none, brand A, and brand B. The

three stain treatments are mud, grass, and gravy. There are thus 54 factor-level combinations.

Each of 108 white-cotton handkerchiefs is numbered with a random code. Nine are selected at random, and these nine are assigned at random to the nine factor-level combinations of stain and prewash. These nine handkerchiefs along with four single sheets make a "tub" of wash. This is repeated twelve times to get twelve tubs. Each tub of wash is assigned at random to one of the six factor-level combinations of detergent and bleach. After washing and drying, the handkerchiefs are graded (in random order) for whiteness by a single evaluator using a 1 to 100 scale, with 1 being whitest (cleanest).

Analyze these data (data set `Handkerchiefs`) and report your findings.

| | | | Stain 1 | | | Stain 2 | | | Stain 3 | | |
Tub	Det.	Bl.	P1	P2	P3	P1	P2	P3	P1	P2	P3
1	1	1	1	3	3	3	3	5	10	3	2
2	1	2	5	3	3	3	5	3	7	3	2
3	2	1	3	2	2	4	6	1	5	1	2
4	2	2	3	1	2	2	4	3	8	1	2
5	3	1	34	29	35	35	34	41	49	25	26
6	3	2	7	5	6	6	6	7	10	5	4
7	1	1	4	4	4	5	7	10	11	5	4
8	1	2	4	6	3	4	7	6	9	7	5
9	2	1	6	8	7	5	6	7	11	6	4
10	2	2	6	6	7	8	7	9	12	5	5
11	3	1	26	28	31	38	30	34	41	27	27
12	3	2	2	4	2	2	5	3	8	3	2

Yellow perch and ruffe are two fish species that compete. An experiment is run to determine the effects of fish density and competition with ruffe on the weight change in yellow perch. There are two levels of fish density (low and high) and two levels of competition (ruffe absent and ruffe present). Sixteen tanks are arranged in four enclosures of four tanks each. Within each enclosure, the four tanks are randomly assigned to the four factor-level combinations of density and competition. The response is the change in the weight of perch after 5 weeks (in grams, data from Julia Frost, data set `Ruffe`).

Problem 16.13

| | | Enclosure | | | |
Ruffe	Density	1	2	3	4
Absent	Low	.0	.4	.9	-.4
	High	.9	-.4	-.6	-1.2
Present	Low	.0	-.4	-.9	-.9
	High	-1.2	-1.5	-1.1	-.7

Analyze these data for the effects of density and competition.

Chapter 17

Designs with Covariates

Covariates are predictive responses, meaning that covariates are responses measured for an experimental unit in anticipation that the covariates will be associated with, and thus predictors for, the primary response. The use of covariates is not design in the sense of treatment structure, unit structure, or the way treatments are assigned to units. Instead, a covariate is an additional response that we exploit by modifying our models to include. Nearly any model can be modified to include covariates.

Covariates are predictive responses

Example 17.1 Keyboarding pain

A company wishes to choose an ergonomic keyboard for its computers to reduce the severity of repetitive motion disorders (RMD) among its staff. Twelve staff known to have mild RMD problems are randomly assigned to three keyboard types. The staff keep daily logs of the amount of time spent keyboarding and their subjective assessment of the RMD pain. After 2 weeks, we get the total number of hours spent keyboarding and the total number of hours in RMD pain.

The primary response here is pain; we wish to choose a keyboard that reduces the pain. However, we know that the amount of pain depends on the amount of time spent keyboarding—more keyboarding usually leads to more pain. If we knew at the outset the amount of keyboarding to be done, we could block on time spent keyboarding. However, we don't know that at the outset of the experiment, we can only measure it along with the primary response. Keyboarding time is a covariate.

17.1 The Basic Covariate Model

Before we show how to use covariates, let's describe what they can do for us. First, we can make comparisons between treatments more precise by including covariates in our model. Thus we get a form of variance reduction through modeling the response-covariate relationship, rather than through

Covariates make treatment comparisons more precise

blocking. The responses we observe are just as variable as without covariates, but we can account for some of that variability using covariates in our model and obtain many of the benefits of variance reduction via modeling instead of blocking.

Second—and this is not completely separate from the first advantage—covariate models allow us to compare predicted treatment responses at a common value of the covariate for all treatments. Thus treatments which by chance received above or below average covariate values can be compared in the center.

<div style="float:right; text-align:right; width:25%;">Treatment comparisons adjusted to common covariate value</div>

One potential pitfall of covariate models is that they assume that the covariate is not affected by the treatment. When treatments affect covariates, the comparison of responses at equal covariate values (our second advantage) may, in fact, obscure treatment differences. For example, one of the keyboards may be so awkward that the users avoid typing; trying to compare it to the others at an average amount of typing hides part of the effect of the keyboard.

<div style="float:right; text-align:right; width:25%;">Treatments should not affect covariates</div>

The key to using covariates is building a model that is appropriate for the design and the data. Covariate models have two parts: a usual treatment effect part and a covariate effect part. The treatment effect part is essentially determined by the design, as usual; but there are several possibilities for the covariate effect part, and our model will be appropriate for the data only when we have accurately modeled the relationship between the covariates and the response.

<div style="float:right; text-align:right; width:25%;">Treatment and covariate effects</div>

Let's begin with the simplest sort of covariance modeling—in fact, the sort usually called *Analysis of Covariance*. We will generalize to more complicated models later. Consider a completely randomized design with a single covariate x; let x_{ij} be the covariate for y_{ij}. For the CRD, the model ignoring the covariate is

<div style="float:right; text-align:right; width:25%;">Analysis of covariance</div>

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \ .$$

We can estimate the ith treatment mean $\widehat{\mu} + \widehat{\alpha}_i$ or a contrast between treatments $\sum w_i \widehat{\alpha}_i$, and we can test the null hypothesis that all the α_i values are zero with the usual F-test by comparing the fit of this model to the fit of a model without the α_i's.

Now consider a model that uses the covariate. We augment the previous model to include a regression-like term for the covariate:

<div style="float:right; text-align:right; width:25%;">Include covariate via regression</div>

$$y_{ij} = \mu^\star + \alpha_i^\star + \beta + \epsilon_{ij}^\star \ .$$

As usual, the treatment effects $\alpha_i^\star$ add to zero. The $\star$'s in this model are shown just this once to indicate that the μ, α_i, and ϵ_{ij} values in this model are different from those in the model without covariates. The $\star$'s will be dropped now for ease of notation.

The difference between the covariate and no-covariate models is the term βx_{ij}. This term models the response as a linear function of the covariate x. The assumption of a linear relationship between x and y is a big one, and writing a model with a linear relationship doesn't make the actual relationship linear. As with any regression, we may need to transform the x or y to

<div style="float:right; text-align:right; width:25%;">Model assumes linear relationship between response and covariate</div>

improve linearity. Plots of the response versus the covariate are essential for assessing this relationship.

Also note that the slope β is assumed to be the same for every treatment. The covariate model for treatment i is a linear regression with slope β and intercept $\mu + \alpha_i$. Because the α_i's can all differ, this is a set of parallel lines, one for each treatment. Thus this covariate model is called the *parallel-lines* model or the *separate-intercepts* model.

<div style="text-align: right">Common slope creates parallel lines</div>

We need to be able to test the same hypotheses and estimate the same quantities as in noncovariate models. To test the null hypothesis of no treatment effects (all the α_i's equal to zero) when covariate effects are present, compare the model with treatment and covariate effects to the reduced model with only covariate effects:

<div style="text-align: right">Test via model comparison</div>

$$y_{ij} = \mu + \beta x_{ij} + \epsilon_{ij} \ .$$

This simpler model is called the *single-line* model, because it is a simple linear regression of the response on the covariate. The reduction in error sum of squares going from the single-line model to the parallel-lines model has $g - 1$ degrees of freedom. The mean square for this reduction is divided by the mean square for error from the larger parallel-lines model to form an F-test of the null hypothesis of no treatment effects. These treatment effects are said to be covariate-adjusted, because the covariate is present in the model. There are formulae for these sums of squares, but I don't think you'll find them enlightening; just let your software do the computations.

<div style="text-align: right">Single-line model

F-test for covariate-adjusted treatment effects</div>

The underlying philosophy of the test is that the covariate relationship with the response is real and exists with or without treatment effects. The test is only to determine if adding treatment effects to a model that already includes a covariate makes any significant improvement in explanatory power. That is, does the parallel-lines model explain significantly more than the single-line model? This test is the classical Analysis of Covariance.

<div style="text-align: right">Analysis of Covariance</div>

Computer software can supply estimates of the effects in our models. The estimated treatment effects $\widehat{\alpha}_i$ describe how far apart the parallel lines are, $\widehat{\mu}$ gives an average intercept, $\widehat{\mu} + \widehat{\alpha}_i$ gives the intercept for treatment i, and $\widehat{\beta}$ is the estimated slope.

How should we answer the question, "What is the mean response in treatment i?" This is a little tricky, because the response depends on the covariate. We need to choose some standard covariate value $\dot{x}$ and evaluate the treatment means there.

<div style="text-align: right">Means depend on covariate</div>

Covariate-adjusted means are the estimated values in each treatment group when the covariate is set to $\overline{x}_{\bullet\bullet}$, the grand mean of the covariates, or

<div style="text-align: right">Covariate adjusted means at grand mean of covariate</div>

$$\widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}\overline{x}_{\bullet\bullet} \ .$$

Covariate-adjusted means give us a common basis for comparison, because all treatments are evaluated at the same covariate level. Note that the difference between two covariate-adjusted means is just the difference between the treatment effects; we would get the same differences if we compare the means at the common covariate value $\dot{x} = 0$.

Table 17.1: Hours keyboarding (x) and hours of repetitive-motion pain (y) during 2 weeks for three styles of keyboards.

1		2		3	
x	y	x	y	x	y
60	85	54	41	56	41
72	95	68	74	56	34
61	69	66	71	55	50
50	58	59	52	51	40
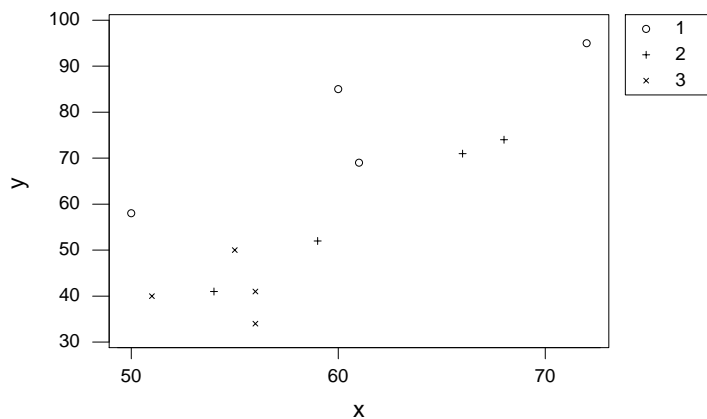
Figure 17.1: Hours of pain versus hours of keyboarding for twelve subjects and three keyboard types, using Minitab.

Example 17.2 Keyboarding pain, continued

Table 17.1 shows hours of keyboarding and hours of pain for the twelve subjects (data set `KeyboardingPain`), and Figure 17.1 shows a plot of the response versus the covariate, with keyboard type indicated by the plotting symbol. The plot clearly shows a strong, reasonably linear relationship between the response and the covariate. The figure also shows that the keyboard 1 responses tend to be above the keyboard 2 responses for similar covariate values, and keyboard 2 and 3 responses are somewhat mixed at the low end of the covariate. We can further see that keyboard 3 covariates tend to be a bit smaller than the other two keyboards, so presumably at least some of the explanation for the low responses for keyboard 3 is the low covariate values.

Minitab output analyzing these data follows. We first check to see if

treatments affect the covariate keyboarding time. The ANOVA ① provides no evidence against the null hypothesis that the treatments have the same average covariate values (p-value .29). In these data, keyboard 3 averages about 6 to 7 hours less than the other two keyboards ②, but the difference is within sampling variability.

Next we do the Analysis of Covariance ③. The model includes the covariate and then the treatment. Minitab produces both sequential and Type III sums of squares; in either case, the sum of squares for treatments is treatments adjusted for covariates, which is what we need. The p-value is .004, indicating strong evidence against the null hypothesis of no treatment effects.

The covariate-adjusted means and their standard errors are given at ⑤. Note that the standard errors are not all equal. We can also construct the covariate adjusted means from the effects ④. For example, the covariate-adjusted mean for keyboard 1 is

$$-48.21 + 14.399 + 1.8199 \times 59 = 73.57 \ .$$

```
Analysis of Variance for x

Source      DF        SS        MS        F      P
type         2    123.50     61.75     1.45  0.286        ①
Error        9    384.50     42.72

Means

type    N         x                                       ②
1       4    60.750
2       4    61.750
3       4    54.500

Analysis of Variance for y, using Adjusted SS for Tests

Source      DF    Seq SS    Adj SS    Adj MS       F      P
x            1    2598.8    1273.5    1273.5   24.79  0.001   ③
type         2    1195.8    1195.8     597.9   11.64  0.004
Error        8     411.0     411.0      51.4

Term         Coef    StDev        T      P
Constant   -48.21    21.67    -2.22  0.057              ④
x          1.8199   0.3655     4.98  0.001
type
1          14.399    2.995     4.81  0.001
2          -4.671    3.094    -1.51  0.170

Means for Covariates

Covariate      Mean    StDev
x             59.00    6.796

Least Squares Means for y

type      Mean     StDev                                 ⑤
1        73.57     3.641
2        54.50     3.722
3        49.44     3.943

Tukey 95.0% Simultaneous Confidence Intervals           ⑥
Response Variable y
All Pairwise Comparisons among Levels of type

type = 1 subtracted from:

type    Lower    Center    Upper  -------+---------+---------+------+---
2      -33.59   -19.07   -4.553        (--------*---------)
3      -40.01   -24.13   -8.244  (----------*----------)
                                 -------+---------+---------+------+---
                                     -30       -15        0
type = 2 subtracted from:

type    Lower    Center    Upper  -------+---------+---------+------+---
3      -21.39   -5.056    11.28               (----------*---------)
                                 -------+---------+---------+------+---
                                     -30       -15        0
```

It appears that keyboards 2 and 3 are about the same, and keyboard 1 is worse (leads to a greater response). This is confirmed by doing a pairwise

comparison of the three treatment effects using Tukey HSD ⑥.

We conclude that there are differences between the three keyboards, with keyboard 1 leading to about 21 more hours of pain in the 2-week period for an average number of hours keyboarding. The coefficient of keyboard hours was estimated to be 1.82, so an additional hour of keyboarding is associated with about 1.82 hours of additional pain.

Before leaving the example, a few observations are in order. First, the linear model is only reliable for the range of data over which it was fit. In these data, the hours of keyboarding ranged from about 50 to 70, so it makes no sense to think that doing no keyboarding with keyboard 1 will lead to -34 hours of pain (34 hours of pleasure?).

Next, it is instructive to compare the results of this Analysis of Covariance with those that would be obtained if the covariate had been ignored. You would not ordinarily do this as part of your analysis, but it helps us see what the covariate has done for us.

```
nalysis of Variance for y, using Adjusted SS for Tests

ource     DF      Seq SS      Adj SS      Adj MS      F      P
ype        2      2521.2      2521.2      1260.6    6.74   0.016    ⑦
rror       9      1684.5      1684.5       187.2

erm            Coef      StDev        T       P
onstant      59.167      3.949    14.98   0.000              ⑧
ype
             17.583      5.585     3.15   0.012
              0.333      5.585     0.06   0.954

east Squares Means for y
ype      Mean      StDev                                     ⑨
         76.75      6.840
         59.50      6.840
         41.25      6.840
```

Two things are noteworthy. First, the error mean square for the analysis without the covariate ⑦ is about 3.6 times larger than that with the covariate. Regression on the covariate has explained much of the variation within treatment groups, so that residual variation is reduced. Second, the covariate-adjusted treatment effects ④ are not the same as the unadjusted treatment effects ⑧; likewise, the covariate-adjusted means 73.565, 54.495, and 49.44 ⑤ differ from the raw treatment means 76.75, 59.5, and 41.25 ⑨. This shows the effect of comparing the treatments at a common value of the covariate. For these data, the covariate-adjusted means are more tightly clustered than the raw means; other data sets may show other patterns.

Some authors prefer to write the covariate model

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij}$$

in the slightly different form

$$y_{ij} = \tilde{\mu} + \alpha_i + \beta(x_{ij} - \overline{x}_{\bullet\bullet}) + \epsilon_{ij} \ .$$

The difference is that the covariate x is centered to have mean zero, so that the covariate-adjusted means in the revised model are just $\tilde{\mu} + \alpha_i$. We can see that there is no essential difference between these two models once we realize that $\tilde{\mu} = \mu + \beta \overline{x}_{\bullet\bullet}$.

Centered covariates

17.2 When Treatments Change Covariates

The usual Analysis of Covariance assumes that treatments do not affect the covariates. When this is true, it makes sense to compare treatments via covariate-adjusted means—that is, to compare treatments at a common value of the covariate—because any differences between covariates are just random variation. When treatments do affect covariates, differences between covariates are partly treatment effect and partly random variation. Forcing treatment comparisons to be at a common value of the covariate obscures the true treatment differences.

Covariate adjustment can obscure the treatment effect

We can make this more precise by reexpressing the covariate in our model. Expand the covariate into a grand mean, deviations of treatment means from the grand mean, and deviations from treatment means to obtain $x_{ij} = \overline{x}_{\bullet\bullet} + (\overline{x}_{i\bullet} - \overline{x}_{\bullet\bullet}) + (x_{ij} - \overline{x}_{i\bullet})$, and substitute it into the model:

$$
\begin{aligned}
y_{ij} &= \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij} \\
&= \mu + \alpha_i + \beta(\overline{x}_{\bullet\bullet} + (\overline{x}_{i\bullet} - \overline{x}_{\bullet\bullet}) + (x_{ij} - \overline{x}_{i\bullet})) + \epsilon_{ij} \\
&= (\mu + \beta\overline{x}_{\bullet\bullet}) + (\alpha_i + \beta(\overline{x}_{i\bullet} - \overline{x}_{\bullet\bullet})) + \beta(x_{ij} - \overline{x}_{i\bullet}) + \epsilon_{ij} \\
&= \tilde{\mu} \qquad\quad + \tilde{\alpha}_i \qquad\qquad\qquad\quad + \beta\tilde{x}_{ij} \qquad\quad + \epsilon_{ij}
\end{aligned}
$$

We have seen that covariate-adjusted treatment effects may not equal covariate-unadjusted treatment effects. In the preceding equations, α_i is the covariate-adjusted treatment effect, and $\tilde{\alpha}_i$ is the unadjusted effect (see Question 17.1). These differ by $\beta(\overline{x}_{i\bullet} - \overline{x}_{\bullet\bullet})$, so adjusted and unadjusted effects are the same if all treatments have the same average covariate. If the treatments are affecting the covariate, these adjustments should not be made.

Covariate adjustment to means is $\beta(\overline{x}_{i\bullet} - \overline{x}_{\bullet\bullet})$

We can obtain the variance reduction property of covariance analysis without also doing covariate adjustment by using the covariate $\tilde{x}$ instead of x. Compute $\tilde{x}$ by treating the covariate x as a response with the treatments as explanatory variables; the residuals from this model are $\tilde{x}$.

Using $\tilde{x}$ gives variance reduction only

Note that the two analyses described here are extremes: ordinary analysis of covariance assumes that treatments cause no variation in the covariate, and the analysis with the altered covariate $\tilde{x}$ assumes that all between treatment variation in the covariates is due to treatment.

Example 17.3 Keyboarding pain, continued

An analysis of variance on the keyboarding times in Table 17.1 showed no evidence that the different keyboards affected keyboarding times. Nonetheless, we use those data here to illustrate the analysis that uses covariates only for variance reduction, and not for covariate adjustment.

The first step is to get the modified covariate as the residuals from a model with treatments and the covariate as the response. The ANOVA for this model is at ① of Example 17.2; the residuals have been saved as $\tilde{x}$, which we next use in a standard Analysis of Covariance.

Here is Minitab output using this modified covariate.

```
nalysis of Variance for y, using Adjusted SS for Tests

ource     DF      Seq SS      Adj SS      Adj MS       F       P
tilde      1      1273.5      1273.5      1273.5    24.79   0.001
ype        2      2521.2      2521.2      1260.6    24.54   0.000
rror       8       411.0       411.0        51.4

east Squares Means for y

ype       Mean       StDev
         76.75       3.584
         59.50       3.584
         41.25       3.584
```

We can see in the ANOVA table that the error mean square is the same in this analysis as it was in the standard Analysis of Covariance in Example 17.2 ③. The mean square for treatments adjusted for this modified covariate is the same as the mean square for treatments alone; in fact, we constructed the modified covariate to make this so. For these data, the treatment mean square adjusted for the modified covariate (same as the unadjusted treatment mean square) is over twice the size of the treatments adjusted for covariate mean square; the p-value in the modified analysis is thus much smaller.

Finally, we see that the covariate-adjusted treatment means using the modified covariate are the same as the simple treatment means in Example 17.2 ⑨. The standard errors for these adjusted means are much smaller than the standard errors for the unadjusted means, however, because the modified covariate accounts for a large amount of response variation within each treatment group. Also, the standard errors for the covariate-adjusted means using $\tilde{x}$ are equal, unlike those using x.

The covariate-adjusted treatment effects can be larger or smaller than the unadjusted effects (depending on the sign of β and the pattern of covariates). Similarly, the covariate-adjusted effects may have a larger or smaller p-value than the treatment effects in a model with the modified covariate. We must not choose between the original and modified covariates based on the results of the analysis; we must choose based on whether we wish to ascribe covariate differences to treatments.

17.3   Other Covariate Models

We have been discussing the simplest possible covariate model: a single covariate with the same slope in all treatment groups. It is certainly possible to have two or more covariates. The standard analysis is still treatments adjusted for covariates, and covariate-adjusted means are evaluated with each covariate at its overall average. If one or more covariates are affected by treatments

More than one
covariate

Constant Mean
|
Single Line

Separate Intercepts   Separate Slopes

Separate Lines
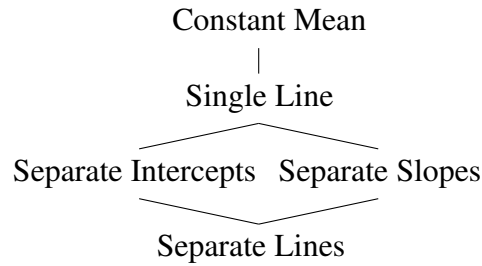
Figure 17.2: Lattice of covariate models.

and we wish to identify the variation associated with treatment differences in those covariates as treatment variation, then each of those covariates should be individually modified as described in the preceding section.

Covariates can also be used in other designs beyond the CRD with a single treatment factor. Blocking designs and fixed-effects factorials can easily accommodate covariates; simply look at treatments adjusted for any blocks and covariates. Note that treatment factors adjusted for covariates will not usually be orthogonal, even for balanced designs, so you will need to do Type II or Type III analyses for factorials.

Covariates with blocks or factorials

Our covariate models have assumed that treatments affect the response by an additive constant that is the same for all values of the covariate. This is the parallel-lines model, and it is the standard model for covariates. It is by no means the only possibility for treatment effects. For example, treatments could change the slope of the response-covariate relationship, or treatments could change both the slope and the intercept.

Treatments could change the covariate slope

We can put covariate models into an overall framework as shown in Figure 17.2. Models are simplest on top and add complexity as you move down an edge. Any two models that can be connected by going down one or more edges can be compared using an Analysis of Variance. The lower model is the full model and the upper model is the reduced model, and the change in error sum of squares between the two models is the sum of squares used to compare the two models. The degrees of freedom for any model comparison is the number of additional parameters that must be fit for the larger model.

Lattice of covariate models

The top model is a constant mean; this is a model with no treatment effects and no covariate effect. We only use this model if we are interested in determining whether there is any covariate effect at all (by comparing it to the single-line model). The single line model is the model where the covariate affects the response, but there are no treatment effects. This model has one more parameter than the constant mean model, so there is 1 degree of freedom in the comparison of the constant-mean and single-line models (and that degree of freedom is the slope parameter).

Constant mean

Single line

Moving down the figure, we have two choices. On the left is the separate-intercepts model. This is the model with a common covariate slope and a different intercept for each treatment. The comparison between the single-line model and the separate-intercepts model is the standard Analysis of Covari-

Separate intercepts

ance, and it has $g - 1$ degrees of freedom for the $g - 1$ additional intercepts that must be fit.

If instead we move down to the right, we get the separate-slopes model:

$$y_{ij} = \mu + \beta_i(x_{ij} - x_0$$

In this model, the relationship between response and covariate has a different slope β_i for each treatment, but all the lines intersect at the covariate value x_0. If you set $x_0 = 0$, then all the lines have the same intercept. Different values of x_0 are like different covariates. This model has $g - 1$ more degrees of freedom than the single-line model.

Separate slopes

At the bottom, we have the separate-lines model:

$$y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \epsilon_{ij}$$

This model has $g - 1$ more degrees of freedom than either the separate-intercepts or separate-slopes models. If we move down the left side of the figure, we add intercepts then slopes, while moving down the right side we add the slopes first, then the intercepts.

Separate lines

Example 17.4 Keyboarding pain, continued

Let's fit the full lattice of covariate models to the keyboarding pain data. Here is MacAnova output for these models; all sums of squares are sequential.

```
odel used is y=x+type+x.type
                DF            SS           MS            F       P-value
                 1        2598.8       2598.8     53.62884   0.00033117   ①
ype              2        1195.8       597.91     12.33835    0.0074822
.type            2        120.27       60.136      1.24095      0.35398
RROR1            6        290.76       48.459

odel used is y=x+x.type+type
                DF            SS           MS            F       P-value
                 1        2598.8       2598.8     57.62884   0.00033117   ②
.type            2        1168.4       584.22     12.05596    0.0079111
ype              2        147.65       73.826      1.52345      0.29171
RROR1            6        290.76       48.459

odel used is y=x59+x59.type
                DF            SS           MS            F       P-value
59               1        2598.8       2598.8     14.66486    0.0050217   ③
59.type          2        189.13       94.566      0.53363      0.60598
RROR1            8        1417.7       177.21
```

ANOVA ① descends the left-hand side of the lattice, starting with the covariate x (time), adding keyboard type adjusted for covariate (separate intercepts), and finally adding separate slopes to get separate lines. The type mean square of 597.91 is the usual Analysis of Covariance mean square. ANOVA ② descends the right-hand side of the lattice, starting with the covariate x, adding separate slopes, and finally adding separate intercepts to get separate lines. Adding separate slopes makes a significant improvement over a single

(a)   (b)

(c)   (d)
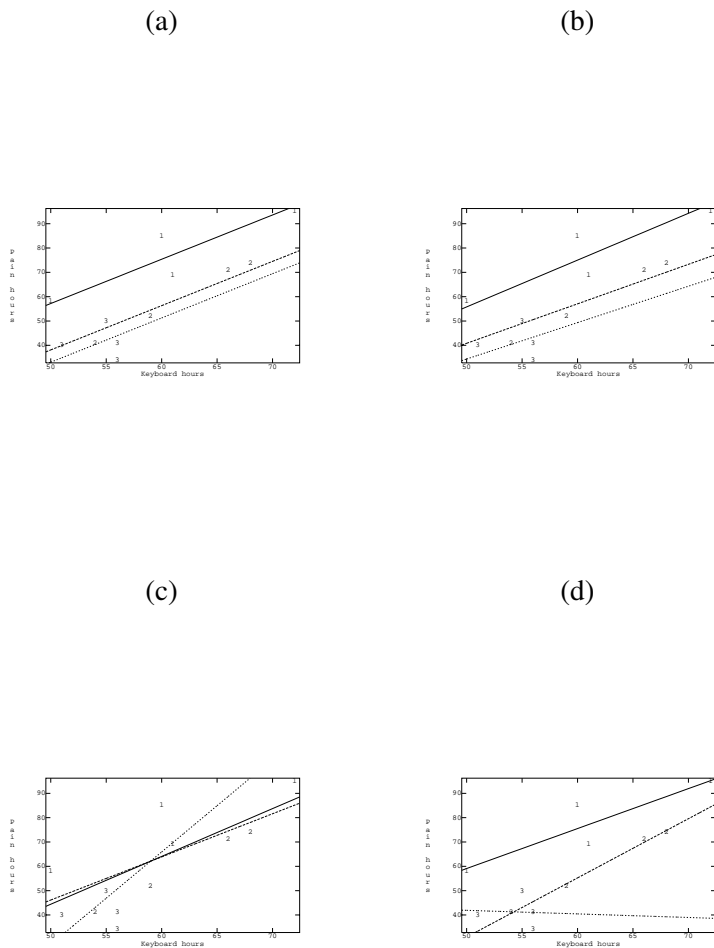
Figure 17.3: Covariate model fits for the keyboarding pain data, using MacAnova: (a) separate intercepts, (b) separate slopes $x_0 = 0$, (c) separate slopes $x_0 = 59$, (d) separate lines.

line (p-value of .0079), but adding separate lines is not a significant improvement over separate slopes. The separate slopes model ② uses $x_0 = 0$, so the fitted lines intersect at 0. ANOVA ③ fits a separate slopes model with $x_0 = 59$. In this case, there is no significant improvement going to separate slopes. Figure 17.3 shows the fits for four models.

The single-line and separate-intercepts models are the most commonly used models of this family. They are analogues of treatment models with blocking. However, not all experimental data will fit nicely into this view of the world, and we need to be ready to consider the less common covariate models if the data require it.

17.4   Further Reading and Extensions

Federer and Meredith (1992) discuss the use of covariates in split-plot and split-block designs. Consider two situations. First, all split plots in a whole plot have the same covariate, so that the covariate only depends on the whole plot. In this case, covariate is a whole-plot effect, and its 1 degree of freedom and sum of squares are computed at the whole-plot level.

Second, consider when each split plot has its own covariate value x_{ijk}. Construct two new covariates from x. The first is a covariate at the whole-plot level formed by taking the average covariate for each whole plot: $\overline{x}_{i\bullet k}$. This covariate acts at the whole-plot level, and its 1 degree of freedom and sum of squares are computed at the whole-plot level. The second is a split-plot covariate: $\tilde{x}_{ijk} = x_{ijk} - \overline{x}_{i\bullet k}$. This split-plot covariate is the deviation of the original covariate x from the whole-plot average value for x. The 1 degree of freedom and sum of squares for this covariate are at the split-plot level. Note that there may be different coefficients (slopes) for the covariates at the whole- and split-plot levels.

Analysis of Covariance for general random- and mixed-effects models is considerably more difficult. Henderson and Henderson (1979) and Henderson (1982) discuss the problems and possible approaches. In fact, the whole September 1982 issue of *Biometrics* that includes Henderson (1982) is devoted to Analysis of Covariance.

17.5   Problems

Exercise 17.1   What is the difference in randomization between a completely randomized design in which a covariate is measured and a completely randomized design in which no covariate is measured?

Exercise 17.2   Briefly discuss the difference in *design* between a randomized complete block design with four treatments and five blocks, and a two-way factorial design with factor A having four levels and factor B having five levels.

Problem 17.1   Briefly describe the experimental design you would choose for each of the following situations, and why. Describe treatments, blocks, etc.

(a) Scientists are studying the use of caffeine sprays (2% caffeine, that will keep you awake!) to kill nonnative frogs in Hawaii. Specifically, we want to compare the populations of frogs after applying the caffeine sprays at three rates (none, a little, or a lot); the frog populations are measured by trapping frogs three days after the spray is applied. We have thirty test

areas that we can use, and no information to group them into homogeneous subgroups. At the same time that we trap frogs, we can also trap insects, because we expect higher frog populations in areas with higher insect densities.

(b) Genetically modified bacteria may be able to produce insulin as they grow and divide. We wish to conduct a small scale experiment to study the effects of growth temperature and culture medium on the production of insulin. The bacteria will grow in a nutrient broth (the medium) in a large beaker; temperature is controlled by putting the beaker in an environmental chamber that maintains a uniform temperature. The environmental chamber is large enough for four beakers. There are three temperatures and four media we wish to study. We have resources to run 48 beakers.

(c) I take horseback riding lessons, and sometimes we will do a "barrel race" for fun. I'm not very good at this, but I do like to win, so I would like to run a little experiment to get my best time. I need to look at the following factors: (1) tight turns (slower but less distance) versus wide turns (faster but longer distance), (2) left handed versus right handed circuit (just mirror images of the pattern), (3) other horses (present or absent).

Now running a pattern at full speed takes a lot out of me and the horse, so we can't do more than two runs during a lesson. Also, I would like to finish my experiment during one "session" of four lessons.

(d) Reactions between water, steel, and chemicals in concrete can degrade the strength of steel reinforced concrete. We wish to compare ordinary reinforcing steel with steel coated with one of three sealants. We have money to make 24 concrete pillars. Each pillar consists of concrete and one of the four versions of reinforcing steel. The finished pillars will be immersed in a brine solution for 90 days, and then tested for strength. We anticipate some batch to batch differences in the concrete we mix, and we can only mix enough concrete for two pillars at one time.

(e) We wish to make experimental observations on the stretching of a fibrous tissue subject to forces in both the longitudinal (along the axis of the fibers) and transverse (across the axis of the fibers) directions. We have four different stretching protocols we want to compare but only eight samples of tissue. The bad news is that we expect substantial variability from sample to sample, but the good news is that we can use each sample more than once. The not quite so good news is that the tissue does gradually distort as we stretch it, so we would expect the results from a first stretch to be different from a second stretch, the second to be different from the third, and so on. It is not realistic to stretch the tissue more than four times, because the distortion becomes too great after that many stretches.

(f) When spring is in the air, a young woman's fancy turns to gardening. We have 16 small flowering plants (all the same variety), and we would like to experiment this year to determine how best to grow this variety in the future. We wish to consider four factors, each at two levels. The factors

are soil type (regular or Miracle Gro^{TM}), pot size (4 inch or 7 inch), buds (pinch off or leave in place), and water (plain or with added fertilizer). We will grow the small plants indoors for three weeks before setting them in the garden outside, and we'll measure the size of the plants after the three week period as our response. While the plants are inside, they will be placed on two shelves, one by an east facing window and one by a north facing window. There will be eight plants per shelf.

(g) The Department of Natural Resources stocks many lakes with walleye for sport fisheries. The fish are grown from eggs in artificial pools at state hatcheries until they are large enough to release into the lakes. As the fish grow, they are moved from pool to pool at the hatchery, generally swimming in four pools before their release; the increase in total fish weight is measured when the fish are removed from each pool (a whole batch of fish, not the individual fish). An important issue for the department is optimizing the efficiency of the food mixture; that is, the department wishes to achieve the maximum growth per dollar spent. In this experiment, the department is considering two different brands of food, and two different feeding rates (kg of food per kg of fish per day). There are eight hatcheries raising walleye, and each hatchery will participate with one batch of walleye from hatch until release. We anticipate some variation from hatchery to hatchery due to weather affecting growth.

(h) An experimental forest has six small watersheds that are available for manipulation. The overall experimental forest is much, much larger than these six watersheds, it's just that these six are the only areas suitable for this experiment. The watersheds are fairly homogeneous in size and other characteristics, but they are far enough apart that local meteorology can differ from watershed to watershed. As part of the experiment, a weather station is set up at each watershed providing daily temperature, precipitation, average wind speed, principal wind direction, and insolation (amount of sunshine).

There are three treatments of interest: clear cutting the trees, selective cutting the trees (less severe than clear cutting), and no cutting (control). Scientists are going to measure many, many response variables, but one of the most important is the total output of stream water from each watershed during the summer growing season after the cuts (which are done in the winter). What kind of experimental design or technique would you use?

(i) Neutral third parties assign punishment to convicted criminals, but does the brain of the judge operate the same way regardless of the facts of the case? We wish to compare the effects of the type of crime (rape, petty theft, or no crime) and any mitigating circumstances (none, duress, or mental illness). Volunteer judges will lie in the fMRI scanner while another volunteer reads the facts of the case. We believe that each judge should consider a single case (type of crime), but he or she can consider punishments for that crime based on all three types of mitigating circumstances. Our response is the activity in the right dorsolateral prefrontal cortex.

(j) We wish to determine the amount of salt to put in a microwave popcorn so that it has the best overall acceptability. We will test three levels of salt: low, medium, and high. We have recruited 25 volunteers to taste popcorn, and while we expect the individuals to be reasonably consistent in their own personal ratings, we expect large volunteer to volunteer differences in overall ratings.

(k) Some brands of golf balls claim to fly farther. To test this claim, you devise a mechanical golf ball whacker which will strike the golf balls with the same power and stroke time after time. Ten balls of each of six brands will be struck once by the device and measured for distance traveled. Wind speed, which will affect the distance traveled, is variable and unpredictable, but can be measured.

(l) We wish to study the effects of two food additives (plus a control treatment for a total of three treatments) on the milk productivity of cows. We have three large herds available, each of a different breed, and we expect breed to breed differences in the response. Furthermore, we expect an age effect, which we make explicit by dividing cows into three groups: those which have had 0, 1, and 2 or more previous calves. We have enough resources to study 27 animals through one breeding cycle.

For each of the following, describe the experimental design used and give a skeleton ANOVA (sources and degrees of freedom only).

Problem 17.2

(a) Resveratrol is a substance found in the skin of red grapes that is being investigated for its health attributes. In our experiment, 120 mice are randomly assigned to three treatments (40 to a treatment): control diet, high fat diet, high fat diet plus resveratrol. The response of interest is how long the mice live. Some researchers believe that any positive effects of resveratrol on lifespan are due to its effect on cholesterol, so we also measure cholesterol level at time of death.

(b) New apple varieties are being introduced; we need to figure out which varieties yield more and how irrigation affects yields for these new varieties. Four commercial orchards have agreed to participate in our experiment evaluating three new apple varieties and three irrigation schedules. At each orchard, three adjacent .5 hectare plots are cleared and then each plot is randomly assigned to one of three new varieties. All trees are allowed to grow in a similar environment for eight years. Yields will be measured in years 9, 10, and 11. The three irrigation schedules are randomly assigned to years for each .5 hectare plot. However, we note that there could be year to year variation, so we restrict the randomization so that each irrigation schedule is used once in each year at each orchard.

(c) LDL is bad cholesterol, and HDL is good cholesterol. There are many drugs on the market to lower LDL. We are developing a drug to raise HDL, because we believe that increased HDL will in turn lower LDL. In our pilot study, all subjects are taking standard medication to lower LDL. We randomly assign 100 patients to our new drug and 100 patients to a placebo. After three months on the drug we measure LDL in all patients

to test for differences. We also measure HDL at the same time, as we are hoping that high HDL will be associated with lower LDL.

(d) Three-spined sticklebacks are a marine fish species in which the males build nests and then attempt to attract females to the nest. It appears that fancy nests attract more females. We compare the mating success of male sticklebacks subjected to four treatments: provision of shiny decorations, provision of nonshiny decorations, provision of sticks, and control. Two nesting males are found in each of twelve bays; we anticipate some differences between bays. We randomly assign the four treatments to the twenty four males such that each pair of treatments occurs in two bays.

(e) There was once there a great deal of concern that COX-2 inhibitor pain relievers (e.g., Vioxx and Celebrex) can increase the risk of heart problems. Just this morning (12-21-2004) there was a news story that naproxen sodium based pain relievers (e.g. Aleve) may do the same. We are just finishing up a study. In this study, 100 patients were randomly assigned to five different pain relievers (Vioxx, Celebrex, and three non COX-2 drugs), for a total of 500 patients. At the beginning of the study, we collected a detailed medical history that for each patient that was processed into the form of a score for the patient's risk of heart problems. Now, three years after beginning the study, we wish to compare the total death rate between the five pain relievers.

(f) The Pollution Control Agency wishes to study the effect of an additional control mechanism on the cadmium concentration of the outflow from a municipal waste water treatment plant. Twelve weeks are randomly divided into two groups of six. In the first group, the treatment plant will be run as usual. In the second group, the new control mechanism will also be used. Measurements will be taken during each week and an average cadmium concentration in the outflow will be computed for each week. In addition, an average nickel concentration will also be measured, as nickel and cadmium are often found together in the waste stream (and the control mechanism should not affect nickel concentrations).

(g) We wish to study the effects of air pressure (low or high) and tire type (radial versus all season radial) on gas mileage. We do this by fitting tires of the appropriate type and pressure on a car, driving the car 150 miles around a closed circuit, then changing the tire settings and driving again. We have obtained eight cars for this purpose and can use each car for one day. Unfortunately, we can only do three of the four tire combinations on one day, so we have each factor-level combination missing for two cars.

(h) Metribuzin is an agricultural chemical that may accumulate in soils. We wish to determine whether the amount of metribuzin retained in the soil depends on the amount applied to the soil. To test the accumulation, we select 24 plots. Each plot is treated with one of three levels of metribuzin, with plots assigned to levels at random. After one growing season, we take a sample of the top three cm of soil from each plot and determine the amount of metribuzin in the soil. We also measure the pH of the soil, as pH may affect the ability of the soil to retain metribuzin.

(i) We wish to test the efficacy of dental sealants for reducing tooth decay on molars in children. There are five treatments (sealants A or B applied at either 6 or 8 years of age, and a control of no sealant). We have 40 children, and the five treatments are assigned at random to the 40 children. As a response, we measure the number of cavities on the molars by age 10. In addition, we measure the number of cavities on the nonmolar teeth (this may be a general measure of quality of brushing or resistance to decay).

(j) A national travel agency is considering new computer hardware and software. There are two hardware setups and three competing software setups. All three software setups will run on both hardware setups, but the different setups have different strengths and weaknesses. Twenty branches of the agency are chosen to take part in an experiment. Ten are high sales volume; ten are low sales volume. Five of the high-sales branches are chosen at random for hardware A; the other five get hardware B. The same is done in the low-sales branches. All three software setups are tried at each branch. One of the three software systems is randomly assigned to each of the first 3 weeks of May (this is done separately at each branch). The measured response for each hardware-software combination is a rating score based on the satisfaction of the sales personnel.

Problem 17.3

Pollutants may reduce the strength of bird bones. We believe that the strength reduction, if present, is due to a change in the bone itself, and not a change in the size of the bone. One measure of bone strength is calcium content. We have an instrument which can measure the total amount of calcium in a 1cm length of bone. Bird bones are essentially thin tubes in shape, so the total amount of calcium will also depend on the diameter of the bone.

Thirty-two chicks are divided at random into four groups. Group 1 is a control group and receives a normal diet. Each other group receives a diet including a different toxin (pesticides related to DDT). At 6 weeks, the chicks are sacrificed and the calcium content (in mg) and diameter (in mm) of the right femur is measured for each chick (data set `BirdBones`).

Control		P #1		P #2		P #3	
C	Dia	C	Dia	C	Dia	C	Dia
10.41	2.48	12.10	3.10	10.33	2.57	10.46	2.6
11.82	2.81	10.38	2.61	10.03	2.48	8.64	2.17
11.58	2.73	10.08	2.49	11.13	2.77	10.48	2.64
11.14	2.67	10.71	2.69	8.99	2.30	9.32	2.35
12.05	2.90	9.82	2.43	10.06	2.56	11.54	2.89
10.45	2.45	10.12	2.52	8.73	2.18	9.48	2.38
11.39	2.69	10.16	2.54	10.66	2.65	10.08	2.55
12.5	2.94	10.14	2.55	11.03	2.73	9.12	2.29

Analyze these data with respect to the effect of pesticide on calcium in bones.

Advertisers wish to determine if program content affects the success of their ads on those programs. They produce two videos, one containing a depressing drama and some ads, the second containing an upbeat comedy and the same ads. Twenty-two subjects are split at random into two groups of eleven, with the first group watching the drama and the second group watching the comedy. After the videos, the subjects are asked several questions, including "How do you feel?" and "How likely are you to buy?" one of the products mentioned in the ads. "How do you feel" was on a 1 to 6 scale, with 1 being happy and 6 being sad. "How likely are you to buy?" was also on a 1 to 6 scale, with 6 being most likely (data set `SadAds`).

Problem 17.4

Drama		Comedy	
Feel	Buy	Feel	Buy
5	1	3	1
1	3	2	2
5	1	3	1
5	3	2	3
4	5	4	1
4	3	1	3
5	2	1	4
6	1	2	4
5	5	3	1
3	4	4	1
4	1	2	2

Analyze these data to determine if program type affects the likelihood of product purchase.

A study has been conducted on the environmental impact of an industrial incinerator. One of the concerns is the emission of heavy metals from the stack, and one way to measure the impact is by looking at metal accumulations in soil and seeing if nearby sites have more metals than distant sites (presumably due to deposition of metals from the incinerator).

Problem 17.5

Eleven sites of one hectare each (100 m by 100 m) were selected around the incinerator. Five sites are on agricultural soils, while the other six are on forested soils. Five of the sites were located near the incinerator (on their respective soil types), while the other sites were located far from the incinerator. At each site, nine locations are randomly selected within the site and mineral soil sampled at each location. We then measure the mercury content in each sample (mg/kg).

Complicating any comparison is the fact that heavy metals are generally held in the organic portion of the soil, so that a soil sample with more carbon will tend to have more heavy metals than a sample with less carbon, regardless of the deposition histories of the samples, soil type, etc. For this reason, we also measure the carbon fraction of each sample (literally the fraction of the soil sample that was carbon).

The data given below (data set `Incinerator`) are site averages for carbon and mercury. Analyze these data to determine if there is any evidence

of an incinerator effect on soil mercury.

Soil	Distance	Carbon	Mercury
Agricultural	Near	.0084	.0128
Agricultural	Near	.0120	.0146
Agricultural	Near	.0075	.0130
Agricultural	Far	.0087	.0133
Agricultural	Far	.0105	.0090
Forest	Near	.0486	.0507
Forest	Near	.0410	.0477
Forest	Far	.0370	.0410
Forest	Far	.0711	.0613
Forest	Far	.0358	.0388
Forest	Far	.0459	.0466

Show that the covariate-adjusted means using the covariate $\tilde{x}$ equal the unadjusted treatment means.   **Question 17.1**

Chapter 18

Fractional Factorials

This chapter and the next deal with *treatment design.* We have been us-
ing treatments that are the factor-level combinations of two or more factors.
These factors may be fixed or random or nested or crossed, but we have a
regular array of factor combinations as treatments. Treatment design investi-
gates other ways for choosing treatments. This chapter investigates fractional
factorials, that is, use of a subset of the factor-level combinations in a facto-
rial treatment structure.

18.1 Why Fraction?

Factorial treatment structure has the benefits that it is efficient and allows us
to study main effects and interactions, but factorials can become really big.
For seven factors, the smallest factorial has $2^7 = 128$ treatments and units.
There are 127 degrees of freedom in such an experiment, with 7 degrees
of freedom for main effects, 21 degrees of freedom for two-factor interac-
tions, 35 degrees of freedom for three-factor interactions, and 64 degrees of
freedom for four-, five-, six-, and seven-factor interactions. In many exper-
iments, we either don't expect high-order interactions or we are willing to
ignore them at the current stage of experimentation, so we construct a surro-
gate error by pooling high-order interactions. For example, pooling fourth-
and higher-order interactions into error in the 2^7 gives us 64 degrees of free-
dom for error.

What does a big factorial such as a 2^7 give us? First, it gives us a large
sample size for estimating main effects and interactions; this is a very good
thing. Second, it allows us to estimate many-way interactions; this may or
may not be useful, depending on the experimental situation. Third, the abun-
dant high-order interactions give us many degrees of freedom for construct-
ing a surrogate error.

Larger sample sizes always give us more precise estimates, but there are
diminishing returns for the second and third advantages. In some experiments
we either do not expect high-order interactions, or we are willing to ignore

them in the current problem. For such an experiment, being able to estimate high-order interactions is not a major advantage. Similarly, more degrees of freedom for error are always better, but the improvement in power and confidence interval length is modest after 15 degrees of freedom for error and very slight after 30.

Thus the full factorial may be wasteful or infeasible if

- We believe there are no high-order interactions or that they are ignorably small, or

- We are just screening a large number of treatments to determine which affect the response and will study interactions in subsequent experiments on the active factors, or

- We have limited resources.

We need a design that retains as many of the advantages of factorials as possible, but does not use all the factor-level combinations.

A *fractional-factorial* design is a modification of a standard factorial that allows us to get information on main effects and low-order interactions without having to run the full factorial design. Fractional factorials are closely related to the confounding designs of Chapter 15, which you may wish to review. In fact, the simplest way to describe a fractional factorial is to confound the factorial into blocks, but only run one of the blocks.

Fractional factorial looks at main effects and low-order interactions

18.2 Fractioning the Two-Series

A 2^k factorial can be confounded into two blocks of size 2^{k-1}, four blocks of size 2^{k-2}, and in general 2^q blocks of size 2^{k-q}. A 2^{k-1} fractional factorial is a design with k factors each at two levels that uses 2^{k-1} experimental units and factor-level combinations. We essentially block the 2^k into two blocks but only run one of the blocks. In general, a 2^{k-q} fractional factorial is a design with k factors each at two levels that uses 2^{k-q} experimental units and factor-level combinations. Again, this design is one block of a confounded 2^k factorial. The principal block of a confounded design becomes the *principal fraction*, and alternate blocks become *alternate fractions*.

A fraction is one block of a confounded design

Principal and alternate fractions

We confound a 2^k factorial by choosing one or more defining contrasts. These defining contrasts are factorial effects that will be confounded with block differences. We construct blocks by partitioning the factor-level combinations into 2^q groups according to whether they are ± 1 on the defining contrasts, or equivalently by whether an even or odd number of factors from the defining contrasts are at the high level in the factor-level combination or by whether the L values are 0 or 1.

Review of confounding

In the confounded 2^k, all possible plus/minus, even/odd, or 0/1 combinations for the defining contrasts occur somewhere in the design, though in different blocks. For example, with two defining contrasts, we will have plus and plus, minus and plus, plus and minus, and minus and minus blocks. A

fractional factorial is a single block of this design, so only a single plus/minus combination of the defining contrasts occurs: for example, the plus and plus combination. Thus a fractional factorial is a subset of factor-level combinations that has a particular pattern of plus and minus signs on the defining contrasts, or equivalently a particular pattern of even/odd or 0/1 values.

The jargon and notation of fractional factorials are slightly different from confounding. Recall the tables of plus and minus signs such as Table 15.1 that we used in two-series design. Augment such tables with a column of all plus signs labeled I. Defining contrasts are the effects that we confound to produce confounded factorials; we call these contrasts *generators* or *words* when we work with just a fraction of the design. In a fraction of a two-series, each generator for the design will always be plus or always be minus; thus for each generating word W, either $I = W$ or $I = -W$ will be true on the fraction. The statement $I = W$ is called a *defining relation.* Note that if $I = W_1$ and $I = -W_2$, then $I = -W_1W_2$; that is, generalized interactions of the generators also have constant sign that can be determined from the defining relations.

Example 18.1 Quarter fraction of a 2^5 design

Construct a 2^{5-2} fractional factorial using ABC and –CDE as generators; I = ABC = –CDE = –ABDE is the full set of defining relations. This is the same as confounding into four blocks using the generators ABC and CDE, but then only using the block where ABC is plus and CDE is minus. Using the even/odd rule, ABC is plus when a factor-level combination has an odd number of factors A, B, or C high, and CDE is minus when a factor-level combination has an even number of C, D, or E high.

The eight factor-level combinations in our fraction are

$$a,\ b,\ ade,\ bde,\ ce,\ abce,\ cd,\ abcd\ .$$

In principle we find the fraction by confounding the full factorial and choosing the correct block. However, we know that we can find alternate blocks from the principal block, so we can find alternate fractions from principal fractions. I found our fraction by first finding the principal fraction,

$$(1),\ ab,\ de,\ abde,\ ace,\ bce,\ acd,\ bcd$$

then finding a factor-level combination in the fraction of interest (a), and multiplying everything in the principal fraction by a to get the alternate fraction.

The natural way to estimate the total effect of factor A in a fractional factorial is to subtract the average response where A is low from the average response where A is high. For the 2^{5-2} of Example 18.1, this is the contrast

$$\frac{\overline{y}_a + \overline{y}_{abce} + \overline{y}_{ade} + \overline{y}_{abcd}}{4} - \frac{\overline{y}_{ce} + \overline{y}_b + \overline{y}_{cd} + \overline{y}_{bde}}{4}\ .$$

This amounts to taking the pattern of pluses and minuses for the A contrast

Table 18.1: Table of pluses and minuses for a 2^{5-2} with I = ABC = –CDE.

	A	B	C	D	E	AB	$\cdots$	ABCDE
ce	–	–	+	–	+	+	$\cdots$	–
a	+	–	–	–	–	–		+
b	–	+	–	–	–	–		+
abce	+	+	+	–	+	+		–
cd	–	–	+	+	–	+		–
ade	+	–	–	+	+	–		+
bde	–	+	–	+	+	–		+
abcd	+	+	+	+	–	+	$\cdots$	–

from the complete factorial and just using the elements in it that correspond to the factor-level combinations that we have in our fraction. Part of this reduced table of pluses and minuses is shown in Table 18.1. Using this table, we can compute contrasts for all the factorial effects.

This sounds as if we've just gotten something for nothing. We only have eight observations, but we've (apparently) just extracted estimates of 31 effects and interactions. The laws of physics and economics argue that you don't get something for nothing, and indeed there is a catch here. To see the catch, look at the patterns of signs we use for the C main effect and the AB interaction. These patterns are the same, so our estimate of the C main effect is the same as our estimate of the AB interaction. If we look further, we will also find that the C contrast is the negative of the DE and ABCDE contrasts.

Same contrast for several effects

We say that C, AB, –DE, and –ABCDE are *aliases*, or aliased to each other. Another way of writing this is C = AB = –DE = –ABCDE, meaning that these contrasts have equal coefficients on this fraction. When we apply that contrast, we are estimating the total effect of C, plus the total effect of AB, minus the total effect of DE, minus the total effect of ABCDE, or C + AB – DE – ABCDE. In a 2^{k-q} design, every degree of freedom is associated with 2^q effects that are aliased to each other. So aliases come in pairs for half-fractions, sets of four for quarter-fractions, and so on.

Fractional factorials have aliased effects

There is a simple rule for determining which effects are aliased. Begin with the defining relations, I = ABC = –CDE = –ABDE in our example. Treat I as an identity, multiply all elements of the defining relations by an effect, and reduce exponents mod 2. For example,

Multiply defining relation to get aliases

$$
\begin{array}{lllll}
\text{C} \times \text{I} & = & \text{C} \times \text{ABC} & = & \text{C} \times \text{–CDE} & = & \text{C} \times \text{–ABDE} \\
\text{C} & = & \text{ABC}^2 & = & \text{–C}^2\text{DE} & = & \text{–ABCDE} \\
\text{C} & = & \text{AB} & = & \text{–DE} & = & \text{–ABCDE}
\end{array}
$$

We can continue this to find the complete set of aliases:

$$
\begin{array}{llllll}
I & = & ABC & = & -CDE & = & -ABDE \\
A & = & BC & = & -ACDE & = & -BDE \\
B & = & AC & = & -BCDE & = & -ADE \\
C & = & AB & = & -DE & = & -ABCDE \\
D & = & ABCD & = & -CE & = & -ABE \\
E & = & ABCE & = & -CD & = & -ABD \\
AD & = & BCD & = & -ACE & = & -BE \\
BD & = & ACD & = & -BCE & = & -AE
\end{array}
$$

It is very important to check the aliasing during the design phase of a fractional factorial. In particular, we do not want to have a two-factor interaction as a generator (or generalized interaction of generators), because that would imply that two main effects will be aliased. The more letters in the generators and their interactions the better.

Check to be sure no important effects are aliased to each other

Aliases for more complicated designs follow the same pattern. The defining relation for the fraction will include I and all $2^q - 1$ of the generators and their interactions. For example, consider a 2^{8-4} with generators BCDE, ACDF, ABDG, and –ABCH; the defining relation is I = BCDE = ACDF = ABEF = ABDG = ACEG = BCFG = DEFG = –ABCH = –ADEH = –BDFH = –CEFH = –CDGH = –BEGH = –AFGH = –ABCDEFGH, which is found as the generators, their 6 two-way interactions, their 4 three-way interactions, and their four-way interaction. Thus every degree of freedom has sixteen names and every effect is aliased to fifteen other effects. The full set of aliases for this design is shown in Table 18.2. We see that no main effect is aliased with a two-factor interaction—only three-way or higher. Thus if we could assume that three-factor and higher interactions are negligible, all main effects would be estimated without aliasing to nonnegligible effects.

All effects have $2^q - 1$ aliases in 2^{k-q} design

Every 2^{k-q} fractional factorial contains a complete factorial in some set of $k - q$ factors (possibly many sets), meaning that if you ignore the letters for the other q factors, all 2^{k-q} factor-level combinations of the chosen $k - q$ factors appear in the design. You can use any set of $k - q$ factors that does not contain an alias of I as a subset. For example, the 2^{5-2} in Example 18.1 has an embedded complete factorial with three factors. This design has defining relation I = ABC = –CDE = –ABDE; there are ten sets of three factors, and any triple except ABC or CDE will provide a complete factorial. Consider A, B, and D. Rearranging the treatments in the fraction, we get

Full factorial in $k - q$ factors embedded in 2^{k-q}

$$ce, \ a, \ b, \ abce, \ cd, \ ade, \ bde, \ abcd;$$

ignoring C and E, we get

$$(1), \ a, \ b, \ ab, \ d, \ ad, \ bd, \ abd,$$

which are in standard order for A, B, and D. We cannot do this with A, B, and C; ignoring D and E, we get

$$c, \ a, \ b, \ abc, \ c, \ a, \ b, \ abc;$$

which is not a complete factorial.

Table 18.2: Aliases for 2^{8-4} with generators BCDE, ACDF, ABDG, and –ABCH.

I = BCDE = ACDF = ABEF = ABDG = ACEG = BCFG = DEFG = -ABCH = -ADEH = -BDFH = -CEFH = -CDGH = -BEGH = -AFGH = -ABCDEFGH

A = ABCDE = CDF = BEF = BDG = CEG = ABCFG = ADEFG = -BCH = -DEH = -ABDFH = -ACEFH = -ACDGH = -ABEGH = -FGH = -BCDEFGH

B = CDE = ABCDF = AEF = ADG = ABCEG = CFG = BDEFG = -ACH = -ABDEH = -DFH = -BCEFH = -BCDGH = -EGH = -ABFGH = -ACDEFGH

AB = ACDE = BCDF = EF = DG = BCEG = ACFG = ABDEFG = -CH = -BDEH = -ADFH = -ABCEFH = -ABCDGH = -AEGH = -BFGH = -CDEFGH

C = BDE = ADF = ABCEF = ABCDG = AEG = BFG = CDEFG = -ABH = -ACDEH = -BCDFH = -EFH = -DGH = -BCEGH = -ACFGH = -ABDEFGH

AC = ABDE = DF = BCEF = BCDG = EG = ABFG = ACDEFG = -BH = -CDEH = -ABCDFH = -AEFH = -ADGH = -ABCEGH = -CFGH = -BDEFGH

BC = DE = ABDF = ACEF = ACDG = ABEG = FG = BCDEFG = -AH = -ABCDEH = -CDFH = -BEFH = -BDGH = -CEGH = -ABCFGH = -ADEFGH

ABC = ADE = BDF = CEF = CDG = BEG = AFG = ABCDEFG = -H = -BCDEH = -ACDFH = -ABEFH = -ABDGH = -ACEGH = -BCFGH = -DEFGH

D = BCE = ACF = ABDEF = ABG = ACDEG = BCDFG = EFG = -ABCDH = -AEH = -BFH = -CDEFH = -CGH = -BDEGH = -ADFGH = -ABCEFGH

AD = ABCE = CF = BDEF = BG = CDEG = ABCDFG = AEFG = -BCDH = -EH = -ABFH = -ACDEFH = -ACGH = -ABDEGH = -DFGH = -BCEFGH

BD = CE = ABCF = ADEF = AG = ABCDEG = CDFG = BEFG = -ACDH = -ABEH = -FH = -BCDEFH = -BCGH = -DEGH = -ABDFGH = -ACEFGH

ABD = ACE = BCF = DEF = G = BCDEG = ACDFG = ABEFG = -CDH = -BEH = -AFH = -ABCDEFH = -ABCGH = -ADEGH = -BDFGH = -CEFGH

CD = BE = AF = ABCDEF = ABCG = ADEG = BDFG = CEFG = -ABDH = -ACEH = -BCFH = -DEFH = -GH = -BCDEGH = -ACDFGH = -ABEFGH

ACD = ABE = F = BCDEF = BCG = DEG = ABDFG = ACEFG = -BDH = -CEH = -ABCFH = -ADEFH = -AGH = -ABCDEGH = -CDFGH = -BEFGH

BCD = E = ABF = ACDEF = ACG = ABDEG = DFG = BCEFG = -ADH = -ABCEH = -CFH = -BDEFH = -BGH = -CDEGH = -ABCDFGH = -AEFGH

ABCD = AE = BF = CDEF = CG = BDEG = ADFG = ABCEFG = -DH = -BCEH = -ACFH = -ABDEFH = -ABGH = -ACDEGH = -BCDFGH = -EFGH

As a second example, the factor-level combinations of the 2^{8-4} in Table 18.2 are

$$h,\ afg,\ beg,\ abefh,\ cef,\ acegh,\ bcfgh,\ abc,$$
$$defgh,\ ade,\ bdf,\ abdgh,\ cdg,\ acdfh,\ bcdeh,\ abcdefg\ ,$$

---

1. Choose q generators and get the aliases of I.

2. Find a set of $k - q$ base factors that has an embedded complete factorial.

3. Write the factor-level combinations of the base factors in standard order.

4. Find the aliases of the remaining q factors in terms of interactions of the $k - q$ base factors.

5. Determine the plus/minus pattern for the q remaining factors from their aliased interactions.

6. Add letters to the factor-level combinations of the base factors to indicate when the remaining factors are at their high levels (plus).

Display 18.1: Constructing fractional factorials

which are in standard order for A, B, C, and D.

The embedded complete factorial is a tool for constructing fractional factorials. Display 18.1 gives the steps. Essentially we start with the factor-level combinations of the embedded factorial. Each additional factor is aliased to an interaction of the embedded factorial, so we can determine the pattern of high and low of the additional factors from the interactions of the embedded factors. Add letters to factor-level combinations of the embedded factorial when the additional factors are at the high level.

Use embedded factorial to build fractions

Example 18.2 Treatments in a 2^{8-4} design

Consider the 2^{8-4} of Table 18.2 with generators BCDE, ACDF, ABDG, and –ABCH. We can see from the aliases of I that this design has an embedded factorial in A, B, C, and D. The remaining factors E, F, G, and H can be expressed in terms of interactions of the base factors as E = BCD, F = ACD, G = ABC, and H = –ABD.

Embedded design	E = BCD	F = ACD	G = ABD	H = –ABC	Final design
(1)	-1	-1	-1	1	h
a	-1	1	1	-1	afg
b	1	-1	1	-1	beg
ab	1	1	-1	1	$abefh$
c	1	1	-1	-1	cef
ac	1	-1	1	1	$acegh$
bc	-1	1	1	1	$bcfgh$
abc	-1	-1	-1	-1	abc
d	1	1	1	1	$defgh$
ad	1	-1	-1	-1	ade
bd	-1	1	-1	-1	bdf
abd	-1	-1	1	1	$abdgh$
cd	-1	-1	1	-1	cdg
acd	-1	1	-1	1	$acdfh$
bcd	1	-1	-1	1	$bcdeh$
$abcd$	1	1	1	-1	$abcdefg$

We can see that each factor-level combination has an even number of letters from the sets BCDE, ACDF, and ABDG, and an odd number of letters from ABCH.

18.3 Analyzing a 2^{k-q}

Analysis of a 2^{k-q} is really much like any 2^k except that we must always keep the alias structure in mind. Most fractional factorials have only a single replication, so there will be no estimate of pure error. We must either compute a surrogate error by pooling interaction terms, use a graphical approach such as the half-normal plot, or use Lenth's PSE. Keep in mind that if we pool interaction terms, we must look at all the aliases for a given degree of freedom; some interaction terms are aliased to main effects! Similarly, a normal plot of effects may show that an interaction appears to be large. Check the aliases for that degree of freedom, because it could be aliased to a main effect.

Analyze like 2^k but remember aliasing

Notice that there is some subjectivity in the analysis of a fractional factorial. For example, we could find that only the degree of freedom D = ABC appears to be significant in a 2^{4-1} design with I = ABCD as a defining relation. The most reasonable interpretation is that we are seeing the main effect of D, not an ABC interaction in the absence of any lower-order effects. It is possible that the ABC interaction is large when the A, B, C, AB, AC, and BC effects are null, so we could be making a mistake ascribing this effect to D; but lower-order aliases are usually the safer bet.

Some subjectivity in interpreting aliases

Example 18.3 Welding strength

Taguchi and Wu (1980) describe an experiment carried out to determine

Table 18.3: Design and responses for welding strength data. Data set `Welding`.

	A	B	C	D	E	F	G	H	J	y
gj	−	−	−	−	−	−	+	−	+	40.2
aef	+	−	−	−	+	+	−	−	−	43.7
bgh	−	+	−	−	−	−	+	+	−	44.7
$abefhj$	+	+	−	−	+	+	−	+	+	42.4
ceh	−	−	+	−	+	−	−	+	−	45.9
$acfghj$	+	−	+	−	−	+	+	+	+	42.4
$bcej$	−	+	+	−	+	−	−	−	+	40.6
$abcfg$	+	+	+	−	−	+	+	−	−	42.2
dfh	−	−	−	+	−	+	−	+	−	45.5
$adeghj$	+	−	−	+	+	−	+	+	+	42.4
$bdfj$	−	+	−	+	−	+	−	−	+	40.6
$abdeg$	+	+	−	+	+	−	+	−	−	43.6
$cdefgj$	−	−	+	+	+	+	+	−	+	40.2
acd	+	−	+	+	−	−	−	−	−	44.0
$bcdefgh$	−	+	+	+	+	+	+	+	−	46.5
$abcdhj$	+	+	+	+	−	−	−	+	+	42.5
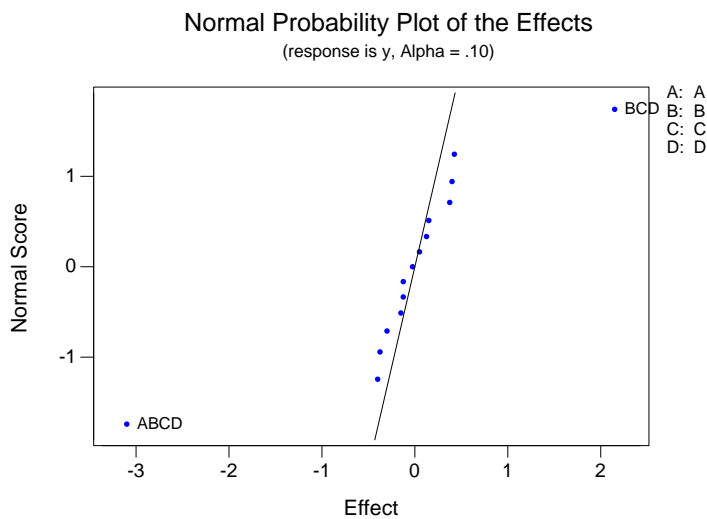
Figure 18.1: Normal plot of effects in welding strength data, using Minitab.

factors affecting the strength of welds. There were nine factors at two levels each to be explored. The full experiment was much too large, so a 2^{9-5} fractional factorial with sixteen units was used. The factors are coded A though J (skipping I); the generators are –ACE, –ADF, –ACDG, BCDH, ABCDJ. The full defining relation is I = –ACE = –ADF = CDEF = –ACDG = DEG

Main Effects Plot (data means) for y
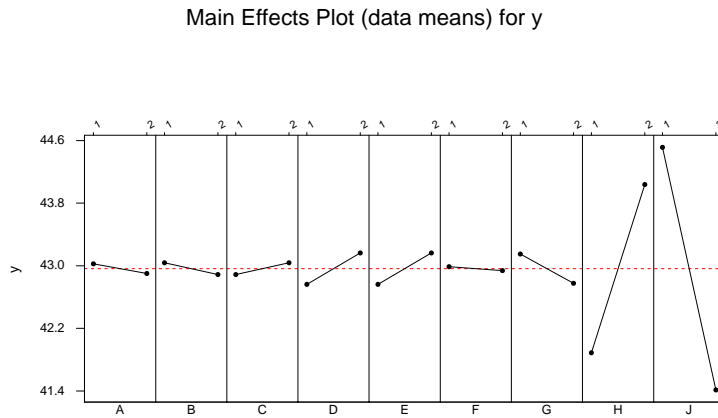
Figure 18.2: Main effects in welding strength data, using Minitab.

= CFG = –AEFG = BCDH = –ABDEH = –ABCFH = BEFH = –ABGH = BCEGH = BDFGH = –ABCDEFGH = ABCDJ = –BDEJ = –BCFJ = ABEFJ = –BGJ = ABCEGJ = ABDFGJ = –BCDEFGJ = AHJ = –CEHJ = –DFHJ = ACDEFHJ = –CDGHJ = ADEGHJ = ACFGHJ = –EFGHJ; every effect is aliased to 31 other effects. The design and responses are given in Table 18.3 (data set `Welding`).

First note that this design has an embedded 2^4 design. A check of the defining relation reveals that ABCD is not aliased to I (nor is any subset of ABCD), so we have a complete embedded factorial in those four factors. The data in Table 18.3 are in standard order for A, B, C, and D, so we may compute the main effects and interactions for A, B, C, and D using Yates' algorithm on the responses in the order presented. Figure 18.1 shows a normal plot of these effects. Only the BCD and ABCD interactions are large. Before we interpret these, we must look at their aliases. We find that BCD is aliased to H, and ABCD is aliased to J, so we are probably seeing main effects of H and J.

Alternatively, we may decide to fit just main effects in an Analysis of Variance and pool all remaining degrees of freedom into error. Minitab output for this approach follows.

```
ractional Factorial Fit

stimated Effects and Coefficients for y (coded units)

erm          Effect        Coef  StDev Coef         T        P
onstant                   42.963      0.1359   316.18    0.000
             -0.125      -0.063      0.1359     -0.46    0.662        ①
             -0.150      -0.075      0.1359     -0.55    0.601
              0.150       0.075      0.1359      0.55    0.601
              0.400       0.200      0.1359      1.47    0.191
              0.400       0.200      0.1359      1.47    0.191
             -0.050      -0.025      0.1359     -0.18    0.860
             -0.375      -0.187      0.1359     -1.38    0.217
              2.150       1.075      0.1359      7.91    0.000
             -3.100      -1.550      0.1359    -11.41    0.000

nalysis of Variance for y (coded units)

ource               DF      Seq SS      Adj SS      Adj MS       F        P
ain Effects          9      59.025      59.025      6.5583   22.20    0.001
esidual Error        6       1.772       1.772      0.2954
otal                15      60.797

lias Structure (up to order 3)
                                                                     ②
 - A*C*E - A*D*F + A*H*J - B*G*J + C*F*G + D*E*G
 - C*E - D*F + H*J - B*G*H - C*D*G - E*F*G
 - G*J - A*G*H + C*D*H - C*F*J - D*E*J + E*F*H
 - A*E + F*G - A*D*G + B*D*H - B*F*J + D*E*F - E*H*J
 - A*F + E*G - A*C*G + B*C*H - B*E*J + C*E*F - F*H*J
 - A*C + D*G - A*F*G - B*D*J + B*F*H + C*D*F - C*H*J
 - A*D + C*G - A*E*G - B*C*J + B*E*H + C*D*E - D*H*J
 - B*J + C*F + D*E - A*B*H - A*C*D - A*E*F
 + A*J - A*B*G + B*C*D + B*E*F - C*E*J - D*F*J
 + A*H - B*G - B*C*F - B*D*E - C*E*H - D*F*H
```

This gives us 9 main-effects degrees of freedom and 6 error degrees of freedom. ① shows the estimated effects, their standard errors, and p-values. Again, only H and J are significant, which can be seen visually in Figure 18.2. Note that Minitab also computes the low-order aliases of any terms in the model ②.

18.4   Resolution and Projection

Fractional factorials are classified according to their *resolution,* which tells us which types of effects are aliased. A resolution R design is one in which no interaction of j factors is aliased to an interaction with fewer than $R - j$ factors. For example, in a resolution three design, no main effect ($j = 1$) is aliased with any other main effect, but main effects can be aliased with two-factor interactions ($R - j = 2$). In a resolution four design, no main effect ($j = 1$) is aliased with any main effect or two-factor interaction, but main effects can be aliased with three-factor interactions ($R - j = 3$), and two-factor interactions ($j = 2$) can be aliased with two-factor interactions

Resolution determines how short aliases can be

$(R - j = 2)$. In a resolution five design, no main effect is aliased with any main effect, two-factor interaction, or three-factor interaction, but main effects can be aliased with four-factor interactions. Two-factor interactions are not aliased with main effects or two-factor interactions, but they may be aliased with three-factor interactions.

A fractional factorial of resolution R has R letters in the shortest alias of I, so we call these R-letter designs. In fact, this is the easy way to remember what resolution means. Resolution is usually written as a Roman numeral subscript for the design. The 2^{8-4} design in Table 18.2 has 14 four-letter aliases of I and an eight-letter alias, so it is resolution IV and is written 2_{IV}^{8-4}.

Resolution equals minimum number of letters in aliases of I

We never want a resolution II design, because such a design would alias two main effects. Thus the minimum acceptable resolution is III. When choosing generators for a 2^{k-p} factorial, we want to obtain as high a resolution as possible so that the aliases of main effects will be interactions with as high an order as possible.

Maximize resolution

Resolution isn't the complete picture. Consider three 2^{7-2} designs, with defining relations I = ABCF = BCDG = ADFG, I = ABCF = ADEG = BCDEFG, and I = ABCDF = ABCEG = DEFG. All four designs are resolution IV, but we prefer the last design because it has only one 4-letter alias, while the others have two or three. Designs that have the minimum possible number of short aliases are called *minimum-aberration* designs. Thus we want maximum resolution and minimum aberration.

Minimize aberration

Resolution III designs have some main effects aliased to two-factor interactions. If we believe that only main effects are present and all interactions are negligible, then a resolution III design is sufficient for estimating main effects. Resolution III designs are called *main-effects designs* for this reason. If we believe that some two-factor interactions may be nonnegligible but all three-way and higher interactions are negligible, then a resolution IV design is sufficient for main effects.

Main-effects designs

Low-resolution fractional factorials are often used as screening designs, where we are trying to screen many factors to see if any of them has an effect. This is usually an early stage of investigation, so we do not usually require information about interactions, though we would not throw away such information if we can get it.

Screening experiments

We have constructed fractional factorials by augmenting an embedded complete factorial. *Projection* of factorials is somewhat the reverse process, in that we collapse a fractional factorial onto a complete factorial in a subset of factors. A 2^{k-q} fractional factorial of resolution R contains a complete factorial in any set of at most $R - 1$ factors. If R is less than $k - q$, then this embedded factorial is replicated. There may also be *some* sets of R or more factors that form a complete factorial, but you are guaranteed a complete factorial for *any* set of $R - 1$ factors.

Projection onto embedded factorial

For example, consider the 2_{IV}^{7-2} design with defining relation I = ABCDF = ABCEG = DEFG. This design contains a replicated complete factorial in any set of three factors. It also contains a complete factorial in all sets of four factors except D, E, F, and G, which cannot form a complete factorial

because their four-factor interaction is aliased to I.

Fractional factorials can be projected onto an embedded factorial during analysis. For example, a half-normal plot of effects in a resolution IV design might indicate that factors A, D, and E look significant. Projection then treats the data as if they were a full factorial in the factors A, D, and E and proceeds with the analysis. Notice that the p-values obtained in this way are somewhat suspect. We have put "big" effects into the model and "small" effects wind up in error, so F-statistics and other tests tend to be too big, and p-values tend to be too small.

Project onto significant factors

Example 18.4 Welding strength, continued

We found in Example 18.3 that factors H and J were significant. This was a resolution III design, so we can project it onto a factorial in H and J. SAS output for this approach follows.

```
ependent Variable: Y
                            Sum of          Mean
ource            DF         Squares         Square    F Value    Pr > F

odel              3      56.9925000     18.9975000      59.91    0.0001

rror             12       3.8050000      0.3170833

ource            DF       Type I SS    Mean Square    F Value    Pr > F

                  1      18.4900000     18.4900000      58.31    0.0001
                  1      38.4400000     38.4400000     121.23    0.0001
*J                1       0.0625000      0.0625000       0.20    0.6650
```

We see an ANOVA for H, J, and their interaction. The main effects are highly significant, as we saw in the earlier analysis. Here we also see that there is no evidence of interaction.

18.5 Confounding a Fractional Factorial

We can run a 2^{k-q} design in incomplete blocks by confounding one or more degrees of freedom with block differences, just as we did for complete two-series factorials. The only difference is that each defining contrast we confound is aliased with $2^q - 1$ other effects. Similarly, the generalized interactions of the defining contrasts and their aliases are also confounded.

Confound fractions using defining contrasts

Example 18.5 2^{8-4} in two blocks of eight

Example 18.2 has generators BCDE, ACDF, ABDG, and –ABCH, and the factor-level combinations of this fraction are

$$h, \ afg, \ beg, \ abefh, \ cef, \ acegh, \ bcfgh, \ abc,$$
$$defgh, \ ade, \ bdf, \ abdgh, \ cdg, \ acdfh, \ bcdeh, \ abcdefg \ .$$

We must choose a degree of freedom to confound, and Table 18.2 shows that all degrees of freedom have either main-effect or two-factor interaction aliases. We don't want to confound a main effect, so we will confound a two-factor interaction, say AB and its aliases ACDE = BCDF = EF = DG = BCEG = ACFG = ABDEFG = –CH = –BDEH = –ADFH = –ABCEFH = –ABCDGH = –AEGH = –BFGH = –CDEFGH.

To do the confounding, we put all the factor-level combinations with an even number of the letters A and B in one block, and those with an odd number in the other block. These blocks are

$$h, \ abefh, \ cef, \ abc, \ defgh, \ abdgh, \ cdg, \ abcdefg$$

and

$$afg, \ beg, \ acegh, \ bcfgh, \ ade, \ bdf, \ acdfh, \ bcdeh \ .$$

We could have used any of the aliases of AB to get the same blocks. For example, the first block has an even number of B, C, D, and F, and the second block has an odd number.

18.6   De-aliasing

Aliasing is the price that we pay for using fractional factorials. Sometimes, aliasing is just a nuisance and it doesn't really affect our analysis. Other times aliasing is crucial. Consider the 2^{5-2} design with defining relation I = ABC = –CDE = –ABDE. This design has eight units and 7 degrees of freedom. Suppose that 3 of these degrees of freedom look significant, namely those associated with the main effects of A, C, and E. We cannot interpret the results until we look at the alias structure, and when we do, we find that A = BC = –ACDE = –BDE, C = AB = –DE = –ABCDE, and E = ABCE = –CD = –ABD. The most reasonable explanation of our results is that the main effects of A, C, and E are significant, because other possibilities such as A, C, and the CD interaction seem less plausible. Here aliasing was a nuisance but didn't hurt much.

Check aliases to interpret results

Suppose instead that the 3 significant degrees of freedom are associated with the main effects of A, B, and C. Now the aliases are A = BC = –ACDE = –BDE, B = AC = –BCDE = –ADE, and C = AB = –DE = –ABCDE. There are four plausible scenarios for significant effects: A, B, and C; A, B, and AB; B, C and BC; or A, C, and AC. All of these interpretations fit the results, and we cannot decide between these interpretations with just these data. We either need additional data or external information that certain interactions are unlikely to choose among the four.

Aliasing can leave unresolved ambiguity

> Fractional factorials can help us immensely by letting us reduce the number of units needed, but they can leave many questions unanswered.

The problem, of course, is that our fractional designs have aliasing. We can *de-alias* by obtaining additional data. Consider the four possible fractions of a 2^5 using ABC and CDE as generators:

De-aliasing breaks aliases by running an additional fraction

ABC	CDE	ABDE								Treatments
−	−	+	(1)	ab	acd	bcd	ace	bce	de	abde
+	−	−	a	b	cd	abcd	ce	abce	ade	bde
−	+	−	ac	bc	d	abd	e	abe	acde	bcde
+	+	+	c	abc	ad	bd	ae	be	cde	abcde

Our original fraction is the second one in this table, where ABC is plus and CDE is minus. If we run an additional fraction, then we will have a half-fraction of a 2^5 run in two blocks of size eight. The aliasing for the half-fraction is the aliasing that is in common to the two quarter-fractions that we use. The defining contrast for blocking is the aliasing that differs between the two fractions.

Suppose that we run the third fraction as an additional fraction. The only aliasing in common to the two fractions is I = −ABDE, so this is the defining relation for the half-fraction. The aliasing that changes between the two fractions is ABC = −CDE, so this is the defining contrast for the confounding.

Note that if we knew ahead of time that we were going to run a second quarter-fraction, we could have designed a resolution V fraction at the start. By proceeding in two steps, we wound up with resolution IV. The advantage of the two-step procedure is that we might have been able to stop at eight units if the three active factors had been any three other than ABC or CDE; we were just unlucky.

18.7 Fold-Over

Resolution III fractions are easy to construct, but resolution IV designs are more complicated. *Fold-over* is a technique related to de-aliasing for producing resolution IV designs from resolution III designs. In particular, fold-over produces a 2_{IV}^{k-q} design from a $2_{III}^{(k-1)-q}$ design.

Resolution III fractions are easy to produce. Choose a set of base factors for an embedded factorial, and alias every additional factor to an interaction of the base factors. This will always be resolution III or higher.

To use fold-over, start with a $2_{III}^{(k-1)-q}$ design in the first $k-1$ factors, and produce the table of plus and minus signs for these $k-1$ factors. Augment this table with an additional column of all minuses, labeled for factor k. Now double the number of runs by adding the inverse of every row. That is, switch all plus signs to minus, and all minus signs to plus, including the column for factor k that was all minus signs. The result is a 2_{IV}^{k-q}. The generators for the full design are the generators from the $2_{III}^{(k-1)-q}$, with reversed signs and factor k appended to any generator with an odd number of letters. Note that even though we have constructed this with two fractions, the design is run in one randomization.

Example 18.6 Fold-over for a 2_{IV}^{15-10}

A 2^{15-10}_{IV} design is too big for most tables, and you will need to work hard to find one by trial and error, but fold-over will do the job easily. Begin with a 2^{14-10} design. We will use the generators AB = E, AC = F, AD = G, BC = H, BD = J, CD = K, ABC = L, ABD = M, ACD = N, BCD = O. This just aliases ten additional factors to interactions of the first four. The factor-level combinations and columns of pluses and minuses for the main effects are in the top half of Table 18.4. This includes a column of all minuses for the fifteenth factor P.

In the bottom half, we reverse all the signs from above to produce the second half of the design. In this half, P is always plus. The generators for the full design are –ABEP, –ACFP, –ADGP, –BCHP, –BDJP, –CDKP, ABCL, ABDM, ACDN, BCDO; the odd-length generators for the resolution III design (ABE, ACF, ADG, BCH, BDJ, CDK, and ABC) gain a –P in the fold-over design. There are 105 four-factor, 280 six-factor, 435 eight-factor, 168 ten-factor, and 35 twelve-factor aliases of I in this fold-over design, a complete enumeration of which you will be spared.

18.8   Sequences of Fractions

De-aliasing makes routine use of fractional factorials possible, because we can always use additional fractions to break any aliases that are giving us trouble. In particular, one thing that makes fractional factorials attractive is the ability to run fractions in sequence.

For example, suppose you have six factors that you wish to explore, and money for 32 experimental units. You could use those 32 units to run a 2^{6-1}_{VI} design. Or you could use 16 of those units and run a 2^{6-2}_{IV} design with ABCE and BCDF as generators and save the remaining 16. Why is the second approach often better? If three or fewer factors are active, then you have a replicated complete factorial in those three factors (projection of a fraction). In this case, these first 16 units may be enough to answer our questions. If more factors are active—in particular if A, B, C, and E or B, C, D, and F are active—we can always use the remaining 16 units to run an additional fraction, and we can choose that fraction to break aliases that appear troublesome in the first fraction. The combined quarter-fractions are as good as the original half-fraction (except for a single degree of freedom between the two blocks), because we can choose our second quarter-fraction after seeing the first.

Sequences of fractions can save money

Use results of first fraction to select later fractions

Thus by using a sequence of fractions, you can often learn everything you need to learn with fewer units; and if you cannot, you can use the first fraction to guide your choice of subsequent fraction for remaining units.

Sequences of fractions make sense when each experiment is of short duration so that running experiments in sequence is feasible. If each experiment takes months to complete (for example, many agronomy experiments), then a sequence of fractions is a poor choice of design.

Sequences need quick turnaround

Table 18.4: Folding over to produce a 2^{15-10}_{IV}.

	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P
$efghjk$	−	−	−	−	+	+	+	+	+	+	−	−	−	−	−
$ahjklmn$	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
$bfgklmo$	−	+	−	−	−	+	+	−	−	+	+	+	−	+	−
$abekno$	+	+	−	−	+	−	−	−	−	+	−	−	+	+	−
$cegjlno$	−	−	+	−	+	−	+	−	+	−	+	−	+	+	−
$acfjmo$	+	−	+	−	−	+	−	−	+	−	−	+	−	+	−
$bcghmn$	−	+	+	−	−	−	+	+	−	−	−	+	+	−	−
$abcefhl$	+	+	+	−	+	+	−	+	−	−	+	−	−	−	−
$defhmno$	−	−	−	+	+	+	−	+	−	−	−	+	+	+	−
$adghlo$	+	−	−	+	−	−	+	+	−	−	+	−	−	+	−
$bdfjln$	−	+	−	+	−	+	−	−	+	−	+	−	+	−	−
$abdegjm$	+	+	−	+	+	−	+	−	+	−	−	+	−	−	−
$cdeklm$	−	−	+	+	+	−	−	−	−	+	+	+	−	−	−
$acdfgkn$	+	−	+	+	−	+	+	−	−	+	−	−	+	−	−
$bcdhjko$	−	+	+	+	−	−	−	+	+	+	−	−	−	+	−
$abcdefghjklmno$	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−
$abcdlmnop$	+	+	+	+	−	−	−	−	−	−	+	+	+	+	+
$bcdefgop$	−	+	+	+	+	+	+	−	−	−	−	−	−	+	+
$acdehjnp$	+	−	+	+	+	−	−	+	+	−	−	−	+	−	+
$cdfghjlmp$	−	−	+	+	−	+	+	+	+	−	+	+	−	−	+
$abdfhkmp$	+	+	−	+	−	+	−	+	−	+	−	+	−	−	+
$bdeghklnp$	−	+	−	+	+	−	+	+	−	+	+	−	+	−	+
$adefjklop$	+	−	−	+	+	+	−	−	+	+	+	−	−	+	+
$dgjkmnop$	−	−	−	+	−	−	+	−	+	+	−	+	+	+	+
$abcgjklp$	+	+	+	−	−	−	+	−	+	+	+	−	−	−	+
$bcefjkmnp$	−	+	+	−	+	+	−	−	+	+	−	+	+	−	+
$aceghkmop$	+	−	+	−	+	−	+	+	−	+	−	+	−	+	+
$cfhklnop$	−	−	+	−	−	+	−	+	−	+	+	−	+	+	+
$abfghjnop$	+	+	−	−	−	+	+	+	+	−	−	−	+	+	+
$behjlmop$	−	+	−	−	+	−	−	+	+	−	+	+	−	+	+
$aefglmnp$	+	−	−	−	+	+	+	−	−	−	+	+	+	−	+
p	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+

18.9   Fractioning the Three-Series

Fractional factorials for the three-series are constructed in the same way as the two-series: confound the full factorial into blocks and then run just one block. Three-series factorials are confounded into 3, 9, 27, and other powers of three blocks, so three-series can be fractioned into fractions of one third, one ninth, and so on.

Recall that the factor levels in a three-series are represented by the digits 0, 1, or 2, and that all degrees of freedom are partitioned into two-degree-of-freedom bundles. The bundles are obtained by splitting the factor-level combinations according to their values on a defining split L. For example,

the defining split $A^1B^1C^2$ separates the factor-level combinations into three groups according to

$$L = 1 \times x_A + 1 \times x_B + 2 \times x_C \bmod 3 \ ,$$

where x_A, x_B, and x_C are the levels of factors A, B, and C; L takes the values 0, 1, or 2. The factor-level combinations that have value 0 for the defining split(s) form the principal block, and all others are alternate blocks. These become principal and alternate fractions. The defining splits are the generators for the fraction.

In a 2^{k-q} factorial, every degree of freedom has 2^q names, and every effect is aliased to $2^q - 1$ other effects. It's just a little more complicated for three-series fractions. In a 3^{k-1}, the constant is aliased to a two-degree-of-freedom split (the generator); all other two-degree-of-freedom bundles have three names, and all other splits are aliased to two other splits. If W is the generator, then the aliases of a split P are PW and PW^2. (Recall that exponents of these products are reduced modulo 3, and if the leading nonzero exponent is a 2, double the exponents and reduce modulo 3 again.) For example, the aliases in a 3^{3-1} with $W = A^1B^2C^2$ as generator are

	W	W^2
I	$A^1B^2C^2$	
A	$A^1B^1C^1 = A(A^1B^2C^2)$	$B^1C^1 = A(A^1B^2C^2)^2$
B	$A^1C^2 = B(A^1B^2C^2)$	$A^1B^1C^2 = B(A^1B^2C^2)^2$
C	$A^1B^2 = C(A^1B^2C^2)$	$A^1B^2C^1 = C(A^1B^2C^2)^2$
A^1B^1	$A^1C^1 = A^1B^1(A^1B^2C^2)$	$B^1C^2 = A^1B^1(A^1B^2C^2)^2$

In a 3^{k-2}, the constant is aliased to four two-degree-of-freedom splits; all other two-degree-of-freedom bundles have nine names, and all other splits are aliased to eight other splits. Using two generators W_1 and W_2, the aliases of I are W_1, W_2, W_1W_2, and $W_1W_2^2$. Which generator is labeled one or two does not matter, because $W_1W_2^2 = W_1^2W_2$ after reducing exponents modulo 3 and making the leading nonzero exponent a 1. The aliases of any other split P are PW_1, PW_2, PW_1W_2, $PW_1W_2^2$, PW_1^2, PW_2^2, $PW_1^2W_2^2$, and $PW_1^2W_2$. (Again, reduce exponents modulo 3; double and reduce modulo 3 again if the leading nonzero exponent is not a 1.) For a 3^{4-2} factorial with generators $A^1B^1C^1$ and $B^1C^2D^1$, the complete alias structure is

	W_1	W_2	W_1W_2	$W_1W_2^2$
I	$A^1B^1C^1$	$B^1C^2D^1$	$A^1B^2D^1$	$A^1C^2D^2$
A	$A^1B^2C^2$	$A^1B^1C^2D^1$	$A^1B^1D^2$	$A^1C^1D^1$
B	$A^1B^2C^1$	$B^1C^1D^2$	A^1D^1	$A^1B^1C^2D^2$
C	$A^1B^1C^2$	B^1D^1	$A^1B^2C^1D^1$	A^1D^2
D	$A^1B^1C^1D^1$	$A^1C^2D^2$	$A^1B^2D^2$	A^1C^2

	W_1^2	W_2^2	$W_1^2W_2^2$	$W_1^2W_2$
I				
A	B^1C^1	$A^1B^2C^1D^2$	B^1D^2	C^1D^1
B	A^1C^1	C^1D^2	$A^1B^1D^1$	$A^1B^2C^2D^2$
C	A^1B^1	$A^1C^1D^1$	$A^1B^2C^2D^1$	$A^1C^1D^2$
D	$A^1B^1C^1D^2$	B^1C^2	A^1B^2	$A^1C^2D^1$

Further fractions require more generators. A 3^{k-q} has q generators W_1 through W_q. The constant is aliased to $1 + 3 + \cdots + 3^{q-1}$ two-degree-of-freedom splits; these splits aliased to I are of the form $W_1^{i_1}W_2^{i_2}\cdots W_q^{i_q}$ where the exponents are 0, 1, or 2, and the first nonzero exponent is a 1. All other two-degree-of-freedom bundles have 3^q names, and all other splits are aliased to $3^q - 1$ other splits. The aliases of a split P are products of the form $PW_1^{i_1}W_2^{i_2}\cdots W_q^{i_q}$, where the exponents i_j are allowed to range over all 3^q combinations of 0, 1, and 2. There are $1 + 3 + \cdots + 3^{k-q-1}$ sets of aliases in addition to the aliases of I.

Resolution in the 3^{k-q} is the same as in the two-series: a fractional factorial has resolution R if no interaction of j factors is aliased to an interaction of fewer than $R - j$ factors. And again like the two-series, the resolution of a 3^{k-q} is the number of letters in the shortest alias of I.

We can construct a 3^{k-q} using embedded factorials as we did for two-series. In the 3^{3-1} described above, recall the aliasing $C = A^1B^2$. Construct a full factorial in A and B, and then set the levels of C according to the A^1B^2 interaction; this will generate the fraction. Consider the following table:

00 0	01 2	02 1
10 1	11 0	12 2
20 2	21 1	22 0

The pairs of digits form a complete 3^2 design, and the single digits are the values of

$$1 \times x_A + 2 \times x_B \bmod 3 \ ,$$

the A^1B^2 interaction. These are also the levels of C for the principal fraction. Group the triples together, and we have the principal fraction of a 3^{3-1} with generator $A^1B^2C^2$. If we want an alternate fraction, use

$$1 \times x_A + 2 \times x_B + 1 \bmod 3$$

or

$$1 \times x_A + 2 \times x_B + 2 \bmod 3$$

to generate the levels of C.

18.10   Problems with Fractional Factorials

Fractional factorials can be extremely advantageous in situations where we want to screen factors, can ignore interactions, or have restricted resources. However, the sophistication of the fractional factorial gives us many ways in which to err, and fractional factorials are a bit more brittle than complete factorials in the face of real-world data. Daniel (1976) discusses these problems in detail.

Fractions offer many chances for mistakes

Here are some common pitfalls that you must try to avoid when using fractional factorials. During the design stage, you can make your fractional factorial too large or too small. A design that is too small tries to estimate too many effects for the number of experimental units used; this is called oversaturation. Designs that are too small tend to be limited in how you can estimate error, because all the degrees of freedom are tied up in interesting effects, and resolution tends to be small. Designs that are too large are being wasteful of resources; you may be able to estimate all terms of interest with a smaller design. This ties in with power. Fractional designs have smaller sample sizes and thus less power for a given set of effects and error variance. When planning the size of the design, we need to keep power in mind. All of these design issues depend on having at least some prior knowledge or belief of how the system works. This will allow us to decide what resolution and replication is needed.

Choose fraction size carefully

In the analysis stage, the most obvious problem is dealing incorrectly with aliasing. You thus wind up with a misinterpretation of which effects are important. You may also miss a need to de-alias. Finally, outliers and missing data tend to cause more problems for fractional factorials than complete factorials. For example, consider an outlier in a 2^{k-q}. In the complete two-series, an outlier can sometimes be detected by a pattern of smallish effects of about the same size, usually high-order interactions. In the fraction, many degrees of freedom have a main effect or low-order interaction in their aliases, so there are few opportunities to see the flat pattern in effects that we expect to be null.

Check aliasing and watch for bad data

18.11   Using Fractional Factorials in Off-Line Quality Control

One of the areas in which fractional factorials and related designs have been used with much success, profit, and acclaim is off-line quality control. Quality control has on-line and off-line aspects. On-line means "on the production line"; on-line quality control includes inspection of manufactured parts to make sure that they meet specifications. Off-line quality control is off the production line; this includes designing the product and manufacturing process so that the product will meet specifications when manufactured. The explicit goal is to have the product on target, with minimum variation around the target.

Goal of off-line quality control is to make products on target with minimum variation

Suppose that you manufacture exhaust tubing for the automotive industry. Your client orders a tubing part that should be 2.1 meters long and bent into a specific shape; parts from 2.09 to 2.11 meters in length are acceptable. One step of the manufacturing process is cutting the tubing to length. On-line quality control will include inspection of the cut tubing and rejection of those tubes out of specification. Off-line quality control designs the tube cutting process so that the average tube length is 2.1 meters and the variation around that average is as small as possible.

Off-line quality control has become quite the rage under the banner of "Taguchi methods," named for Genechi Taguchi, the Japanese statistician who developed and advocated the methods. The principle of off-line quality control is to put a product on target with minimum variation. This principle is absolutely golden, but the exact methods Taguchi recommended for achieving this have flaws and inefficiencies in both design and analysis (see Box, Bisgaard, and Fung 1988 or Pignatiello and Ramberg 1991). What we discuss here is very much in the spirit of Taguchi, but the analysis approach is closer to Box (1988).

Taguchi methods

Most manufacturing processes have many controllable design parameters. For the exhaust tubes, design parameters include the speed at which tubing moves down the line, the air pressure for tubing clamps, cutting saw speed, the type of sensor for recognizing the end of a tube, and so on. These parameters might influence product quality, but we generally don't know which ones are important. Manufacturing processes also have uncontrollable aspects, including variation in raw materials and environmental variation such as temperature and humidity. Some of these "uncontrollables" can actually be controlled under laboratory or testing conditions. Taguchi uses the term "inner noise" for variation that arises from changes in the controllable parameters and the term "outer noise" for variation due to the uncontrollable parameters.

Inner noise controllable, outer noise uncontrollable

18.11.1   Designing an off-line quality experiment

We want to find settings for the controllable variables so that the product is on target and the variation due to the outer noise is as small as possible. This implies that we need experiments that can study both means *and* variances. We are also explicitly considering the possibility that the variance will not be constant, so we will need some form of replication at all design points to allow us to estimate the variances separately.

Study means and variances

Replicated two- and three-series factorials are the basic designs for off-line quality control. From these we can estimate mean responses as usual, and replication allows us to estimate the variance at each factor-level combination as well. There are often ten to fifteen or more factors identified as potentially important. A complete factorial with this many factors would be prohibitively large, so off-line quality control designs are frequently highly-fractioned factorials, but with replication.

Use replicated fractional factorials

Two situations present themselves. In the first situation, the outer noise is at something of a micro scale, meaning that you tend to experience the full

range of outer noise whenever you experiment. One of Taguchi's early suc-
cesses was at the Ina Tile Company, where there was temperature variation in
the kilns. This noise was always present, as tiles in different parts of the kiln
experienced different temperatures. In the second situation, the outer noise is
at a more macro scale, meaning that you tend to experience only part of the
range of outer noise in one experiment. In the exhaust tubing, for example,
temperature and humidity in the factory may affect the machinery, but you
tend not to get hot and cold, dry and humid conditions scattered randomly
among your experimental runs. It is hot and humid in the summer and cold
and dry in the winter.

Is outer noise
micro or macro
scale?

These two situations require different experimental approaches. When
you have outer noise at the micro level, it is generally enough to plan an
experiment using the controllable variables and let the outer noise appear
naturally during replication. When the outer noise is at the macro level, you
must take steps to make sure that the range of outer noise is included in your
experiment. If the outer-noise factors can be controlled under experimental
conditions, then these factors should also be included in the design to ensure
their full range.

Design plan
should include
macro-level outer
noise

Let's return to the exhaust tube problem to make things explicit. Our
controllable factors are tube speed, air pressure, saw speed, and sensor type;
the outer-noise factors are temperature and humidity. Assume for simplicity
that we can choose two levels for all factors, so that there are sixteen combi-
nations for the controllable factors and four combinations for the outer-noise
factors. We need to include the outer-noise factors in our design, because we
are unlikely to see the full range of outer-noise variation if we do not.

There are several possibilities for this experiment. For example, we could
run the full 2^6 design. This gives four "replications" at each combination of
the controllable factors, and these replications span the range of the noise
factors. Or we could run a 2^{6-1} fraction with 32 points. This is smaller
(and possibly quicker and cheaper), but with a smaller sample size we have
less power for detecting effects and only 1 degree of freedom for estimating
variation at each of the sixteen combinations of controllable factors.

18.11.2   Analysis of off-line quality experiments

Analysis is based on the following idea. Some of the controllable factors
affect the variance and the mean, and an additional set of controllable factors
affects only the mean. The factors that affect the variance and mean are
called *design* variables; those that affect only the mean are called *adjustment*
variables. The idea is to use the design variables to minimize the variance,
and then use the adjustment variables to bring the mean on target.

Design variables
affect mean and
variation,
adjustment
variables affect
only mean

This approach is complicated by the fact that mean and variance are often
linked in the usual nonconstant-variance sense that we check with residual
plots and remove using a transformation. If we have this kind of nonconstant
variance, then every variable that affects the mean also affects the variance,
and we will have no adjustment variables. Therefore we need to accom-

Table 18.5: Variance of natural-log sample variances from normal data for 1 through 10 degrees of freedom.

1	2	3	4	5	6	7	8	9	10
4.93	1.64	.93	.64	.49	.39	.33	.28	.25	.22

modate this kind of nonconstant variance before dealing with variation that depends on controllable variables but not directly through the mean.

First, find a transformation of the responses that removes the dependence of variance on mean as much as possible. This is essentially a Box-Cox transformation analysis. On this transformed scale, we hope that there are variables that affect the mean but not the variance.

Transform to "constant" variance

Next, compute the log of the variance of the transformed data at every factor-level combination of the controllable factors. Treat these log variances as responses, and analyze them via ANOVA to see which, if any, controllable factors affect the variance; these are the design variables. Find the factor-level combination that minimizes the variance. For highly-fractioned designs we may only be able to do this by looking at main effects and hoping that there are no interactions. One complication that arises in this step is that once we have log variance as a response, there is no replication. Thus we must use a method for unreplicated factorials to assess whether treatments affect variances.

Analyze log variances to determine design variables

If we can assume that the (transformed) responses that go into each of these variances are independent and normally distributed, then we can calculate an approximate MS_E for the ANOVA with log variances as the responses. Suppose that there are n experimental units at each factor-level combination of the controllable factors; then each of these sample variances has $n - 1$ degrees of freedom. The variance of the (natural) log of a sample variance depends only on the degrees of freedom. Table 18.5 lists the variance of the log of a sample variance for up to 10 degrees of freedom. Note that the variances in that table are *very* sensitive to the normality assumption.

Variance of log sample variance is known for normally distributed data

Finally, return to the original scale. Analyze the response to determine which factors affect the mean response, and find settings for the adjustment variables that put the response on target when the design variables are at their variance-minimizing settings. This step generally makes the assumptions that the adjustment factors can be varied continuously and that the response is linear between the two observed levels of a factor. Please note that adjusting a transformation of y to a target T, say $\sqrt{y}$ to $\sqrt{T}$, will result in a bias on the original scale and thus a deviation from the target.

Put response on target using adjustment variables with design variables set to minimum variance

Example 18.7 Free height of leaf springs

Pignatiello and Ramberg (1985) present a set of data from a quality experiment on the manufacture of leaf springs for trucks. The free height should be as close to 8 inches as possible, with minimum variation. There are four inner noise factors, each at two levels: furnace temperature (B),

Table 18.6: Free height of leaf springs. Data set `LeafSprings`.

B	C	D	E	O low			O high			$\overline{y}$	s^2
−	−	−	−	7.78	7.78	7.81	7.50	7.25	7.12	7.54	.0900
+	−	−	+	8.15	8.18	7.88	7.88	7.88	7.44	7.90	.0707
−	+	−	+	7.50	7.56	7.50	7.50	7.56	7.50	7.52	.0010
+	+	−	−	7.59	7.56	7.75	7.63	7.75	7.56	7.64	.0079
−	−	+	+	7.94	8.00	7.88	7.32	7.44	7.44	7.67	.0908
+	−	+	−	7.69	8.09	8.06	7.56	7.69	7.62	7.79	.0529
−	+	+	−	7.56	7.62	7.44	7.18	7.18	7.25	7.37	.0380
+	+	+	+	7.56	7.81	7.69	7.81	7.50	7.59	7.66	.0173

heating time (C), transfer time (D), and hold-down time (E). There was one outer noise factor: quench oil temperature (O). A 2^{5-1} design with three replications was conducted. We will analyze this as a 2^{4-1} design in the inner noise factors with six replications, because quench-oil temperature is not easily controlled in factory conditions. Table 18.6 shows the results.

We first examine whether the data should be transformed. A plot of log treatment variance against log treatment mean shows no pattern, and Box-Cox does not indicate the need for a transformation, so we use the data on the original scale.

We now do a factorial analysis using log treatment variance as response. (If we had transformed the data, the response would be the log of the variance of the transformed data.) Figure 18.3 shows a half-normal plot of the dispersion effects, that is, the factorial effects with log variance as response. Only factor C appears to affect dispersion, and inspection of Table 18.6 shows that the high level of C has lower variance.

Now examine how the treatments affect average response. Figure 18.4 shows a half-normal plot of the location effects. Here we see that B, C, and the BCD interaction are significant. Recalling the aliasing, the BCD interaction is also the main effect of E. Thus heating time is a design variable that we will set to a high level to keep variance low, and furnace temperature and hold-down time are adjustment variables.

Here are the location effects for these variables (using MacAnova).

```
component: CONSTANT
(1)        7.636
component: b
(1)      -0.11062      0.11063
component: c
(1)       0.088125    -0.088125
component: e
(1)      -0.051875     0.051875
```

We have set C to the high level to get a small variance. To get the mean close to the target of 8, we need B and E to be at their high levels as well; this gives us 7.636 + .111 − .088 + .052, or 7.711, as our estimated response. This is still a little low, so we may need to explore the possibility of expanding the ranges for factors B and E to get the response closer to target.
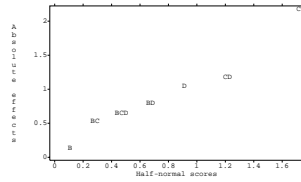
Figure 18.3: Half-normal plot of dispersion effects for leaf spring data, using MacAnova.
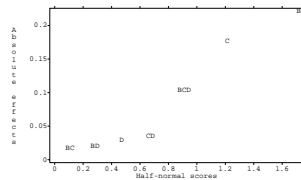
Figure 18.4: Half-normal plot of location effects for leaf spring data, using MacAnova.

18.12   Further Reading and Extensions

Orthogonal-main-effects plans are resolution III designs constructed so that the main effects are orthogonal. Resolution III two- and three-series fraction factorials are orthogonal-main-effects plans, but there are several addi-

tional families of designs that have these properties as well. Plackett-Burman designs (Plackett and Burman 1946) are orthogonal-main-effects plans for $N - 1$ factors at two levels each using N experimental units when N is an integer multiple of 4. When N is a power of 2, these are resolution III fractions of the kind discussed in this chapter. Addelman (1962) constructs orthogonal-main-effects plans for mixed factorials by collapsing factors. For example, start with a 3^{4-2} fraction. Replace factor A by a two level factor E, using the low level of E when A is 0 or 2, and the high level of E when A is 1. This produces a fraction of a $2^1 3^3$ design in nine units. John (1971) discusses these two classes, as well as some other mixed factorial fractions. The aliasing structure of these designs can be quite complex.

Orthogonal arrays are a third class of orthogonal-main-effects plans that are often used in quality experiments. An orthogonal array for k factors in N units is described by an N by k matrix of integers; rows for units, columns for factors, and integers giving factor levels. To be an orthogonal array, all possible pairs of factor levels must occur together an equal number of times for any pair of factors. Standard two- and three-series fractional factorials of resolution III meet this criterion, but so do many additional designs. Hedayat and Wallis (1978) review some of the theory and applications of these arrays.

Fractional factorials can also be run using split-plot and related unit structures. See Miller (1997).

18.13   Problems

Food scientists are trying to determine what chemical compounds make heated butter smell like heated butter. If they could figure that out, then they could make foods that smell like butter without having all the fat of butter. There are eight compounds that they wish to investigate, with each compound at either a high or low level. They use a 2^{8-4} fractional factorial design with I = ABDE = ABCF = -ACDG = -BCDH.

Exercise 18.1

  (a) Find the factor-level combinations used in this design.

  (b) Find the aliases of I and A.

  (c) If A, B, D, E, and AB look big, are there any unresolved ambiguities? If so, which further fraction would you run to resolve the ambiguity?

Consider a 2^{6-2} fractional factorial using I=ABDF = -BCDE.

Exercise 18.2

  (a) Find the aliases of the main effects.

  (b) Find the factor-level combinations used.

  (c) Show how you would block these combinations into two blocks of size eight.

Consider the 2^{8-4} fractional factorial with generator I = BCDE = ACDF = ABCG = ABDH. Find the aliases of C.

Exercise 18.3

Design a 2^{7-2} resolution IV fractional factorial. Give the factor-level combinations used in the principal fraction and show how you would block these combinations into two blocks of size sixteen. | **Exercise 18.4**

Design an experiment. There are eight factors, each at two levels. However, we can only afford 64 experimental units. Furthermore, there is considerable unit to unit variability, so blocking will be required, and the maximum block size possible is 16 units. You may assume that three-way and higher-order interactions are negligible, but two-factor interactions may be present. | **Exercise 18.5**

Find the factor-level combinations used in the principal fraction of a 3^{4-1} with the generator $A^1B^1C^1D^1$. Report the alias structure, and show how you would block the design into blocks of size nine. | **Exercise 18.6**

A 2^{4-1} fractional factorial is created by the aliasing $I = ABD$, and is then blocked into two blocks of size four using $AC = BCD$. Find the factor-level combinations in the two blocks. | **Exercise 18.7**

Consider a 2^{4-1} factorial with I=–ABCD as generator, blocked into two blocks of size four using AB=–CD. | **Exercise 18.8**

(a) Give a skeleton ANOVA for this design.

(b) Say which treatments are assigned to each block.

You ran a 2^{7-3} fractional factorial with aliasing generated by A=EFG, B=DEG, C=DEF. You then do a Daniel plot for the model `y˜A*B*C*D` (A, B, C, D work as a base factorial even though A, B, C, D was not the base factorial used in generating the design). The terms that look big are ABC, ABD, and CD. How would you interpret this result? | **Exercise 18.9**

Briefly describe the experimental design you would choose for each of the following situations, and why. Describe treatments, blocks, etc. | **Problem 18.1**

(a) Many high tech products are extremely expensive to build, so various versions of the product are simulated in a computer before any are physically constructed. These simulations are cheap compared to physical construction, but they are still very time consuming. In this case, we are building turbine blades for jet engines. There are 14 factors that we need to vary (two levels each), and we have resources on the supercomputer to simulate up to 16 factor/level combinations.

(b) Graphene is a nano material composed of a sheet of carbon one atom thick with atoms arranged in a hexagonal lattice. It is very light and very strong, and people seemingly find new applications for it every day. One recent discovery about graphene is that single protons can pass through a sheet of graphene, but apparently not electrons; this makes it a candidate for use as an ultra thin membrane in fuel cells. Our experiment seeks to test graphene in the fuel cell application.

The factors of interest are temperature, hydrogen fuel purity, and catalyst. Temperature will be set at 60, 70, or 80 degrees C; purity will be set at 50%, 60%, or 70%; catalyst will be type A or B. We assemble a test

fuel cell with a catalyst and a membrane and then test it at some temperature with some fuel. A single membrane should only be used with one combination of purity and catalyst, but it can be used for all three temperatures. Removing a membrane from the test fuel cell (as you would need to do to change the catalyst) will destroy the membrane. We have 18 membranes available for test.

(c) An aquapod is a small robot that can "swim" through water; aquapods are intended for environmental monitoring. We want to study four treatments, which are the combinations of two factors, each at two levels. The factors are the size of the flippers (large or small) and whether the flippers are used in phase or out of phase (sort of like butterfly versus freestyle swimming). The robot is battery powered, and the battery must be recharged between each run. Given the length of the test run and the recharge time, we can only do three runs per day. The main response of interest is how far the robot can swim before it exhausts its battery.

We have secured the use of the test tank for four consecutive Saturdays. The test tank is adjacent to a river and uses river water as its filler (in fact, the water in the tank is constantly being exchanged for fresh water from the river). One concern we have is that the water could be different temperatures on the different days, and we know that water temperature will affect battery performance.

(d) Motor oil in an automotive engine will change viscosity over time. We wish to study if the viscosity change depends the oil spending time at an elevated temperature. Our basic idea is to take a sample of oil, measure its viscosity, heat it to one of four temperatures and hold it at that temperature for 72 hours, and then measure the viscosity again. The response is the change in viscosity.

We suspect that there will be brand-of-oil differences in viscosity change, with the name brand oils expected to change viscosity less; we will have to deal with this, but this is not of interest. We also expect that there will be differences in viscosity change based on the original viscosity of the oil; this is also not of interest. We have available oil of four different label viscosities (5W30, 10W30, 5W40, and 10W40) from eight different brands (three are name brands and five are private store labels). We have the capacity to run 32 viscosity-drop trials.

(e) My wife likes to garden, and we have created four raised garden beds, each four feet by four feet in size. My wife likes tomatoes, so we constantly strive to get higher tomato production. We can squeeze four tomato plants into one raised bed, but certainly no more. This year we are going to find the way to get the most tomato yield by weight. We can vary the variety of tomato (Big Boy vs German Stripe), the brand of fertilizer (Scott's vs store brand), and planting schedule (early vs late). Unfortunately, we had access to different kinds of soil when we built our four raised beds, so the soil conditions in the four raised beds are very different.

Design an experiment to help us determine the combination of factors that will give us good tomato yield.

(f) Functional magnetic resonance imaging (fMRI) is a technique that allows us to monitor brain activity (technically, monitor blood oxygenation levels in the brain) during mental tasks. There are six different "tuning" parameters that I can set in the procedure that will affect the quality of results. For simplicity, assume that each parameter has only two levels. The magnet is very busy and very expensive to use, so I can afford only 16 runs to use to select my tuning parameters. How should I design my experiment to get the information I need from only 16 runs?

(g) When iron oxide nanoparticles are placed in an alternating magnetic field they produce heat. These particles offer a potential method for uniformly warming frozen tissues, including large tissues such as organs frozen for transport to an organ recipient. We need to study the concentration of nanoparticles that should be used to infuse an organ sample before it is frozen, and the frequency of the alternating magnetic field that should be used when it is thawed. In order to understand how the infusion of nanoparticles will work, we need to attempt to infuse whole organs. However, to understand the frequency of the magnetic field during thawing, we can use portions of the organ.

We want to consider three different infusion concentrations and two different field frequencies. We have twelve swine livers available for study.

Describe an appropriate experimental design for this situation.

(h) We are concerned about agricultural chemicals and their effects on amphibian growth, specifically frogs. We have eight 290-gallon cattle tanks in which we construct artificial ponds. We add tadpoles and other native life to all eight tanks. We want to study the effects of atrazine (absent or present), phosphate fertilizer (low or high level), glyphosate (absent or present), and organochlorides (absent or present) on the health of the frogs after six weeks of growth.

(i) A high blood concentration of homocysteine is associated with increased risk of cardiovascular disease. We wish to study the effect of three treatments on the homocystein concentration (control, 870 mg per day caffeine, or filtered coffee containing 870 mg caffeine per day). Forty-eight subjects have agreed to participate, and we expect large subject to subject variation in their levels of homocysteine. Each subject should be on a treatment for two weeks to get a stable blood concentration of homocysteine. Subjects participate for twelve weeks.

(j) Preliminary studies indicate that the anti-oxidants vitamin C and vitamin E may help prevent cancer, specifically, prostate cancer. We have a volunteer group of more than 10,000 male physicians in their early 50s with no known individual risk factors who can be given either, neither, or both of the vitamins blindly. They will be followed for a minimum of 10 years to find the number who contract prostate cancer.

(k) Asbestos fiber concentrations in air are measured by drawing a fixed volume of air through a disk-shaped filter, taking a wedge of the filter (generally 1/4 of the filter), preparing it for microscopic analysis, and then counting the number of asbestos fibers found on the prepared wedge when looking through an optical microscope. (Actually, we only count on a random subsample of the area of the prepared wedge, but for the purposes of the question, consider the wedge counted.) We wish to compare four methods of preparing the wedges for their effects on the subsequent fiber counts. We have available 24 filters from a broad range of asbestos air concentrations; we can use each filter entirely, so that we can get four wedges from each filter. We can also use four trained microscopists. Despite the training, we anticipate considerable microscopist to microscopist variation in the counts (that is, some tend to count high, and some tend to count low).

(l) A food scientist wishes to study the effect that eating a given food will have on the ratings given to a similar food (sensory-specific satiety). There is a pool of 24 volunteers to work with. Each volunteer must eat a "load food" (a large portion of hamburger or potato), and then eat and rate two "test foods" (small portions of roast beef and rice). After eating, the volunteer will rate the appeal of the roast and rice.

(m) Scientists studying the formation of tropospheric ozone believe that five factors might be important: amount of hydrocarbon present, amount of NO_X present, humidity, temperature, and level of ultraviolet light. They propose to set up a "model atmosphere" with the appropriate ingredients, "let it cook" for 6 hours, and then measure the ozone produced. They only have funding sufficient for sixteen experimental units, and their ozone-measuring device can only be used eight times before it needs to be cleaned and recalibrated.

(n) A school wishes to evaluate four reading texts for use in the sixth grade. One of the factors in the evaluation is a student rating of the stories in the texts. The principal of the school decides to use four sixth-grade rooms in the study, and she expects large room to room differences in ratings. Due to the length of the reading texts and the organization of the school year into trimesters, each room can evaluate three texts. The faculty do not expect systematic differences in ratings between the trimesters.

(o) The sensory quality of prepared frozen pizza can vary dramatically. Before the quality control department begins remedial action to reduce the variability, they first attempt to learn where the variability arises. Three broad sources are production (variation in quality from batch to batch at the factory), transportation (freeze/thaw cycles degrade the product, and our five shipping/warehouse companies might not keep the product fully frozen), and stores (grocery store display freezers may not keep the product frozen). Design an experiment to estimate the various sources of variability from measurements made on pizzas taken from grocery freezers. All batches of pizza are shipped by all shipping companies, but each grocery store is served by only one shipping company. You should buy no more than 500 pizzas.

(p) Food scientists are trying to figure out what makes cheddar cheese smell like cheddar cheese. To this end, they have been able to identify fifteen compounds in the "odor" of the cheese, and they wish to make a preliminary screen of these compounds to see if consumers identify any of these compounds or combinations of compounds as "cheddary." At this preliminary stage, the scientists are willing to ignore interactions. They can construct test samples in which the compounds are present or absent in any combination. They have resources to test sixteen consumers, each of whom should sample at most sixteen combinations.

(q) The time until germination for seeds can be affected by several variables. In our current experiment, a batch of seeds is pretreated with one of three chemicals and stored for one of three time periods in one of two container types. After storage time is complete, the average time to germination is measured for the batch. We have 54 essentially uniform batches of seeds, and wish to understand the relationships between the chemicals, storage times, and storage containers.

(r) Our company is creating a biodegradable polymer coating that includes nano-scale structures. There are 12 process factors that we can vary, and we would like to know which, if any, of these 12 factors affect the total mass of the polymer that gets applied to a surface. Our boss will allow us 16 experimental runs. Design an experiment to screen these 12 factors in 16 experimental units.

(s) The U.S. Department of Transportation needs to compare five new types of pavement for durability. They do this by selecting "stretches" of highway, installing an experimental pavement in the stretch, and then measuring the condition of the stretch after 3 years. There are resources allocated for 25 stretches of highway. From past experience, the department knows that traffic level and weather patterns affect the durability of pavement. The department is organized into five regional districts, and within each district the weather patterns are reasonably uniform. Also within each district are highways from each of the five traffic level groups.

(t) We are designing a nasal spray for the delivery of a drug (HU). Two of the issues in how well the system will work for drug delivery are aerosol particle size and absorbability. Other compounds are added to the mixture to help adjust those responses. This experiment will study the aerosol particle size as a response, and we will vary the concentrations of five compounds in the solution. The five compounds are the drug HU, two polymers (HEC and PEU) and two salts ($CaCl_2$ and NaCl). We can set the concentrations of the salts at 0%, 15%, or 30%, and we can set the concentrations of the other factors at 0%, 2%, or 4%. We need to be able to fit a quadratic model to describe how particle size varies as a function of the five factors.

a. Design an experiment using n=50 observations.

b. Can you design an experiment to fit this model using only n=40 observations? If so, how?

Briefly describe the experimental design used in each of the follow- **Problem 18.2**
ing situations (list units, blocks, covariates, factors, whole/split plots, and
so forth). Give a skeleton ANOVA (sources and degrees of freedom only).

(a) Biosolids are nutrient rich, organic by-products of the sewage treatment
process with pathogens removed. Biosolids have been used as soil treat-
ments for many years, but the current experiment explores the use of
biosolids for reducing the bioavailability of lead in soils. (Bioavailable
lead is essentially the amount of lead that can be absorbed or metabolized
by an organism, rather than simply the total amount of lead.) Biosolids
can be modified by adding lime (or not), by adding iron (or not), and by
adding phosphorus (or not).

We have eight plots of residential soil; four are in a suburban area, and the
other four are in an urban, industrialized area. Each plot will be treated
with one of the eight combinations of biosolids (eight combinations of
lime, iron, and phosphorus). The biosolids will be tilled into the soils,
and the plots will then be planted with grass. After one year, we will
sample each plot and measure the bioavailability of lead. We randomize
the four treatments control, iron and lime, iron and phosphorus, lime and
phosphorus to the urban sites; we randomize the four treatments iron,
lime, phosphorus, and (iron, lime, and phosphorus) to the suburban sites.

(b) We wish to study the effect of maternal condition on the survival of off-
spring in deer in the wild. Two factors felt to contributed to maternal
condition are food availability and winter severity (as measured by av-
erage snow depth). Twenty winter yards (areas where deer congregate)
are found, and ten yards are selected at random to receive food supple-
mentation via corn feeders that will be refilled once a week. In addition,
snow depth is measured each week at all yards so that an average snow
depth can be computed. After the fawns are born, one fawn at each yard
is caught and fit with a radio collar. The radio collar will change its pulse
rate if the fawn dies (detected by temperature change), so we can monitor
survival by listening for the pulses.

(c) We wish to compare four types of running shoes. We have 100 high
school boys to use as subjects. We want them to run for 800 meters, and
their time will be the response. We randomly assign the shoes to the boys,
25 to each type of shoe. We also expect general cardio-vascular health
to be associated with time, and we measure heart rate of each individual
when they finish the run in addition to just their time.

(d) Snellingdale Mall uses a lot of cut Christmas trees (real, not artificial)
as decoration during the Christmas season. These trees are placed in
clusters of three at ten locations around the mall. One important issue
to them is how tree species affects how long the trees will retain their
needles. This year they run an experiment. They get six each of Frasier
Fir, Balsam Fir, Scotch Pine, White Pine, and Blue Spruce; all trees
are the same size. The trees are then randomly spread around the mall
subject to the restriction that each combination of three species occurs at

one of the decoration locations. The response measured is how long each tree lasts before it begins to drop an unacceptable number of needles.

(e) Hazel nuts contain phenolic compounds that have food preserving properties (they are antioxidants). We are interested in whether nuts from four different varieties contain different amounts of a particular phenol. The following experiment is done on three separate days. On each of the three days, all four varieties are analyzed in a random order. Each day, a batch of nuts from a variety is ground to a fine power. Five grams of the powder are stirred into a solvent and allowed to soak for an hour. After the wait, the solvent is run through a separation column, and the extract between 20 and 21 minutes into the separation is collected. This extract is then analyzed with magnetic resonance spectroscopy to determine the amount of the phenol of interest present in the extract; this amount is the response for a given variety on a given day.

(f) Perfume is expensive, and we'd like it to retain its odor. A newcomer to Minnesota remarked that perfume seems to lose its odor faster in the cold weather than in warm weather, so we explore this in an experiment. We will have 36 identical pieces of cotton cloth, and each piece will be treated with the same amount of perfume. The pieces of cloth are then assigned to four temperature treatments (nine for each treatment). The treatments are two hours at 0°, 35°, 70°, or 95° (freezer, refrigerator, room, low oven). After the treatment, the pieces are allowed 5 minutes to come to room temperature; then the clothes are sniffed by judges and given a strength rating from 1 to 10. There are 18 judges, and each judge sniffs 2 pieces of cloth (in random order), with each pair of temperatures sniffed by three judges.

(g) We wish to study the effects of stress and activity on the production of a hormone present in the saliva of children. The high-stress treatment is participation in a play group containing children with whom the subject child in unacquainted; the low-stress treatment is participation in a play group with other children already known to the subject child. The activities are a group activity, where all children play together, and an individual activity, where each child plays separately. Thirty-two children are split at random into two groups of sixteen. The first group is assigned to high stress, the other to low stress. For each child the order of group or individual activity is randomized, and a saliva sample is taken during each activity.

(h) Neighbors near the municipal incinerator are concerned about mercury emitted in stack gasses. They want a measure of the accumulation rate of mercury in soil at various distances and directions from the incinerator. They collect a bunch of soil, mix it as well as they can, divide it into 30 buckets, and have a lab measure the mercury concentration in each bucket. The buckets are then randomly divided into fifteen sets of two; the pairs are placed in fifteen locations around the incinerator, left for 2 years, and then analyzed again for mercury. The response is the increase in mercury. The lab informed the activists that the amount of increase

will be related to the amount of carbon in the soil, because mercury is held in the organic fraction; so they also take a carbon measurement.

(i) We wish to discover the effects of food availability on the reproductive success of anole lizards as measured by the number of new adults appearing after the breeding season. There are twelve very small islands with anole populations available for the study. The islands are man-made and more or less equally spaced along a north-south line. The treatments will be manipulation of the food supply on the islands during peak breeding season. There are three treatments: control (leave natural), add supplemental food, and reduced food (set out traps to deplete the population of insects the anoles eat). One potential source of variation is that the lizards are eaten by birds, and there is a wildlife refuge with a large bird population near the northern extreme of the study area. To control for this, we group the islands into the northern three, the next three, and so on, and randomize the treatments within these groups.

(j) A fast-food restaurant offers both smoking and non-smoking sections for its customers. However, there is considerable smoke "leakage" from the smoking section to the non-smoking section. The manager wants to minimize this leakage by finding a good division of the restaurant into the two sections. She has three possible divisions of the tables, and conducts an experiment by assigning divisions at random to days for 3 weeks (7 days per division) and surveying non-smoking patrons about the amount of smoke. In addition, she monitors the number of smokers per day, as that has an obvious effect on the amount of leakage.

Problem 18.3

Avocado oil may be extracted from avocado paste using the following steps: (1) dilute the paste with water, (2) adjust the pH of the paste, (3) heat the paste at 98°C for 5 minutes, (4) let the paste settle, (5) centrifuge the paste. We may vary the dilution rate (3:1 water or 5:1 water), pH (4.0 or 5.5), settling (9 days at 23°C or 4 days at 37°C), and centrifugation (6000g or 12000g). Briefly describe experimental designs for each of the following situations. You may assume that the paste (prior to any of the five steps mentioned) may be used any time up to a week after its preparation. You may also assume that the primary cost is the processing; the cost of the paste is trivial.

(a) We wish to study effects of the four factors mentioned on the extraction efficiency. Avocado paste is rather uniform, and we have enough money for 48 experimental units.

(b) We wish to study effects of the four factors mentioned on the extraction efficiency. Avocado paste is not uniform but varies from individual fruit to fruit. Each fruit produces enough paste for about 20 experimental units, and we have enough money for 48 experimental units.

(c) We wish to study effects of the four factors mentioned on the extraction efficiency. Avocado paste is not uniform but varies from individual fruit to fruit. Each fruit produces enough paste for about 10 experimental units, and we have enough money for 48 experimental units.

(d) We wish to determine the effects of the pH, settling, and centrifugation treatments on the concentration of α-tocopherol (vitamin E) in the oil. Each fruit produces enough paste for about six experimental units, and we have enough money for 32 experimental units. Furthermore, we can only use four experimental units per day and the instruments need to be recalibrated each day.

Problem 18.4

Here are the factor/level combinations used in a fractional factorial: def, af, be, abd, cd, ace, bcf, abcdef. I assert that this fraction was formed using the generators D = AB, E = ABC, and F = AC. Am I correct or not? Explain your answer.

Problem 18.5

An experiment was conducted to determine the factors that affect the amount of shrinkage in speedometer cable casings. There were fifteen factors, each at two levels, but the design used only sixteen factor-level combinations (2_{III}^{15-11}). The generators were I = –DHM = –BHK = BDF = BDHO = –AHJ = –ADE = ADHN = –ABC = ABHL = ABDG = –ABDHP, and the factors were: liner OD (A); liner die (B); liner material (C); liner line speed (D); wire braid type (E); braiding tension (F); wire diameter (G); liner tension (H); liner temperature (J); coating material (K); coating die type (L); melt temperature (M); screen pack (N); cooling method (O); and line speed (P). The response is the average of four shrinkage measurements (data from Quinlan 1985, data set `Shrinkage`).

A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	y
−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	.4850
−	−	−	−	−	+	−	+	+	+	+	+	+	+	+	.5750
−	−	−	+	+	−	+	−	−	−	−	+	+	+	+	.0875
−	−	−	+	+	−	+	+	+	+	+	−	−	−	−	.1750
−	+	+	−	−	−	+	−	−	+	+	−	−	+	+	.1950
−	+	+	−	−	−	+	+	+	−	−	+	+	−	−	.1450
−	+	+	+	+	+	−	−	−	+	+	+	+	−	−	.2250
−	+	+	+	+	+	−	+	+	−	−	−	−	+	+	.1750
+	−	+	−	+	+	+	−	+	−	+	−	+	−	+	.1250
+	−	+	−	+	+	+	+	−	+	−	+	−	+	−	.1200
+	−	+	+	−	−	−	−	+	−	+	+	−	+	−	.4550
+	−	+	+	−	−	−	+	−	+	−	−	+	−	+	.5350
+	+	−	−	+	−	−	−	+	+	−	−	+	+	−	.1700
+	+	−	−	+	−	−	+	−	−	+	+	−	−	+	.2750
+	+	−	+	−	+	+	−	+	+	−	+	−	−	+	.3425
+	+	−	+	−	+	+	+	−	−	+	−	+	+	−	.5825

Analyze these data to determine which factors affect shrinkage, and how they affect shrinkage.

Problem 18.6

Seven factors are believed to control the softness of cold-foamed car seats, and an experiment was conducted to determine how these factors influence the softness. A 2_{III}^{7-4} design was run with generators I = ABD = ACE = BCF = ABCG. The response is the average softness of the seats (data from Bergman and Hynén 1997, data set `SeatSoftness`).

A	B	C	D	E	F	G	y
−	−	−	+	+	−	−	25.3
+	−	−	−	−	+	+	20.6
−	+	−	−	+	−	+	26.7
+	+	−	+	−	+	−	23.8
−	−	+	+	−	−	+	23.5
+	−	+	−	+	+	−	24.0
−	+	+	−	−	−	−	23.5
+	+	+	+	+	+	+	24.2

Analyze these data to determine how the factors affect softness.

Problem 18.7

Silicon wafers for integrated circuits are grown in a device called a susceptor, and a response of interest is the thickness of the silicon. Eight factors, each at two levels, were believed to contribute: rotation method (A), wafer code (B), deposition temperature (C), deposition time (D), arsenic flow rate (E), HCl etch temperature (F), HCl flow rate (G), and nozzle position (H). A 2^{8-4}_{IV} design was run with generators I = ABCD = BCEF = ACEG = BCEH. The average thickness of the silicon follows (data from Shoemaker, Tsui, and Wu 1991, data set `SiliconThickness`).

A	B	C	D	E	F	G	H	y
−	−	−	−	−	−	−	−	14.80
−	−	−	−	+	+	+	+	14.86
−	−	+	+	−	+	+	+	14.00
−	−	+	+	+	−	−	−	13.91
−	+	−	+	−	+	−	+	14.14
−	+	−	+	+	−	+	−	13.80
−	+	+	−	−	−	+	−	14.73
−	+	+	−	+	+	−	+	14.89
+	−	−	+	−	−	+	−	13.93
+	−	−	+	+	+	−	+	14.09
+	−	+	−	−	+	−	+	14.79
+	−	+	−	+	−	+	−	14.33
+	+	−	−	−	+	+	+	14.77
+	+	−	−	+	−	−	−	14.88
+	+	+	+	−	−	−	−	13.76
+	+	+	+	+	+	+	+	13.97

Analyze these data to determine how silicon thickness depends on the factors.

Problem 18.8

The responses shown in Problem 18.6 are the averages of sixteen individual units. The variances among those units were: 3.24, .64, 1.00, 2.56, 1.96, 1.00, 1.00, and 2.56 for the eight factor-level combinations used in the design. Which factor-levels should we use to reduce variation?

Problem 18.9

We have a replicated 2^3 design with data (in standard order, first replicate then second replicate) 6, 10, 32, 60, 4, 15, 26, 60, 8, 12, 34, 60, 16, 5, 37, 52. We would like the mean response to be about 30, with minimum variability. How should we choose our factor levels?

A product is produced that should have a score as close to 2 as possible. Eight factors are believed to influence the score, and a completely randomized experiment is conducted using 64 units and sixteen treatments in a 2^{8-4}_{IV} fractional-factorial treatment structure. Analyze these data and report how you would achieve the score of 2. You may assume that the treatments are continuous and can take any level between -1 (low) and 1 (high). Increasing any factor costs more money, and factors are named in order of increasing expense (data set `ProductScore`).

Problem 18.10

Treatment		Score		
(1)	2.50	2.85	2.80	2.92
$aefg$	1.83	1.87	1.87	1.70
$befh$	1.55	1.56	1.64	1.56
$abgh$	1.12	1.14	1.23	1.18
$cegh$	1.67	1.65	1.83	1.89
$acfh$	2.79	2.75	2.95	3.18
$bcfg$	1.15	1.19	1.18	1.16
$abce$	1.55	1.52	1.62	1.66
$dfgh$	2.95	4.05	2.73	2.13
$adeh$	9.41	4.37	5.06	4.20
$bdeg$	1.38	1.88	2.05	1.54
$abdf$	2.14	2.79	2.65	1.85
$cdef$	7.48	5.79	3.55	13.63
$acdg$	3.13	1.98	2.24	3.14
$bcdh$	2.48	1.87	2.92	2.21
$abcdefgh$	2.00	1.42	1.36	1.23

We have run a 2^{6-3}_{III} fractional factorial with generators ABCD, ACE, and BCF.

Problem 18.11

(a) List all of the aliases of I.

(b) In the Daniel plot, the main effects of B, D, and E looked large. What can you conclude?

(c) Your colleague claims that the following factor/level combinations were run in the experiment: ace, cd, bcf, abcdef, be, def, af, abd. Is he correct? Explain why or why not.

Bacteriocin is a food preservative that can be extracted from some bacterial cultures. We have five factors, each at two levels (Glucose, Inoculum size, Aeration, Temperature, Sodium, as factors A through E). We run a 2^{5-2}_{III} fraction with D=AC and E=BC. The factor level combinations are de, ae, bd, ab,c,acd, bce, abcde.

Problem 18.12

If factors A, C, and D appear to be large, which additional fraction should you run to break the ambiguity?

You and your coworkers Jim and Joe have been asked to review an experiment that was run 10 years ago. It is clearly a 2^{6-3} fractional factorial with the following factor/level combinations: abd, af, bcf, abcdef, cd,

Problem 18.13

`be, ace, def`. Jim says that the aliasing structure was generated by I = ABD = ACE = BCF. Joe says that the aliasing structure was generated by I = ABCD = BCE = ABF. Which, if either, of these two is correct, and why?

Suppose you have seven factors to study, each at two levels, but that you can only afford 32 runs. Further assume that at most four of the factors are active, and the rest inert. You may safely assume that all three-factor or higher-order interactions are negligible, but many or all of the two-factor interactions in the active factors are present.

Problem 18.14

(a) Design a single-stage experiment that uses all 32 runs. Show that this experiment may not be able to estimate all effects of interest.

(b) Design a two-stage experiment, where you use 16 runs in the first stage, and then use an additional 16 runs if needed. Show that you can always estimate the effects of interest with the two-stage design.

(c) Suppose that we had assigned the seven labels A, B, C, D, E, F, and G to the seven factors at random. There are 35 (seven choose four) ways of assigning the four active factors to labels, ignoring the order. What is the probability that you can estimate main effects and all two-factor interactions in the active factors with your design from part (a)? What is the probability that you can estimate main effects and all two factor interactions in the active factors with your first 16-point design from (b) and your full two-stage design from part (b)?

(d) What is the main lesson you draw from (a), (b), and (c)?

We wish to determine the tolerance of icings to ingredient changes and variation in the preparation. Ingredient changes are represented by factors C, D, E, F, G, and H. All are at two levels. C and D are two types of sugars; E, F, and G are three stabilizers; and H is a setting agent. The levels of these factors represent changes in the amounts of these constituents in the mix. Variation in preparation is modeled as the amount of water added to the product. This has four levels and is represented as the combinations of factors A and B. The response we measure is (coded) viscosity of the icing. A quarter-fraction with 64 observations was run; data follow (Carroll and Dykstra 1958) (data set `Icings`):

Problem 18.15

(1)	26	agh	6	bh	43	abg	-3
cg	16	ach	10	$bcgh$	69	abc	-5
dgh	12	ad	13	bdg	45	$abdh$	-13
cdh	22	$acdg$	17	bcd	45	$abcdgh$	-4
eh	29	aeg	13	be	54	$abegh$	4
$cegh$	30	ace	17	$bceg$	54	$abceh$	5
deg	29	$adeh$	16	$bdegh$	43	$abde$	-2
cde	34	$acdegh$	16	$bcdeh$	67	$abcdeg$	-3
fgh	32	af	19	bfg	64	$abfh$	6
cfh	30	$acfg$	18	bcf	57	$abcfgh$	6
df	27	$adfgh$	29	$bdfh$	50	$abdfg$	6
$cdfg$	35	$acdfh$	22	$bcdfgh$	53	$abcdf$	7
efg	53	$aefh$	29	$befgh$	74	$abef$	8
cef	46	$acefgh$	21	$bcefh$	73	$abcefg$	13
$defh$	35	$adefg$	23	$bdef$	69	$abdefgh$	20
$cdefgh$	42	$acdef$	27	$bcdefg$	69	$abcdefh$	10

Determine which factors affect the viscosity of the icing, and in what ways. The response should lie between 25 and 30; what does the experiment tell us about the icing's tolerance to changes in ingredients?

Question 18.1 Use the fact that the shortest alias of I in a resolution R design has R letters to show that a 2^{k-p} design of resolution R contains a complete factorial in any $R-1$ factors.

Question 18.2 Show that fold-over breaks all aliases of odd length.

Question 18.3 Show that (1) there are $1 + 3 + 3^2 + \cdots + 3^{k-1}$ two-degree-of-freedom splits in a 3^k factorial; (2) there are $1 + 3 + 3^2 + \cdots + 3^{k-q-1}$ two-degree-of-freedom splits in a 3^{k-q} fractional factorial, each with 3^q labels; and (3) there are $1 + 3 + \cdots + 3^{q-1}$ two-degree-of-freedom splits aliased to I in a 3^{k-q} fractional factorial.

Chapter 19

Response Surface Designs

Many experiments have the goals of describing how the response varies as a function of the treatments and determining treatments that give optimal responses, perhaps maxima or minima. Factorial-treatment structures can be used for these kinds of experiments, but when treatment factors can be varied across a continuous range of values, other treatment designs may be more efficient. *Response surface methods* are designs and models for working with continuous treatments when finding optima or describing the response is the goal.

Response surface methods

19.1   Visualizing the Response

In some experiments, the treatment factors can vary continuously. When we bake a cake, we bake for a certain time x_1 at a certain temperature x_2; time and temperature can vary continuously. We could, in principle, bake cakes for any time and temperature combination. Assuming that all the cake batters are the same, the quality of the cakes y will depend on the time and temperature of baking. We express this as

Response is a function of continuous design variables

$$y_{ij} = f(x_{1i}, x_{2i}) + \epsilon_{ij} \ \ ,$$

meaning that the response y is some function f of the design variables x_1 and x_2, plus experimental error. Here j indexes the replication at the ith unique set of design variables.

One common goal when working with response surface data is to find the settings for the design variables that optimize (maximize or minimize) the response. Often there are complications. For example, there may be several responses, and we must seek some kind of compromise optimum that makes all responses good but does not exactly optimize any single response. Alternatively, there may be constraints on the design variables, so that the goal is to optimize a response, subject to the design variables meeting some constraints.

Compromise or constrained optimum
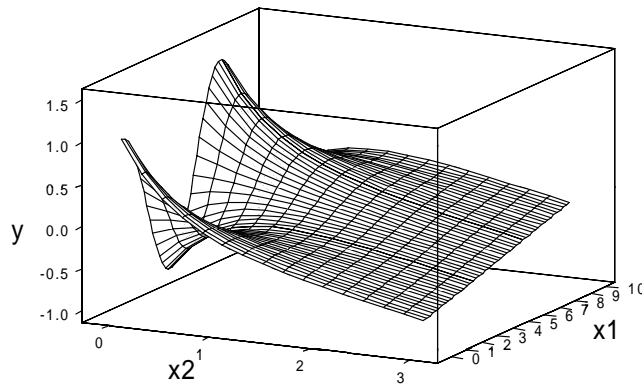
Figure 19.1: Sample perspective plot, using Minitab.

A second goal for response surfaces is to understand "the lie of the land." Where are the hills, valleys, ridge lines, and so on that make up the topography of the response surface? At any give design point, how will the response change if we alter the design variables in a given direction?

Describe the shape of the response

We can visualize the function f as a surface of heights over the x_1, x_2 plane, like a relief map showing mountains and valleys. A perspective plot shows the surface when viewed from the side; Figure 19.1 is a perspective plot of a fairly complicated surface that is wiggly for low values of x_2, and flat for higher values of x_2. A contour plot shows the contours of the surface, that is, curves of x_1, x_2 pairs that have the same response value. Figure 19.2 is a contour plot for the same surface as Figure 19.1.

Perspective plots and contour plots

Graphics and visualization techniques are some of our best tools for understanding response surfaces. Unfortunately, response surfaces are difficult to visualize when there are three design variables, and become almost impossible for more than three. We thus work with models for the response function f.

Use models for f

19.2   First-Order Models

All models are wrong; some models are useful. *George Box*

We often don't know anything about the shape or form of the function f, so any mathematical model that we assume for f is surely wrong. On the other hand, experience has shown that simple models using low-order polynomial terms in the design variables are generally sufficient to describe sections of

Polynomials are often adequate models
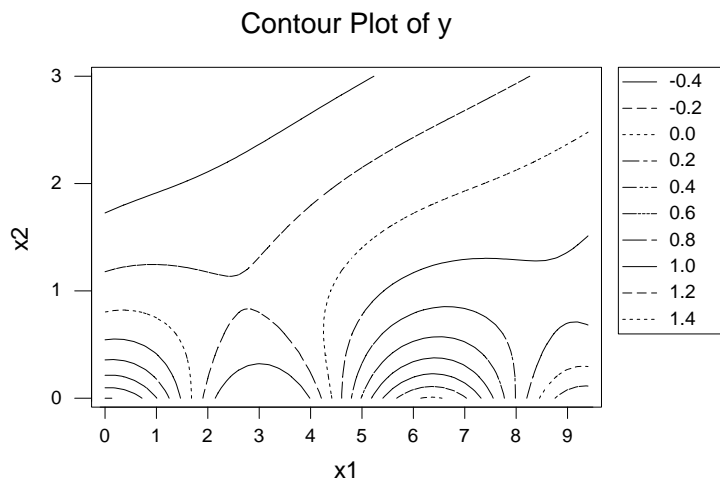
Contour Plot of y

Figure 19.2: Sample contour plot, using Minitab.

a response surface. In other words, we know that the polynomial models described below are almost surely incorrect, in the sense that the response surface f is unlikely to be a true polynomial; but in a "small" region, polynomial models are usually a close enough approximation to the response surface that we can make useful inferences using polynomial models.

We will consider *first-order models* and *second-order models* for response surfaces. A first-order model with q variables takes the form

First-order model has linear terms

$$
\begin{aligned}
y_{ij} &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_q x_{qi} + \epsilon_{ij} \\
&= \beta_0 + \sum_{k=1}^{q} \beta_k x_{ki} + \epsilon_{ij} \\
&= \beta_0 + \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_{ij} \ ,
\end{aligned}
$$

where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{qi})'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$. The first-order model is an ordinary multiple-regression model, with design variables as predictors and β_k's as regression coefficients.

First-order models describe inclined planes: flat surfaces, possibly tilted. These models are appropriate for describing portions of a response surface that are separated from maxima, minima, ridge lines, and other strongly curved regions. For example, the side slopes of a hill might be reasonably approximated as inclined planes. These approximations are local, in the sense that you need different inclined planes to describe different parts of the mountain. First-order models can approximate f reasonably well as long as the region of approximation is not too big and f is not too curved in that region. A first-order model would be a reasonable approximation for the part

First-order models describe flat, but tilted, surfaces

of the surface in Figures 19.1 or 19.2 where x_2 is large; a first-order model would work poorly where x_2 is small.

Bearing in mind that these models are only approximations to the true response, what can these models tell us about the surface? First-order models can tell us which way is up (or down). Suppose that we are at the design variables $\mathbf{x}$, and we want to know in which direction to move to increase the response the most. This is the direction of *steepest ascent*. It turns out that we should take a step proportional to $\boldsymbol{\beta}$, so that our new design variables are $\mathbf{x} + r\boldsymbol{\beta}$, for some $r > 0$. If we want the direction of *steepest descent,* then we move to $\mathbf{x} - r\boldsymbol{\beta}$, for some $r > 0$. Note that this direction of steepest ascent is only approximately correct, even in the region where we have fit the first-order model. As we move outside that region, the surface may change and a new direction may be needed.

Contours or level curves are sets of design variables that have the same expected response. For a first-order surface, design points $\mathbf{x}$ and $\mathbf{x} + \boldsymbol{\delta}$ are on the same contour if $\sum \beta_k \delta_k = 0$. First-order model contours are straight lines for $q = 2$, planes for $q = 3$, and so on. Note that directions of steepest ascent are perpendicular to contours.

19.3   First-Order Designs

We have three basic needs from a response surface design. First, we must be able to estimate the parameters of the model. Second, we must be able to estimate *pure error* and *lack of fit*. As described below, pure error and lack of fit are our tools for determining if the first-order model is an adequate approximation to the true mean structure of the data. And third, we need the design to be efficient, both from a variance of estimation point of view and a use of resources point of view.

The concept of pure error needs a little explanation. Data might not fit a model because of random error (the ϵ_{ij} sort of error); this is pure error. Data also might not fit a model because the model is misspecified and does not truly describe the mean structure; this is lack of fit. Our models are approximations, so we need to know when the lack of fit becomes large relative to pure error. This is particularly true for first-order models, which we will then replace with second-order models. It is also true for second-order models, though we are more likely to reduce our region of modeling rather than move to higher orders.

We do not have lack of fit for factorial models when the full factorial model is fit. In that situation, we have fit a degree of freedom for every factor-level combination—in effect, a mean for each combination. There can be no lack of fit in that case because all means have been fit exactly. We can get lack of fit when our models contain fewer degrees of freedom than the number of distinct design points used; in particular, first- and second-order models may not fit the data.

Response surface designs are usually given in terms of *coded variables*. Coding simply means that the design variables are rescaled so that 0 is in

the center of the design, and ± 1 are reasonable steps up and down from the center. For example, if cake baking time should be about 35 minutes, give or take a couple of minutes, we might rescale time by $(x_1 - 35)/2$, so that 33 minutes is a –1, 35 minutes is a 0, and 37 minutes is a 1.

First-order designs collect data to fit first-order models. The standard first-order design is a 2^q factorial with *center points*. The (coded) low and high values for each variable are ± 1; the center points are m observations taken with all variables at 0. This design has $2^q + m$ points. We may also use any 2^{q-k} fraction with resolution III or greater.

The replicated center points serve two uses. First, the variation among the responses at the center point provides an estimate of pure error. Second, the contrast between the mean of the center points and the mean of the factorial points provides a test for lack of fit. When the data follow a first-order model, this contrast has expected value zero; when the data follow a second-order model, this contrast has an expectation that depends on the pure quadratic terms.

Example 19.1 Cake baking

Our cake mix recommends 35 minutes at 350°, but we are going to try to find a time and temperature that suit our palate better. We begin with a first-order design in baking time and temperature, so we use a 2^2 factorial with three center points. Use the coded values –1, 0, 1 for 33, 35, and 37 minutes for time, and the coded values –1, 0, 1 for 340, 350, and 360 degrees for temperature. We will thus have three cakes baked at the package-recommended time and temperature (our center point), and four cakes with time and temperature spread around the center. Our response is an average palatability score, with higher values being desirable (data set `CakeBaking1`):

x_1	x_2	y
-1	-1	3.89
1	-1	6.36
-1	1	7.65
1	1	6.79
0	0	8.36
0	0	7.63
0	0	8.12

19.4 Analyzing First-Order Data

Here are three possible goals when analyzing data from a first-order design:

- Determine which design variables affect the response.

- Determine whether there is lack of fit.

- Determine the direction of steepest ascent.

Some experimental situations can involve a sequence of designs and all these goals. In all cases, model fitting for response surfaces is done using multiple linear regression. The model variables (x_1 through x_q for the first-order model) are the "independent" or "predictor" variables of the regression. The estimated regression coefficients are estimates of the model parameters β_k. For first-order models using data from 2^q factorials with or without center points, the estimated regression slopes using coded variables are equal to the ordinary main effects for the factorial model. Let **b** be the vector of estimated coefficients for first-order terms (an estimate of $\boldsymbol{\beta}$).

<div style="float:right; text-align:right;">Multiple regression to estimate β_k's</div>

Model testing is done with F-tests on mean squares from the ANOVA of the regression; each term has its own line in the ANOVA table. Predictor variables are orthogonal to each other in many designs and models, but not in all cases, and certainly not when there is missing data; so it seems easiest just to treat all testing situations as if the model variables were nonorthogonal.

To test the null hypothesis that the coefficients for a set of model terms are all zero, get the error sum of squares for the full model and the error sum of squares for the reduced model that does not contain the model terms being tested. The difference in these error sums of squares is the improvement sum of squares for the model terms under test. The improvement mean square is the improvement sum of squares divided by its degrees of freedom (the number of model terms in the multiple regression being tested). This improvement mean square is divided by the error mean square from the full model to obtain an F-test of the null hypothesis. The sum of squares for improvement can also be computed from a sequential (Type I) ANOVA for the model, provided that the terms being tested are the last terms entered into the model. The F-test of $\beta_k = 0$ (with one numerator degree of freedom) is equivalent to the t-test for β_k that is printed by most regression software.

<div style="float:right; text-align:right;">Test terms of interest adjusted for other terms in model</div>

In many response surface experiments, all variables are important, as there has been preliminary screening to find important variables prior to exploring the surface. However, inclusion of noise variables into models can alter subsequent analysis. It is worth noting that variables can look inert in some parts of a response surface, and active in other parts.

<div style="float:right; text-align:right;">Test to exclude noise variables from model</div>

The direction of steepest ascent in a first-order model is proportional to the coefficients $\boldsymbol{\beta}$. Our estimated direction of steepest ascent is then proportional to **b**. Inclusion of inert variables in the computation of this direction increases the error in the direction of the active variables. This effect is worst when the active variables have relatively small effects. The net effect is that our response will not increase as quickly as possible per unit change in the design variables, because the direction could have a nonnegligible component on the inert axes.

<div style="float:right; text-align:right;">Direction of steepest ascent proportional to estimated β's</div>

Residual variation can be divided into two parts: pure error and lack of fit. Pure error is variation among responses that have the same explanatory variables (and are in the same blocks, if there is blocking). We use replicated points, usually center points, to get an estimate of pure error. All the rest of residual variation that is not pure error is lack of fit. Thus we can make the

<div style="float:right; text-align:right;">Divide residual into pure error and lack of fit</div>

decompositions

$$\begin{aligned}
\mathrm{SS_{Tot}} &= \mathrm{SS_{Model}} + \mathrm{SS_{LoF}} + \mathrm{SS_{PE}} \\
N - 1 &= \mathrm{df_{Model}} + \mathrm{df_{LoF}} + \mathrm{df_{PE}} \ .
\end{aligned}$$

The mean square for pure error estimates σ^2, the variance of ϵ. If the model we have fit has the correct mean structure, then the mean square for lack of fit also estimates σ^2, and the F-ratio $\mathrm{MS_{LoF}}/\mathrm{MS_{PE}}$ will have an F-distribution with $\mathrm{df_{LoF}}$ and $\mathrm{df_{PE}}$ degrees of freedom. If the model we have fit has the wrong mean structure—for example, if we fit a first-order model and a second-order model is correct—then the expected value of $\mathrm{MS_{LoF}}$ is larger than σ^2. Thus we can test for lack of fit by comparing the F-ratio $\mathrm{MS_{LoF}}/\mathrm{MS_{PE}}$ to an F-distribution with $\mathrm{df_{LoF}}$ and $\mathrm{df_{PE}}$ degrees of freedom.

Pure error estimates σ^2; lack of fit measures deviation of model from true mean structure

For a 2^q factorial design with m center points, there are $2^q + m - 1$ degrees of freedom, with q for the model, $m - 1$ for pure error, and all the rest for lack of fit.

Quantities in the analysis of a first-order model are not very reliable when there is significant lack of fit. Because the model is not tracking the actual mean structure of the data, the importance of a variable in the first-order model may not relate to the variable's importance in the mean structure of the data. Likewise, the direction of steepest ascent from a first-order model may be meaningless if the model is not describing the true mean structure.

All bets off when lack of fit present

Example 19.2 Cake baking, continued

Example 19.1 was a 2^2 design with three center points. Our first-order model includes a constant and linear terms for time and temperature. With seven data points, there will be 4 residual degrees of freedom. The only replication in the design is at the three center points, so we have 2 degrees of freedom for pure error. The remaining 2 residual degrees of freedom are lack of fit.

Here are the results for this analysis as done in Minitab.

```
Estimated Regression Coefficients for y

Term           Coef        StDev           T       P
Constant     6.9714       0.5671      12.292   0.000
x1           0.4025       0.7503       0.536   0.620        ①
x2           1.0475       0.7503       1.396   0.235


S = 1.501      R-Sq = 35.9%      R-Sq(adj) = 3.8%


Analysis of Variance for y

Source               DF     Seq SS      Adj SS      Adj MS      F       P
Regression            2     5.0370      5.0370      2.5185   1.12   0.411
  Linear              2     5.0370      5.0370      2.5185   1.12   0.411
Residual Error        4     9.0064      9.0064      2.2516
  Lack-of-Fit         2     8.7296      8.7296      4.3648  31.53   0.031   ②
  Pure Error          2     0.2769      0.2769      0.1384
Total                 6    14.0435
```

Using the 4-degree-of-freedom residual mean square, neither time nor temperature has an F-ratio much bigger than one, so neither appears to affect the response ①. However, look at the test for lack of fit ②. This test has an F-ratio of 31.5 and p-value of .03, indicating that the first-order model is missing some of the mean structure.

The 2 degrees of freedom for lack of fit are the interaction in the factorial points and the contrast between the factorial points and the center points. The sums of squares for these contrasts are 2.77 and 5.96, so most of the lack of fit is due to the center points not lying on the plane fit from the factorial points. In fact, the center points are about 1.86 higher on average than what the first-order model predicts.

The direction of steepest ascent in this model is proportional to (.40, 1.05), the estimated β_1 and β_2. That is, the model says that a maximal increase in response can be obtained by increasing x_1 by .38 (coded) units for every increase of 1 (coded) unit in x_2. However, we have already seen that there is significant lack of fit using the first-order model with these data, so this direction of steepest ascent is not reliable.

19.5   Second-Order Models

We use second-order models when the portion of the response surface that we are describing has curvature. A second-order model contains all the terms in the first-order model, plus all quadratic terms like $\beta_{11}x_{1i}^2$ and all cross product terms like $\beta_{12}x_{1i}x_{2i}$. Specifically, it takes the form

Second-order models include quadratic and cross product terms

$$
\begin{aligned}
y_{ij} &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_q x_{qi} + \\
&\quad \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \cdots + \beta_{qq} x_{qi}^2 + \\
&\quad \beta_{12} x_{1i} x_{2i} + \beta_{13} x_{1i} x_{3i} + \cdots + \beta_{1q} x_{1i} x_{qi} + \\
&\quad \beta_{23} x_{2i} x_{3i} + \beta_{24} x_{2i} x_{4i} + \cdots + \beta_{2q} x_{2i} x_{qi} + \\
&\quad \cdots + \beta_{(q-1)q} x_{(q-1)i} x_{qi} + \epsilon_{ij} \\
&= \beta_0 + \sum_{k=1}^{q} \beta_k x_{ki} + \sum_{k=1}^{q} \beta_{kk} x_{ki}^2 + \sum_{k=1}^{q-1} \sum_{l=k+1}^{q} \beta_{kl} x_{ki} x_{li} + \epsilon_{ij} \\
&= \beta_0 + \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{x}_i' \mathcal{B} \mathbf{x}_i + \epsilon_{ij} \;,
\end{aligned}
$$

where once again $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{qi})'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$, and $\mathcal{B}$ is a $q \times q$ matrix with $\mathcal{B}_{kk} = \beta_{kk}$ and $\mathcal{B}_{kl} = \mathcal{B}_{lk} = \beta_{kl}/2$ for $k < l$. Note that the model only includes the kl cross product for $k < l$; the matrix form with $\mathcal{B}$ includes both kl and lk, so the coefficients are halved to take this into account.

Second-order models describe quadratic surfaces, and quadratic surfaces can take several shapes. Figure 19.3 shows four of the shapes that a quadratic surface can take. First, we have a simple minimum and maximum.   Then we have a ridge; the surface is curved (here a maximum) in one direction,

Quadratic surfaces take many shapes
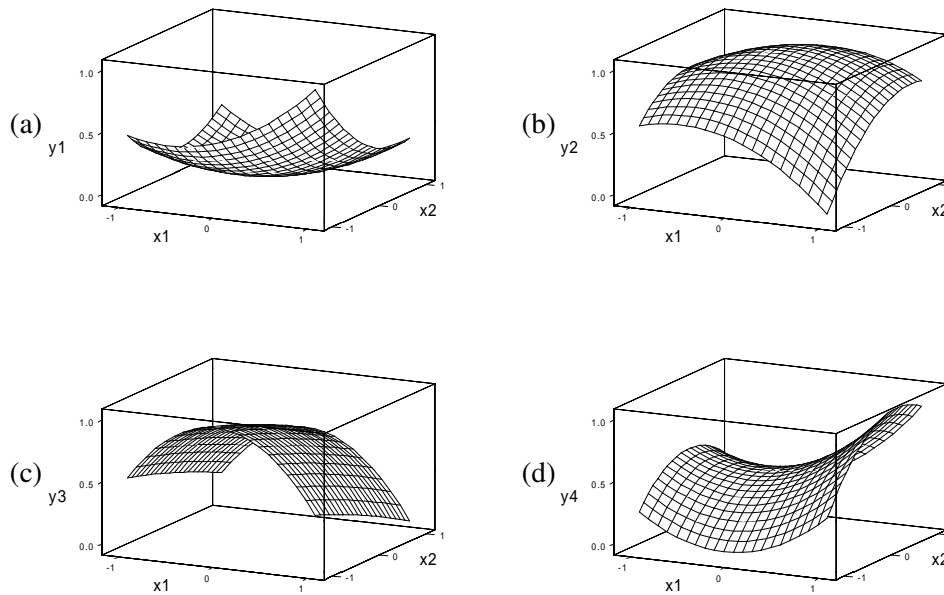
Figure 19.3: Sample second-order surfaces: (a) minimum, (b) maximum, (c) ridge, and (d) saddle, using Minitab.

but is fairly constant in another direction. Finally, we see a saddle point; the surface curves up in one direction and curves down in another.

Second-order models are easier to understand if we change from the original design variables x_1 and x_2 to *canonical variables* v_1 and v_2. Canonical variables will be defined shortly, but for now consider that they shift the origin (the zero point) and rotate the coordinate axes to match the second-order surface; the second-order model is very simple when expressed in canonical variables:

$$f_v(\mathbf{v}) = f_v(0) + \sum_{k=1}^{q} \lambda_k v_k^2 \ ,$$

Use canonical variables

where $\mathbf{v} = (v_1, v_2, \ldots, v_q)'$ is the design variables expressed in canonical coordinates; f_v is the response as a function of the canonical variables; and λ_k's are numbers computed from the $\mathcal{B}$ matrix. The $\mathbf{x}$ value that maps to 0 in the canonical variables is called the *stationary point* and is denoted by $\mathbf{x}_0$; thus $f_v(0) = f(\mathbf{x}_0)$.

The key to understanding canonical variables is the stationary point of the second-order surface. The stationary point is that combination of design variables where the surface is at either a maximum or a minimum in all directions. If the stationary point is a maximum in all directions, then the

Stationary point is maximum, minimum, or saddle point

stationary point is the maximum response on the whole modeled surface. If the stationary point is a minimum in all directions, then it is the minimum response on the whole modeled surface. If the stationary point is a maximum in some directions and a minimum in other directions, then the stationary point is a saddle point, and the modeled surface has no overall maximum or minimum. If a ridge surface is absolutely level in some direction, then it does not have a unique stationary point; this rarely happens in practice.

The stationary point will be the origin (0 point) for our canonical variables. Now imagine yourself situated at the stationary point of a second-order surface. The first canonical axis is the direction in which you would move so that a step of unit length yields a response as large as possible (either increase the response as much as possible or decrease it as little as possible). The second canonical axis is the direction, among all those directions perpendicular to the first canonical axis, that yields a response as large as possible. There are as many canonical axes as there are design variables. Each additional canonical axis that we find must be perpendicular to all those we have already found.

Figure 19.4 shows contours, stationary points, and canonical axes for the four sample second-order surfaces. As shown in this figure, contours for surfaces with maxima or minima are ellipses. The stationary point $\mathbf{x}_0$ is the center of these ellipses, and the canonical axes are the major and minor axes of the elliptical contours. For the ridge system, we still have elliptical contours, but they are very long and skinny, and the stationary point is outside the region where we have fit the model. If the ridge is absolutely flat, then the contours are parallel lines. For the saddle point, contours are hyperbolic instead of elliptical. The stationary point is in the center of the hyperbolas, and the canonical axes are the axes of the hyperbolas.

This description of second-order surfaces has been geometric; pictures are an easy way to understand these surfaces. It is difficult to calculate with pictures, though, so we also have an algebraic description of the second-order surface. Recall that the matrix form of the response surface is written

$$f(\mathbf{x}) = \beta_0 + \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\mathcal{B}\mathbf{x} \ \ .$$

Our algebraic description of the surface depends on the following facts:

1. The stationary point for this quadratic surface is at

$$\mathbf{x}_0 = -\frac{1}{2}\mathcal{B}^{-1}\boldsymbol{\beta} \ \ ,$$

   where $\mathcal{B}^{-1}$ is the matrix inverse of $\mathcal{B}$.

2. For the $q \times q$ symmetric matrix $\mathcal{B}$, we can find a $q \times q$ matrix H such that $H'H = HH' = I_q$ and $H'\mathcal{B}H = \Lambda$, where I_q is the $q \times q$ identity matrix and Λ is a matrix with elements $\lambda_1, \ldots, \lambda_q$ on the diagonal and zeroes off the diagonal.
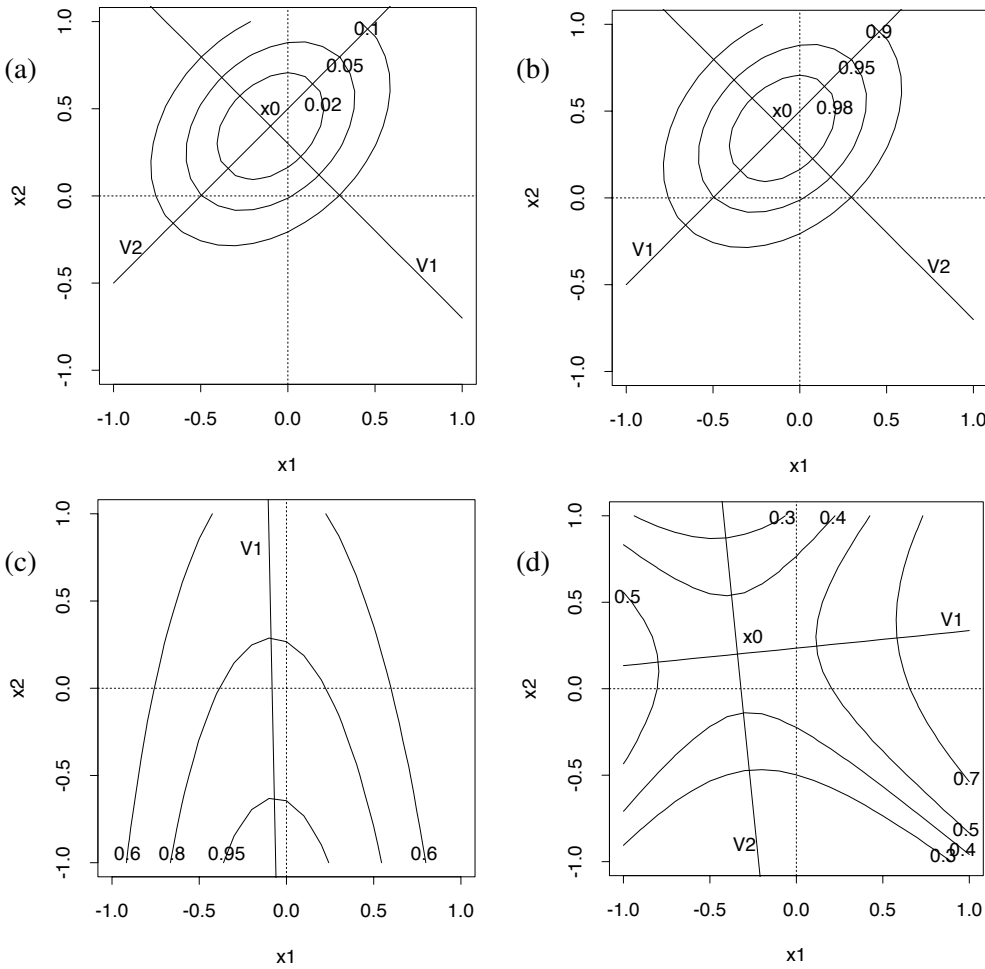
Figure 19.4: Contours, stationary points, and canonical axes for
sample second-order surfaces: (a) minimum, (b) maximum, (c) ridge,
and (d) saddle, using S-Plus.

The numbers λ_k are the *eigenvalues* of $\mathcal{B}$, and the columns of H are the
corresponding *eigenvectors*.

We saw in Figure 19.4 that the stationary point and canonical axes give us
a new coordinate system for the design variables. We get the new coordinates
$\mathbf{v}' = (v_1, v_2, \ldots, v_q)$ via

Get canonical
coordinates

$$\mathbf{v} = H'(\mathbf{x} - \mathbf{x}_0) \ .$$

Subtracting $\mathbf{x}_0$ shifts the origin, and multiplying by H' rotates to the canoni-
cal axes.

Finally, the payoff: in the canonical coordinates, we can express the response surface as

$$f_v(\mathbf{v}) = f_v(0) + \sum_{k=1}^{q} \lambda_k v_k^2 \ ,$$

where

$$f_v(0) = f(\mathbf{x}_0) = \beta_0 + \frac{1}{2}\mathbf{x}_0'\boldsymbol{\beta} \ .$$

That is, when looked at in the canonical coordinates, the response surface is a constant plus a simple squared term from each of the canonical variables v_i. If all of the λ_k's are positive, $\mathbf{x}_0$ is a minimum. If all of the λ_k's are negative, $\mathbf{x}_0$ is a maximum. If some are negative and some are positive, $\mathbf{x}_0$ is a saddle point. If all of the λ_k's are of the same sign, but some are near zero in value, we have a ridge system. The λ_k's for our four examples in Figure 19.4 are (.31771, .15886) for the surface with a minimum, (-.31771, -.15886) for the surface with a maximum, (-.021377, -.54561) for the surface with a ridge, and (.30822, -.29613) for the surface with a saddle point.

In principal, we could also use third- or higher-order models. This is rarely done, as second-order models are generally sufficient.

19.6   Second-Order Designs

There are several choices for second-order designs. One of the most popular is the *central composite design* (CCD). A CCD is composed of factorial points, *axial* points, and center points. Factorial points are the points from a 2^q design with levels coded as ± 1 or the points in a 2^{q-k} fraction with resolution V or greater; center points are again m points at the origin. The axial points have one design variable at $\pm \alpha$ and all other design variables at 0; there are $2q$ axial points. Figure 19.5 shows a CCD for $q = 3$.

One of the reasons that CCD's are so popular is that you can start with a first-order design using a 2^q factorial and then augment it with axial points and perhaps more center points to get a second-order design. For example, we may find lack of fit for a first-order model fit to data from a first-order design. Augment the first-order design by adding axial points and center points to get a CCD, which is a second-order design and can be used to fit a second-order model. We consider such a CCD to have been run in two incomplete blocks.

We get to choose α and the number of center points m. Suppose that we run our CCD in incomplete blocks, with the first block having the factorial points and center points, and the second block having axial points and center points. Block effects should be orthogonal to treatment effects, so that blocking does not affect the shape of our estimated response surface. We can achieve this orthogonality by choosing α and the number of center points in the factorial and axial blocks as shown in Table 19.1 (Box and Hunter 1957).

Table 19.1 deserves some explanation. When blocking the CCD, factorial points and axial points will be in different blocks. The factorial points may
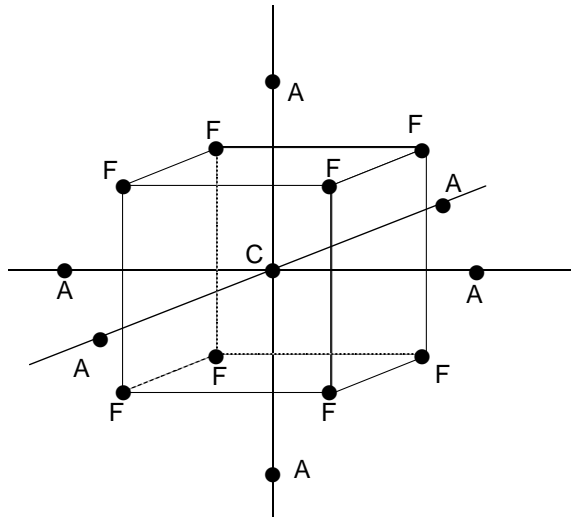
Figure 19.5: A central composite design in three dimensions, showing center (C), factorial (F), and axial (A) points.

Table 19.1: Design parameters for Central Composite Designs with orthogonal blocking.

q	2	3	4	5	5	6	6	7	7
rep	1	1	1	1	$\frac{1}{2}$	1	$\frac{1}{2}$	1	$\frac{1}{2}$
Number of blocks in factorial	1	2	2	4	1	8	2	16	8
Center points per factorial block	3	2	2	2	6	1	4	1	1
α for axial points	1.414	1.633	2.000	2.366	2.000	2.828	2.366	3.364	2.828
Center points for axial block	3	2	2	4	1	6	2	11	4
Total points in design	14	20	30	54	33	90	54	169	80

also be blocked using the confounding schemes of Chapter 15. The table gives the maximum number of blocks into which the factorial portion can be confounded, while main effects and two-way interactions are confounded only with three-way and higher-order interactions. The table also gives the number of center points for *each* of these blocks. If fewer blocks are desired, the center points are added to the combined blocks. For example, the 2^5 can be run in four blocks, with two center points per block. If we instead use two blocks, then each should have four center points; with only one block, use all eight center points. The final block consists of all axial points and additional center points.

There are a couple of heuristics for choosing α and the number of center

Table 19.2: Parameters for rotatable, uniform precision Central Composite Designs.

q	2	3	4	5	5	6	6	7	7
Replication	1	1	1	1	$\frac{1}{2}$	1	$\frac{1}{2}$	1	$\frac{1}{2}$
Number of center points	5	6	7	10	6	15	9	21	14

points when the CCD is not blocked, but these are just guidelines and not overly compelling. If the precision of the estimated response surface at some point **x** depends only on the distance from **x** to the origin, not on the direction, then the design is said to be *rotatable*. Thus rotatable designs do not favor one direction over another when we explore the surface. This is reasonable when we know little about the surface before experimentation. We get a rotatable design by choosing $\alpha = 2^{q/4}$ for the full factorial or $\alpha = 2^{(q-k)/4}$ for a fractional factorial. Some of the blocked CCD's given in Table 19.1 are exactly rotatable, and all are nearly rotatable.

α for rotatable design

Rotatable designs are nice, and I would probably choose one as a default. However, I don't obsess on rotatability, for a couple of reasons. First, rotatability depends on the coding we choose. The property that the precision of the estimated surface does not depend on direction disappears when we go back to the original, uncoded variables. It also disappears if we keep the same design points in the original variables but then express them with a different coding. Second, rotatable designs use five levels of every variable, and this may be logistically awkward. Thus choosing $\alpha = 1$ so that all variables have only three levels may make a more practical design. Third, using $\alpha = \sqrt{q}$ so that all the noncenter points are on the surface of a sphere (only rotatable for $q = 2$) gives a better design when we are only interested in the response surface within that sphere.

Rotatable designs need five levels of every factor and depend on coding

A second-order design has *uniform precision* if the precision of the fitted surface is the same at the origin and at a distance of 1 from the origin. Uniform precision is a reasonable criterion, because we are unlikely to know just how close to the origin a maximum or other surface feature may be; (relatively) too many center points give us much better precision near the origin, and too few give us better precision away from the origin. It is impossible to achieve this exactly; Table 19.2 shows the number of center points to get as close as possible to uniform precision for rotatable CCD's.

m for uniform precision

Example 19.3 Cake baking, continued

We saw in Example 19.2 that the first-order model was a poor fit; in particular, the contrast between the factorial points and the center points indicated curvature of the response surface. We will need a second-order model to fit the curved surface, so we will need a second-order design to collect the data for the fit.

We already have factorial points and three center points. Looking in Table 19.1, we see that adding three more center points and axial points at

$\alpha = 1.414$ will give us a design with two blocks with blocks orthogonal to treatments. This design is also rotatable, but not uniform precision.

Here is the complete design, including responses for the seven additional cakes we bake to complete the CCD (data set `CakeBaking2`):

Block	x_1	x_2	y
1	−1	−1	3.89
1	1	−1	6.36
1	−1	1	7.65
1	1	1	6.79
1	0	0	8.36
1	0	0	7.63
1	0	0	8.12
2	1.414	0	8.40
2	−1.414	0	5.38
2	0	1.414	7.00
2	0	−1.414	4.51
2	0	0	7.81
2	0	0	8.44
2	0	0	8.06

There are several other second-order designs in addition to central composite designs. The simplest are 3^q factorials and fractions with resolution V or greater. These designs are not much used for $q \geq 3$, as they require large numbers of design points.

<div style="float:right">3^q designs</div>

Box-Behnken designs are rotatable, second-order designs that are incomplete 3^q factorials, but not ordinary fractions. Box-Behnken designs are formed by combining incomplete block designs with factorials. For q factors, find an incomplete block design for q treatments in blocks of size two. (Blocks of other sizes may be used, we merely illustrate with two.) Associate the "treatment" letters A, B, C, and so on with "factor" letters A, B, C, and so on. When two factor letters appear together in a block, use all combinations where those factors are at the ± 1 levels, and all other factors are at 0. The combinations from all blocks are then joined with some center points to form the Box-Behnken design.

<div style="float:right">Box-Behnken designs</div>

For example, for $q = 3$, we can use the BIBD with three blocks and (A,B), (A,C), and (B,C) as assignment of treatments to blocks. From the three blocks, we get the combinations:

A	B	C		A	B	C		A	B	C
x_1	x_2	x_3		x_1	x_2	x_3		x_1	x_2	x_3
−1	−1	0		−1	0	−1		0	−1	−1
−1	1	0		−1	0	1		0	−1	1
1	−1	0		1	0	−1		0	1	−1
1	1	0		1	0	1		0	1	1

To this we add some center points, say five, to form the complete design. This design takes only 17 points, instead of the 27 (plus some for replication) needed in the full factorial.

19.7   Second-Order Analysis

Here are three possible goals for the analysis of second-order models:

- Determine which design variables affect the response.

- Determine whether there is lack of fit.

- Determine the stationary point and surface type.

As with first-order models, fitting is done with multiple linear regression, and testing is done with F-tests. Let $\mathbf{b}$ be the estimated coefficients for first-order terms, and let $\mathbf{B}$ be the estimate of the second-order terms.

Use regression and F-tests

The goal of determining which variables affect the response is a bit more complex for second-order models. To test that a variable—say variable 1—has no effect on the response, we must test that its linear, quadratic, and cross product coefficients are all zero: $\beta_1 = \beta_{11} = \cdots = \beta_{1q} = 0$. This is a $q + 1$-degree-of-freedom null hypothesis which we must test using an F-test.

Test all coefficients to exclude a variable

Testing for lack of fit in the second-order model is completely analogous to the first-order model. Compute an estimate of pure error variability from the replicated points; all other residual variability is lack of fit. Significant lack of fit indicates that our model is not capturing the mean structure in our region of experimentation. When we have significant lack of fit, we should first consider whether a transformation of the response will improve the quality of the fit. For example, a second-order model may be a good fit for the log of the response. Alternatively, we can investigate higher-order models for the mean or obtain data to fit the second-order model in a smaller region.

Canonical analysis is the determination of the type of second-order surface, the location of its stationary point, and the canonical directions. These quantites are functions of the estimated coefficients $\mathbf{b}$ and $\mathbf{B}$ computed in the multiple regression. We estimate the stationary point as $\widehat{x}_0 = -\mathbf{B}^{-1}\mathbf{b}/2$, and the eigenvectors and eigenvalues of $\mathcal{B}$ are estimated by the eigenvectors and eigenvalues of $\mathbf{B}$ using special software.

Canonical analysis for shape of surface

Example 19.4 Cake baking, continued

We now fit a second-order model to the data from the blocked central composite design of Example 19.3. This model will have linear terms, quadratic terms, a cross product term, and a block term. Here are the results in Minitab.

```
Estimated Regression Coefficients for y

Term            Coef        StDev          T      P
Constant        8.070       0.1842     43.809  0.000          ①
Block          -0.057       0.1206     -0.473  0.651
x1              0.735       0.1595      4.608  0.002
x2              0.964       0.1595      6.042  0.001
x1*x1          -0.628       0.1661     -3.779  0.007
x2*x2          -1.195       0.1661     -7.197  0.000
x1*x2          -0.832       0.2256     -3.690  0.008


S = 0.4512      R-Sq = 95.0%      R-Sq(adj) = 90.8%


Analysis of Variance for y

Source             DF     Seq SS      Adj SS      Adj MS      F      P
Blocks              1     0.0457      0.0455      0.04546   0.22  0.651
Regression          5    27.2047     27.2047      5.44094  26.72  0.000
  Linear            2    11.7562     11.7562      5.87808  28.87  0.000
  Square            2    12.6763     12.6763      6.33816  31.13  0.000
  Interaction       1     2.7722      2.7722      2.77223  13.62  0.008
Residual Error      7     1.4252      1.4252      0.20359
  Lack-of-Fit       3     0.9470      0.9470      0.31567   2.64  0.186   ②
  Pure Error        4     0.4781      0.4781      0.11953
Total              13    28.6756
```

At ① we see that all first- and second-order terms are significant, so that no variables need to be deleted from the model. We also see that lack of fit is not significant ②, so the second-order model should be a reasonable approximation to the mean structure in the region of experimentation.

Figure 19.6 shows a contour plot of the fitted second-order model. We see that the optimum is at about .4 coded time units above 0, and .2 coded temperature units above zero, corresponding to 35.8 minutes and 352°. We also see that the ellipse slopes northwest to southeast, meaning that we can trade time for temperature and still get a cake that we like.

Here is a canonical analysis for this surface done in MacAnova.

```
component: b0              ①
(1)          8.07
component: b               ②
(1)      0.73515        0.964
component: B               ③
(1,1)    -0.62756    -0.41625
(2,1)    -0.41625    -1.1952
component: x0              ④
(1,1)     0.41383
(2,1)     0.25915
component: y0              ⑤
(1,1)      8.347
component: H               ⑥
(1,1)    -0.88413    -0.46724
(2,1)     0.46724    -0.88413
component: lambda          ⑦
(1)     -0.40758      -1.4152
```
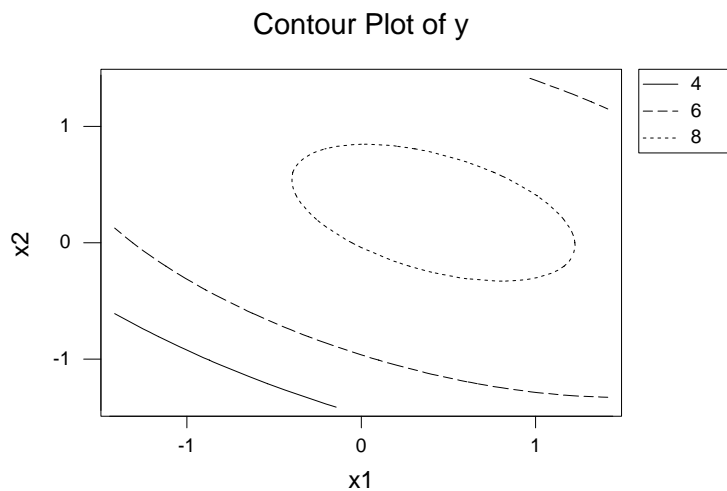
Figure 19.6: Contour plot of fitted second-order model for cake baking data, using Minitab.

The estimated coefficients are at ① ($\widehat{\beta}_0$), ② (**b**), and ③ (**B**). The estimated stationary point and its response are at ④ and ⑤; I guessed (.4, .2) for the stationary point from Figure 19.6—it was actually (.42, .26). The estimated eigenvectors and eigenvalues are at ⑥ and ⑦. Both eigenvalues are negative, indicating a maximum. The smallest decrease is associated with the first eigenvector (-.884, .467), so increasing the temperature by .53 coded units for every decrease in 1 coded unit of time keeps the response as close to maximum as possible.

The results of a canonical analysis have an aura of precision that is often not justified. Many software packages can compute and print the estimated stationary point, but few give a standard error for this estimate. In fact, the standard error is difficult to compute and tends to be rather large. Likewise, there can be considerable error in the estimated canonical directions.

19.8 Mixture Experiments

Mixture experiments are a special case of response surface experiments in which the response depends on the proportions of the various components, but not on absolute amounts. For example, the taste of a punch depends on the proportion of ingredients, not on the amount of punch that is mixed, and the strength of an alloy may depend on the proportions of the various metals in the alloy, but not on the total amount of alloy produced.

Mixtures depend on proportions

The design variables $x_1, x_2, \ldots, x_q$ in a mixture experiment are propor-

Table 19.3: Blends of fruit punch. Data set `FruitPunch`.

x_1	x_2	x_3	Appeal		
1	0	0	4.3	4.7	4.8
0	1	0	6.2	6.5	6.3
.5	.5	0	6.3	6.1	5.8
0	0	1	7.0	6.9	7.4
.5	0	.5	6.1	6.5	5.9
0	.5	.5	6.2	6.1	6.2

tions, so they must be nonnegative and add to one:

$$x_k \geq 0, \qquad k = 1, 2, \cdots, q$$

and

$$x_1 + x_2 + \cdots + x_q = 1 \ .$$

This design space is called a *simplex* in q dimensions. In two dimensions, the design space is the segment from (1,0) to (0,1); in three dimensions, it is bounded by the equilateral triangle (0,0,1), (0,1,0), and (1,0,0); and so on. Note that a point in the simplex in q dimensions is determined by any $q-1$ of the coordinates, with the remaining coordinate determined by the constraint that the coordinates add to one.

Mixtures have a simplex design space

Example 19.5 Fruit punch

Cornell (1985) gave an example of a three-component fruit punch mixture experiment, where the goal is to find the most appealing mixture of watermelon juice (x_1), pineapple juice (x_2), and orange juice (x_3). Appeal depends on the recipe, not on the quantity of punch produced, so it is the proportions of the constituents that matter. Six different punches are produced, and eighteen judges are assigned at random to the punches, three to a punch. The recipes and results are given in Table 19.3 (data set `FruitPunch`).

As in ordinary response surfaces, we have some response y that we wish to model as a function of the explanatory variables:

$$y_{ij} = f(x_{1i}, x_{2i}, \cdots, x_{qi}) + \epsilon_{ij} \ .$$

We use a low-order polynomial for this model, not because we believe that the function really is polynomial, but rather because we usually don't know what the correct model form is; we are willing to settle for a reasonable approximation to the underlying function. We can use this model for various purposes:

Model response with low-order polynomial

- To predict the response at any combination of design variables,

- To find combinations of design variables that give best response, and

- To measure the effects of various factors on the response.

(a)   (b)
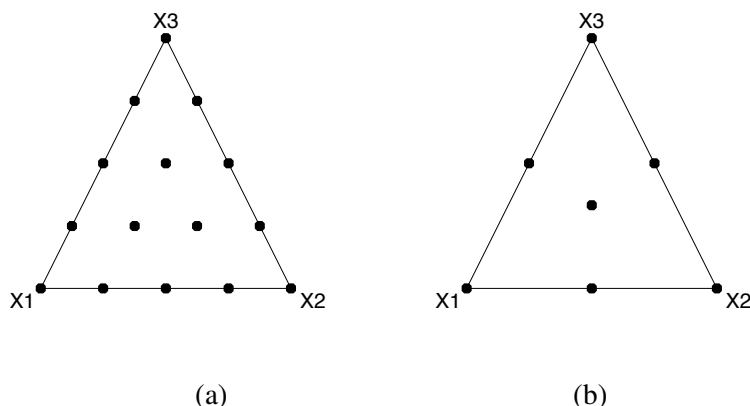
Figure 19.7: (a) {3,4} simplex lattice and (b) three variable simplex centroid designs.

19.8.1   Designs for mixtures

A {q,m} *simplex lattice* design for q components consists of all design points on the simplex where each component is of the form r/m, for some integer $r = 0, 1, 2, \ldots, m$. For example, the {3,2} simplex lattice consists of the six combinations (1, 0, 0), (0, 1, 0), (0, 0, 1), (1/2, 1/2, 0), (1/2, 0, 1/2), and (0, 1/2, 1/2). The fruit punch experiment in Example 19.5 is a {3,2} simplex lattice. The {3,3} simplex lattice has the ten combinations (1, 0, 0), (0, 1, 0), (0, 0, 1), (2/3, 1/3, 0), (2/3, 0, 1/3), (1/3, 2/3, 0), (0, 2/3, 1/3), (1/3, 0, 2/3), (0, 1/3, 2/3), and (1/3, 1/3, 1/3). In general, m needs to be at least as large as q to get any points in the interior of the simplex, and m needs to be larger still to get more points into the interior of the simplex. Figure 19.7(a) illustrates a {3,4} simplex lattice.

> Simplex lattice design

The second class of models is the *simplex centroid* designs. These designs have $2^q - 1$ design points for q factors. The design points are the pure mixtures, all the 1/2-1/2 two-component mixtures, all the 1/3-1/3-1/3 three-component mixtures, and so on, through the equal mixture of all q components. Alternatively, we may describe this design as all the permutations of $(1, 0, \ldots, 0)$, all the permutations of $(1/2, 1/2, \ldots, 0)$, all the permutations of $(1/3, 1/3, 1/3, \ldots, 0)$, and so on to the point $(1/q, 1/q, \ldots, 1/q)$. A simplex centroid design only has one point in the interior of the simplex; all the rest are on the boundary. Figure 19.7(b) illustrates a simplex centroid in three factors.

> Simplex centroid design

Mixtures in the interior of the simplex—that is, mixtures which include at least some of each component—are called *complete* mixtures. We sometimes need to do our experiments with complete mixtures. This may arise for several reasons, for example, all components may need to be present for a chemical reaction to take place.

> Complete mixtures have all $x_k > 0$

Factorial ratios provide one class of designs for complete mixtures. This design is a factorial in the ratios of the first $q - 1$ components to the last

> Factorial ratios vary x_k/x_q

Table 19.4: Harvey Wallbanger mixture experiment. Data set
`Wallbangers`.

O/G	V/G	G	V	O	Rating
4.0	1.2	.161	.194	.645	3.6
9.0	1.2	.089	.107	.804	5.1
4.0	2.8	.128	.359	.513	3.8
9.0	2.8	.078	.219	.703	3.8
6.5	2.0	.105	.211	.684	4.7
4.0	2.0	.143	.286	.571	2.4
9.0	2.0	.083	.167	.750	4.0

component. We may want to reorder our components to obtain a convenient "last" component. The design points will have ratios x_k/x_q that take a few fixed values (the factorial levels) for each k, and we then solve for the actual proportions of the components. For example, if $x_1/x_3 = 4$ and $x_2/x_3 = 2$, then $x_1 = 4/7$, $x_2 = 2/7$, and $x_3 = 1/7$. Only complete mixtures occur in a factorial ratios design with all ratios greater than 0.

Example 19.6 Harvey Wallbangers

Sahrmann, Piepel, and Cornell (1987) ran an experiment to find the best proportions for orange juice (O), vodka (V), and Galliano (G) in a mixed drink called a Harvey Wallbanger. Only complete mixtures are considered, because it is the mixture of these three ingredients that defines a Wallbanger (as opposed to say, orange juice and vodka, which is a drink called a screwdriver). Furthermore, preliminary screening established some approximate limits for the various components.

The authors used a factorial ratios model, with three levels of the ratio V/G (1.2, 2.0, and 2.8) and two levels of the ratio O/G (4 and 9). They also ran a center point at V/G = 2 and O/G = 6.5. Their actual design included incomplete blocks (so that no evaluator consumed more than a small number of drinks). However, there were no apparent evaluator differences, so the average score was used as response for each mixture, and blocks were ignored. Evaluators rated the drinks on a 1 to 7 scale. The data are given in Table 19.4, which also shows the actual proportions of the three components.

A second class of complete-mixture designs arises when we have lower bounds for each component: $x_k \geq d_k > 0$, where $\sum d_k = D < 1$. Here, we define *pseudocomponents*

Pseudocomponents

$$x'_k = \frac{x_k - d_k}{1 - D}$$

and do a simplex lattice or simplex centroid design in the pseudocomponents. The pseudocomponents map back to the original components via

$$x_k = d_k + (1 - D)x'_k \ .$$

Many realistic mixture problems are constrained in some way so that the available design space is not the full simplex or even a simplex of pseudo-

components. A regulatory constraint might say that ice cream must contain at least a certain percent fat, so we are constrained to use mixtures that contain at least the required amount of fat; and an economic constraint requires that our recipe cost less than a fixed amount. Mixture designs can be adapted to such situations, but we often need special software to determine a good design for a specific model over a constrained space.

<div align="right">Many mixture problems have constrained design spaces</div>

19.8.2 Models for mixture designs

Polynomial models for a mixture response have fewer parameters than the general polynomial model found in ordinary response surfaces for the same number of design variables. This reduction in parameters arises from the simplex constraints on the mixture components—some terms disappear due to the linear restrictions among the mixture components. For example, consider a first-order model for a mixture with three components. In such a mixture, we have $x_1 + x_2 + x_3 = 1$. Thus,

<div align="right">Mixture constraints reduce parameter count</div>

$$
\begin{aligned}
f(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
&= \beta_0(x_1 + x_2 + x_3) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
&= (\beta_1 + \beta_0)x_1 + (\beta_2 + \beta_0)x_2 + (\beta_3 + \beta_0)x_3 \\
&= \tilde{\beta}_1 x_1 \qquad + \tilde{\beta}_2 x_2 \qquad + \tilde{\beta}_3 x_3
\end{aligned}
$$

In this model, the linear constraint on the mixture components has allowed us to eliminate the constant from the model. This reduced model is called the *canonical form* of the mixture polynomial. We will simply use β in place of $\tilde{\beta}$ in the sequel.

<div align="right">Canonical form of first-order model</div>

Mixture constraints also permit simplifications in second-order models. Not only can we eliminate the constant, but we can also eliminate the pure quadratic terms! For example:

$$
\begin{aligned}
x_1^2 &= x_1 x_1 \\
&= x_1(1 - x_2 - x_3 - \cdots - x_q) \\
&= x_1 - x_1 x_2 - x_1 x_3 - \cdots - x_1 x_q \ .
\end{aligned}
$$

By making similar substitutions for all pure quadratic terms, we get the canonical form:

<div align="right">Canonical form of second-order model</div>

$$
f(x_1, x_2, \cdots, x_q) = \sum_{k=1}^{q} \beta_k x_k + \sum_{k<l}^{q} \beta_{kl} x_k x_l \ .
$$

Third-order models are sometimes fit for mixtures; the canonical form for the full third-order model is:

<div align="right">Canonical form of third-order model</div>

$$f(x_1, x_2, \cdots, x_q) \quad = \quad \sum_{k=1}^{q} \beta_k x_k + \sum_{k<l}^{q} \beta_{kl} x_k x_l$$
$$+ \sum_{k<l}^{q} \delta_{kl} x_k x_l (x_k - x_l) + \sum_{k<l<m}^{q} \beta_{klm} x_k x_l x_m \quad .$$

A subset of the full cubic model called the *special cubic* model sometimes appears:

<div style="text-align: right">Special cubic model</div>

$$f(x_1, x_2, \cdots, x_q) = \sum_{k=1}^{q} \beta_k x_k + \sum_{k<l}^{q} \beta_{kl} x_k x_l + \sum_{k<l<n}^{q} \beta_{kln} x_k x_l x_n \quad .$$

Coefficients in mixture canonical polynomials have interpretations that are somewhat different from standard polynomials. If the mixture is pure (that is, contains only a single component, say component k), then x_k is 1 and the other components are 0. The predicted response is β_k. Thus the "linear" coefficients give the predicted response when the mixture is simply a single component. If the mixture is a 50-50 mix of components k and l, then the predicted response is $\beta_k/2 + \beta_l/2 + \beta_{kl}/4$. Thus the bivariate interaction terms correspond to deviations from a simple additive fit, and in particular show how the response for pairwise blends varies from additive. The three-way interaction term β_{klm} has a similar interpretation for triples. The cubic interaction term δ_{kl} provides some asymmetry in the response to two-way blends.

<div style="text-align: right">Mixture coefficients have special interpretations</div>

We may use ordinary polynomial models in $q-1$ factors instead of reduced polynomial models in q factors. For example, the canonical quadratic model in $q = 3$ factors is

<div style="text-align: right">Fewer factors as an alternative to reduced models</div>

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \quad .$$

We can instead use the model

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{\beta}_{12} x_1 x_2 + \tilde{\beta}_{11} x_1^2 + \tilde{\beta}_{22} x_2^2 \quad ,$$

which is the usual quadratic model for $q = 2$ factors. The models are equivalent mathematically, and which model you choose is personal preference. There are linear relations between the models that allow you to transfer between the representations. For example, $\tilde{\beta}_0 = \beta_3$ ($x_3 = 1$, $x_1 = x_2 = 0$), and $\tilde{\beta}_0 + \tilde{\beta}_1 + \tilde{\beta}_{11} = \beta_1$ ($x_1 = 1$, $x_2 = x_3 = 0$).

Factorial ratios experiments also have the option of using polynomials in the components, polynomials in the ratios, or a combination of the two. The choice of model can sometimes be determined *a priori* but will frequently be determined by choosing the model that best fits the data.

Example 19.7 Harvey Wallbangers, continued

Example 19.6 introduced the Harvey Wallbanger data. Here are the results from fitting the canonical second-order model in MacAnova.

```
              Coef         StdErr             t
           -518.14         41.143       -12.594
           -12.625         1.1111       -11.363
            100.56         5.8373        17.226
g           812.73        55.472        14.651
g           126.64        56.449         2.2435
y          -101.53         5.8706       -17.294

: 7, MSE: 0.0042851, DF: 1, R^2: 0.99996
egression F(6,1): 4344.4, Durbin-Watson: 2.1195
```

All terms are significant with the exception of the vodka by Galliano interaction (though there is only 1 degree of freedom for error, so significance testing is rather dubious).

It is difficult to interpret the coefficients directly. The usual interpretations for coefficients are for pure mixtures and two-component mixtures, but this experiment was conducted on a small region in the interior of the design space. Thus using the model for pure mixtures or two-component mixtures would be an unwarranted extrapolation. The best approach is to plot the contours of the fitted response surface, as shown in Figure 19.8. We see that there is a saddle point near the fifth design point (the center point), and the highest estimated responses are on the boundary between the first two design points. This has the V/G ratio at 1.2 and the O/G ratio between 4.0 and 9.0, but somewhat closer to 9.

19.9   Further Reading and Extensions

As might be expected, there is much more to the subjects discussed in this chapter. Box and Draper (1987) and Cornell (1990) provide excellent book-length coverage of response surfaces and mixture experiments respectively.

Earlier we alluded to the issue of constraints on the design space. These constraints can make it difficult to run standard response surface or mixture designs. Special-purpose computer software (for example, Design-Expert) can construct good designs for constrained situations. These designs are generally chosen to be optimal in the sense of minimizing the estimation variance. See Cook and Nachtsheim (1980) or Cook and Nachtsheim (1989). A second interesting area is trying to optimize when there is more than one response. Multiple responses are common in the real world, and methods have been proposed to compromise among the competing criteria. See Myers, Khuri, and Carter (1989) and the references cited there.

19.10   Problems

We run a central composite design and fit a second-order model. The fitted coefficients are: **Exercise 19.1**

$$y = 86 + 9.2x_1 + 7.3x_2 - 7.8x_1^2 - 3.9x_2^2 - 6.0x_1x_2 \ \ .$$
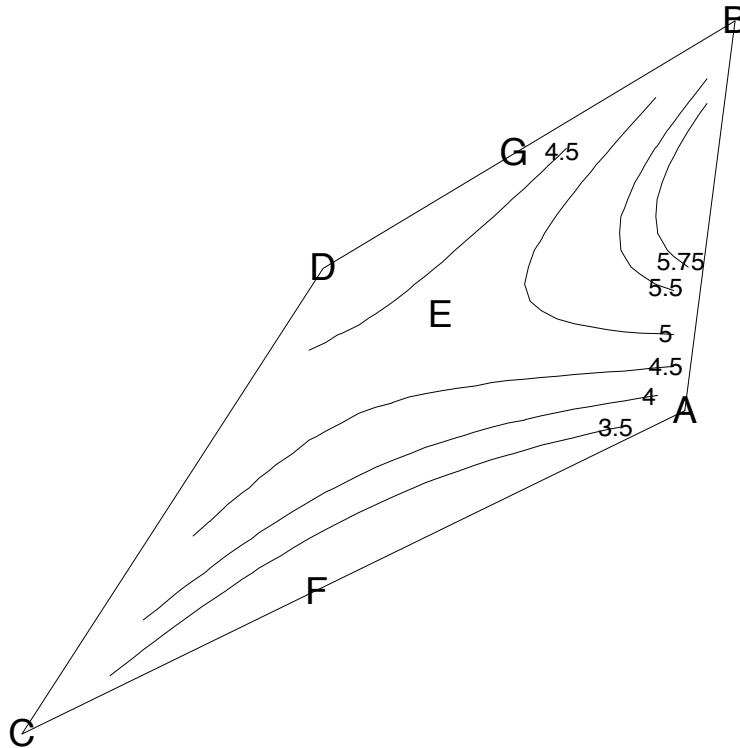
Figure 19.8: Contour plot for Harvey Wallbanger data, using S-Plus. Letters indicate the points of Table 19.4 in the table order.

Perform the canonical analysis on this response surface.

Exercise 19.2 Fit the second-order model to the fruit punch data of Example 19.5. Which mixture gives the highest appeal?

Exercise 19.3 The whiteness of acrylic fabrics after being washed at different detergent concentrations (.09 to .21 percent) and temperatures (29 to 41°C) was measured and the following model was obtained (Prato and Morris 1984):

$$y = -116.27 + 819.58x_1 + 1.77x_2 - 1145.34x_1^2 - .01x_2^2 - 3.48x_1x_2 \ .$$

Perform the canonical analysis on this response surface.

Problem 19.1 Briefly describe an experimental design appropriate for each of the following situations. Describe treatments, blocks, etc.

(a) I like tortilla chips and salsa, but if you're not careful at the store you can wind up with tortilla chip crumbs instead of chips. What I want to understand is the variability in the fraction of unbroken chips from bag to bag (measured as fraction of whole chips by weight). It could just be

variability from bag to bag, but it could also be variability from store to store (some clumsy guy might keep dropping the bags). I'm willing to buy 20 bags and do the separation and weighing in order to solve this issue of vital importance.

(b) An Ankle Foot Orthosis (AFO) is a device worn by a person with significant ankle impairment. An active AFO actually assists the person in walking by moving the foot through the typical pointing and flexing as required in various parts of the stride. However, any such active device needs to be optimized for the individual user, as every user prefers something a little different from the standard settings. This involves finding the best compromise settings for the motor speed and the length of a lever that is part of the device.

Testing is done by setting the speed and length and then having the user walk around for five minutes. At the end of that time, the user gives a "comfort rating." Users will test multiple possible settings, then the technician will select the best speed and length for the user based on all of the collected comfort ratings. Given the time required to do this customization, we must be able to choose the best settings after just 12 trials.

(c) Alfalfa is grown and cut for animal fodder. We wish to investigate the effects of variety of alfalfa and cutting schedule on the quality of the alfalfa produced. We have four varieties to compare, and three cutting schedules. There are 12 test sites available for our use, four at each of three experiment stations. The seeds tend to get strewn pretty broadly during planting, so we do not want to seed areas smaller than a test site. However, the cutting can be done on a small scale.

(d) Computer systems are compared by their performance on standard benchmark programs. Computer source code for the benchmark is compiled to make an executable which is then run to measure speed. The trick is that our compiler has 24 options that can be turned off or on, and we would like to find which options should be used for good performance. Our boss wants the answer right away, and we will only have time to try 32 different combinations of compiler options.

(e) One test for whether a chemical causes mutations uses bacteria that cannot produce an essential amino acid and thus cannot grow unless they mutate to produce that acid. A measured dose of bacteria are spread on a petri dish containing the mutagen and a nutritional broth (which does *not* contain the missing amino acid!). The dish is placed in a gentle environment; three days later, we detect the presence of a mutation if there is any bacterial growth. We wish to test five concentrations of the suspected mutagen and two different broths using 50 petri dishes. We have no reason to expect systematic differences between dishes.

(f) The Department of Natural Resources is under political pressure to keep deer populations fairly high so that there will be plenty of deer for hunters. The issue under consideration is whether farmers leaving some standing crop in corn fields over winter decreases winter mortality of deer. (That

is, does the additional food source make any difference?) Researchers have identified twelve forested locations where the deer tend to group (called "yards") during the winter. These locations are near corn fields, and the farmers have agreed to participate in the study. There are three of each of four kinds of sites: evergreen forest cover immediately adjacent to the fields, deciduous forest cover immediately adjacent to the fields, evergreen forest cover across a highway from the fields, and deciduous forest cover across a highway from the fields. We anticipate that both forest cover type and vehicle traffic may affect the mortality rate over the winter. We can assign crop left standing or no crop left standing to any field, and measure the difference in fall and spring populations as the response for any yarding area.

(g) The cookbook *Joy of Cooking* has a dynamite recipe for egg nog (and repeats an adage attributed to Mark Twain about how too much of anything is bad, except for whisky, for which too much is just right). We want to design an experiment that will compare the use of light rum, dark rum, and bourbon in this recipe; we make a batch using each liquor. A rater should drink at least four ounces of one of the egg nogs before giving his or her rating, and we expect substantial rater to rater variability. Furthermore, this recipe contains a lot of alcohol, and, to put it gently, ratings after the second cup are not reliable. We have twelve people at our party willing to dedicate a few moments to science and participate in the experiment.

The recipe is great, but it uses raw eggs. If you decide to try it, please use eggs certified as salmonella free. Egg substitutes will work, but they're not as good as fresh eggs.

(h) Medical tablets (pills) contain active ingredients plus other constituents including a binder, diluents, disintegrants, lubricants, and so on. The manufacturer mixes up a batch and then presses the mixture into tablets using a press. In the current experiment, we wish to study how the concentration of disintegrant affects how long it takes a tablet to dissolve in water. Specifically, we want to study five different concentrations.

The standard protocol for measuring dissolve time is to produce a batch of tablets with the given formula, randomly choose 12 of the produced tablets, and take the average of the 12 dissolve times as the response. However, since we are making these tablets under lab rather than factory conditions, we can only make two batches per day. Furthermore, environmental conditions (e.g., humidity when pressing the tablets) can affect dissolve times, and we would expect these conditions to vary from day to day.

The boss wants this all finished in 10 days. Choose an appropriate design for this experiment.

(i) Every year at Christmas I make "thumbprint" cookies. I want to perfect them, and I need to find the right time and temperature for baking them. It should be about 375 degrees for about 11 minutes, but it may not be that

exactly. I'm going to make 12 trays of these cookies this week, and from those trays I want to be able to estimate the best time and temperature for baking. Each tray can have a different recipe, but every cookie on the tray must be the same recipe. The response is how good the cookies taste to me.

(j) Atmospheric carbon dioxide is increasing and causing global warming. To add insult to very serious injury, increased carbon dioxide causes poison ivy to grow faster. We want to determine the growth rate of poison ivy as a function of carbon dioxide concentration and temperature. Our range of interest is current average summer temperature up to current average plus 4 degrees C, and current carbon dioxide concentration up to current concentration plus 25%. We have 12 environmental chambers in which we can set temperature and carbon dioxide concentration. We put poison ivy plants into the chambers, and then measure growth after three months.

(k) The ramp meter controversy refuses to die! Someone finally decided that just turning all the meters off for six weeks was a fairly crude way to assess the effects of meters. This time around, we're going to look at a stretch of I-35 containing 7 metered on-ramps. Each of these meters can be individually set to a fast or slow setting. We are trying to find which, if any, of the meter settings affects the vehicle-miles per hour (vmph: number of vehicles times speed) on the stretch of highway during rush periods. Each setting should be used for a full week to get a reliable response. Unfortunately, we have to finish the study in eight weeks.

(l) Whole house air exchangers have become important as houses become more tightly sealed and the dangers of indoor air pollution become known. Exchangers are used primarily in winter, when they draw in fresh air from the outside and exhaust an equal volume of indoor air. In the process, heat from the exhausted indoor air is used to warm the incoming air. The design problem is to construct an exchanger that maximizes energy efficiency while maintaining air flow volume within tolerances. Energy efficiency is energy saved by heating the incoming air minus energy used to power the fan. There are two design variables: the pore size of the exchanger and the fan speed. In general, as the pore size decreases the energy saved through heat exchange increases, but for smaller pores the fan must be run faster to maintain air flow, thus using more energy.

We have a current guess as to the best settings for maximum energy efficiency (pore size P and fan speed S). Any settings with 15% of P and S will provide acceptable air flow, and we feel that the optimum is probably within about 5% of these current settings.

(m) Neuropeptide Y (NPY) is believed to be involved in the regulation of feeding and basal metabolism. When rat brains are perfused with NPY, the rats dramatically increase their food intake over the next 24 hours. Naloxone (NLX) may potentially block the effects of NPY. If so, it could be an important line of research in obesity studies. We wish to test the effect of four treatments, the factorial combinations of brain perfusion

by either NPY or saline (as a control), and the subcutaneous injection of either NLX or saline (as a control) on 24-hour post-treatment food intake. We have available 32 male inbred, essentially similar rats.

(n) We are trying to produce a new cleaning solvent for circuit boards. We anticipate that a combination of three standard solvents will work as well as the specialty solvent currently in use, but beyond knowing that we want each of the three to be at least 10% of the combination, we don't know how much of each to use.

(o) Child development specialists are interested in factors affecting the ability of children to solve "ten questions" puzzles. In these puzzles the child is given a set of pictures, one of which has been chosen by the researcher. The child gets to ask questions that the researcher answers either yes or no; on the basis of these answers the child tries to determine which of the pictures has been chosen. The response the researchers are looking at is the number of questions (ten maximum) that the child asks before determining the chosen picture. Two factors are under study: the number of pictures to choose from (either fifteen or twenty), and the familiarity of the objects in the pictures (either dinosaurs or birds, and oddly enough, I think the dinosaurs are the familiar objects!). The researchers have funds to study twelve children, and they expect substantial child to child variation. All children will do four puzzles, one of each type. They expect learning to take place, so that the later puzzles will generally be solved more quickly.

(p) A fertilizer company is developing a rose fertilizer which consists of a nitrogen compound N, a phosphorus compound P, a potassium compound K, and an inert binder to hold it all together. (The binder can be disregarded in the experiment.) The company believes that there are optimum levels of N, P, and K to give best rose yield, and they believe that their current settings $N_0 = 6$, $P_0 = 6$, and $K_0 = 4$ (kg per 100 kg of fertilizer) are pretty close to optimal; probably each is within 10% of the optimal values. They want to find the optimal values.

For each of the following, briefly describe the design used and give a skeleton ANOVA.

Problem 19.2

(a) Laser light is scattered as it passes through a transparent PVC sample. The amount of light that passes through the sample may depend on the sample thickness and possibly on the degree of polishing given the surface of the sample. In this experiment, there are five thicknesses of PVC and three surface polishing treatments. We have 90 PVC blanks. Eighteen of these blanks are randomly chosen for the first thickness and shaved to that thickness. These 18 pieces are then randomly assigned to the three polishing treatments, six per treatment. This procedure is then repeated for the remaining blanks until all have been shaped and polished (and all thicknesses have been used). The 90 pieces are then measured (in random order) for laser light transmission.

(b) Smooth operation of an automobile depends in part on the ease with which the driver can reach and manipulate the controls. You are designing the controls for a car and you have two possibilities (old and new) for each of the following controls: windshield wiper switch, cruise control switch, and headlights switch. Eight driving simulators have been set up, one with each of the combinations of the three factors. The simulators are held at separate locations. Those simulators with an even number of new features are at location A, whereas those with an odd number of new features are at location B. Twenty-four subjects are nonrandomly divided into two groups, with the first 12 subjects sent to location A and the second 12 to location B. Each subject will use all four simulators at his or her location in random order and rate the overall ease-of-use for each control setup as the response.

(c) Doctors wish to assess the influence of three anti-viral drugs (two actual drugs and a control) on the survival of SARS patients. As new patients are identified, they are randomly assigned to one of the three drugs. Survival over the next month is noted as the response. It is suspected that patient age affects survival; age was noted for each patient, but it was not logistically feasible to block on age.

(d) More and more states are requiring that students pass a major exam before they can get their high school diplomas. However, there is little research into whether these exams improve achievement, affect graduation rates, or have other consequences. One forward thinking state decides to run an experiment to explore some of these issues. Twenty moderate sized school districts will take part; each of these districts has two high schools. The twenty districts are divided at random into two sets of ten. Incoming ninth graders in the first group of districts will be required to pass an exit exam before they can graduate. No test requirement is made in the second set of districts. In each district, the two high schools are randomly assigned to two different curricula. One curriculum is the standard curriculum that the district was already using, and the other curriculum is tailored to the exit exam. After four years (that is, when this year's freshmen are scheduled to graduate), we measure the graduation rate at each of the 40 high schools.

(e) One criterion for a paper airplane is that it should glide for a long time. This experiment compares three different designs for a paper airplane. All of the airplanes are made from new standard US Letter paper from the same package. Eighteen sheets of paper are randomly assigned to three different airplane designs, six sheets per design. On Friday I fold all of the planes. On Saturday, my two daughters and I test the planes. Each of us flies one plane of each type, and each individual plane is only flown once. The flights are randomized in order, but each type of plane is flown once by each of us, and each type of plane is flown once in the living room, once in the front yard, and once in the back yard. On Sunday, the three of us repeat the experiment using the remaining nine planes. The response is the length of time the plane glides.

(f) There is considerable concern about how drainage from farm fields de-

grades water quality in the Minnesota river. One suggested treatment is to install settling ponds on the drainage ditches just before they flow into the river. The idea is that the nutrient-rich sediments will settle in the ponds rather than go into the river. Twenty drainage ditches that are suitable for settling pond installation have been located. Ten of these are chosen at random for ponds; the other ten are left as is. For each ditch we measure the nutrient flow into the river over a growing season as the response. We also measure the volume of water flowing from each ditch, as rainfall volume affects water volume, which probably affects nutrient flow.

(g) National forests are managed for multiple uses, including wildlife habitat. Suppose that we are managing our multiple-use forest, and we want to know how snowmobiling and timber harvest method affect timber wolf reproductive success (as measured by number of pups surviving to 1 year of age over a 5-year interval). We may permit or ban snowmobiles; snowmobiles cover a lot of area when present, so we can only change the snowmobile factor over large areas. We have three timber harvest methods, and they are fairly easy to change over small areas. We have six large, widely dispersed forest sections that we may use for the experiment. We choose three sections at random and ban snowmobiles there. The other three sections allow snowmobiles. Each of these sections is divided into three zones, and we randomly assign one of the three harvest methods to each zone within each section. (Note that we do not harvest the entire zone; we merely use that harvest method when we do harvest within the zone.) We observe timber wolf success in each zone.

(h) Some aircraft have in-flight deicing systems that are designed to prevent or remove ice buildup from the wings. A manufacturer wishes to compare three different deicing systems. This is done by installing the system on a test aircraft and flying the test aircraft behind a second plane that sprays a fine mist into the path of the test aircraft. The wings are photographed, and the ice buildup is estimated from interpretation of the photographs. They make five test flights for each of the three systems. The amount of buildup is influenced by temperature and humidity at flight altitude. The flights will be made at constant temperature (achieved by slightly varying the altitude); relative humidity cannot be controlled, but will be measured at the time of the flight.

(i) We wish to study new varieties of corn for disease resistance. We start by taking four varieties (A, B, C, D) and cross them (pollen from type A, B, C or D fertilizing flowers from type A, B, C, or D), getting sixteen crosses. (This is called a diallel cross experiment, and yes, four of the sixteen "crosses" are actually pure varieties.) The sixteen crosses produce seed, and we now treat the crosses as varieties for our experiment. We have 48 plots available, 16 plots in each of St. Paul, Crookston, and Waseca. We randomly assign each of the crosses to one of the sixteen plots at each location.

(j) A political scientist wishes to study how polling methods affect results. Two candidates (A and B) are seeking endorsement at their party con-

vention. A random sample of 3600 voters has been taken and divided at random into nine sets of 400. All voters were asked if they support candidate A. However, before the question was asked, they were either told (a) that the poll is funded by candidate A, (b) that the poll is funded by candidate B, or (c) nothing. Due to logistical constraints, all voters in a given set (of 400) were given the same information; the response for a set of 400 is the number supporting candidate A. The three versions of information were randomly assigned to the nine sets.

Problem 19.3

Three components of a rocket propellant are the binder (x_1), the oxidizer (x_2), and the fuel (x_3). We want to find the mixtures that yield coefficients of elasticity (y) less than 3000. All components must be present and there are minimum proportions, so the investigators used a pseudocomponents design, with the following pseudocomponent values and results (data from Kurotori 1966 via Park 1978, data set `RocketFuel`):

x_1	x_2	x_3	y
1	0	0	2350
0	1	0	2450
0	0	1	2650
1/2	1/2	0	2400
1/2	0	1/2	2750
0	1/2	1/2	2950
1/3	1/3	1/3	3000
2/3	1/6	1/6	2690
1/6	2/3	1/6	2770
1/6	1/6	2/3	2980

Does this design correspond to any of our standard mixture designs? Does it have an estimate of pure error? Fit the second-order mixture model. Is the estimated maximum above 3000? Where is the estimated maximum, and where is the region that has elasticity less than 3000?

Problem 19.4

Millers want to make bread flours that bake into large loaves. They need to mix flours from four varieties of wheat, so they run an experiment with different mixtures and measure the volume of the resulting loaves (ml/100 g dough). The experiment was performed on 2 separate days, obtaining the following results (data from Draper et al. 1993, data set `LoafVolume`):

		Day 1					Day 2		
x_1	x_2	x_3	x_4	Volume	x_1	x_2	x_3	x_4	Volume
0	.25	0	.75	403	0	.75	0	.25	423
.25	0	.75	0	425	.25	0	.75	0	417
0	.75	0	.25	442	0	.25	0	.75	388
.75	0	.25	0	433	.75	0	.25	0	407
0	.75	.25	0	445	0	0	.25	.75	338
.25	0	0	.75	435	.25	.75	0	0	435
0	0	.75	.25	385	0	.25	.75	0	379
.75	.25	0	0	425	.75	0	0	.25	406
.25	.25	.25	.25	433	.25	.25	.25	.25	439

Analyze these data to determine which mixture of flours yields the largest loaves.

An experiment is performed to determine how a gasoline engine responds to various factors. The response of interest is CO emissions in grams per hour. The design factors are engine load, in Newton meters, range (30,70); engine speed, in rpm, range (1000, 4000); spark advance, in degrees, range (10, 30); air-to-fuel ratio, dimensionless, range (13, 16.4); and exhaust gas recycle, in percent, range (0, 10). The experimental design has 46 observations in two blocks of 23 each. The design factors have been coded to the range (-1, 1) in the table below (data from Draper et al. 1994, data set COEmissions). Analyze these data and describe how CO emissions depend on engine settings.

Problem 19.5

Load	Speed	Advance	Ratio	Recycle	Block	Response
−1	−1	0	0	0	1	81
1	−1	0	0	0	1	148
−1	1	0	0	0	1	348
1	1	0	0	0	1	530
0	0	−1	−1	0	1	1906
0	0	1	−1	0	1	1717
0	0	−1	1	0	1	91
0	0	1	1	0	1	42
0	−1	0	0	−1	1	86
0	1	0	0	−1	1	435
0	−1	0	0	1	1	93
0	1	0	0	1	1	474
−1	0	−1	0	0	1	224
1	0	−1	0	0	1	346
−1	0	1	0	0	1	147
1	0	1	0	0	1	287
0	0	0	−1	−1	1	1743
0	0	0	1	−1	1	46
0	0	0	−1	1	1	1767
0	0	0	1	1	1	73
0	0	0	0	0	1	195
0	0	0	0	0	1	233
0	0	0	0	0	1	236
0	−1	−1	0	0	2	100
0	1	−1	0	0	2	559
0	−1	1	0	0	2	118
0	1	1	0	0	2	406
−1	0	0	−1	0	2	1255
1	0	0	−1	0	2	2513
−1	0	0	1	0	2	53
1	0	0	1	0	2	54
0	0	−1	0	−1	2	270
0	0	1	0	−1	2	277
0	0	−1	0	1	2	303
0	0	1	0	1	2	213
−1	0	0	0	−1	2	171
1	0	0	0	−1	2	344
−1	0	0	0	1	2	180
1	0	0	0	1	2	280
0	−1	0	−1	0	2	548
0	1	0	−1	0	2	3046
0	−1	0	1	0	2	13
0	1	0	1	0	2	123
0	0	0	0	0	2	228
0	0	0	0	0	2	201
0	0	0	0	0	2	238

Curing time and temperature affect the shear strength of an adhesive that bonds galvanized steel bars. The following experiment was repeated on 2 separate days. Twenty-four pieces of steel are obtained by random sampling from warehouse stock. These are grouped into twelve pairs; the twelve pairs are glued and then cured with one of nine curing treatments assigned at random. The treatments are the three by three factorial combinations of temperature (375°, 400°, and 450^oF, coded -1, 0, 2) and time (30, 35, or 40 seconds, coded -1, 0, 1). Four pairs were assigned to the center point, and one pair to all other conditions. The response is shear strength (in psi, data from Khuri 1992, data set `SteelBars`):

Problem 19.6

Temp.	Time	Day 1	Day 2
-1	-1	1226	1213
0	-1	1898	1961
2	-1	2142	2184
-1	0	1472	1606
0	0	2010	2450
0	0	1882	2355
0	0	1915	2420
0	0	2106	2240
2	0	2352	2298
-1	1	1491	2298
0	1	2078	2531
2	1	2531	2609

Determine the temperature and time settings that give strong bonds.

Suppose we are fitting a first-order model using data from a 2^q design with m center points, but a second-order model is actually correct. Show that the contrast formed by taking the average response at the factorial points minus the average at the center points estimates the sum of the quadratic coefficients of the second-order model. Show that the two-factor interaction effects in the factorial points estimate the cross product terms in the second-order model.

Question 19.1

Chapter 20

On Your Own

Adult birds push their babies out of the nest to force them to learn to fly. As I write this, I have a 16-year-old daughter learning to drive. And you, our statistical children, must leave the cozy confines of textbook problems and graduate to the real world of designing and analyzing your own experiments for your own goals. This final chapter is an attempt at a framework for the experimental design process, to help you on your way to designing real-world experiments.

20.1 Experimental Context

An individual experiment is usually part of a larger research enterprise; thus planning an experiment takes place within this larger context. One way to frame this larger context is hierarchically, with goals, objectives, and hypotheses. The (overall) goals are for the large research enterprise. For example, we might have the goal of developing artificial heated-butter aromas for the food industry. The (immediate) objective is a refinement of the goals to narrow the scope of investigation. Continuing the butter aroma example, we might have the objective of determining which naturally occurring odorants in heated butter influence the perceived butter aroma. Finally, hypotheses are specific, answerable questions regarding an objective that can be addressed in an experiment. We might ask, can human subjects detect the difference in aroma between heated butter and this particular mixture of compounds?

Goals, objectives, and hypotheses

> We design experiments to answer the questions raised in our hypotheses.

20.2 Experiments by the Numbers

Many authors have presented guidelines for designing experiments. Noteworthy among these are Kempthorne (1952), Cochran and Cox (1957), Cox

(1958), Daniel (1976), and Box, Hunter, and Hunter (1978). I have tried to synthesize a number of these recommendations into a sequence of steps for designing an experiment, which are presented below. Experimentation, like all science, is not one-size-fits-all, but these steps will work for many investigations.

I have two basic rules when planning an experiment. The first is "Use all the information you have available to you." Most of this information is subject matter information (what you know about treatments, units, and so on) rather than statistical tactics. The second is "Use the simplest possible design that gets the job done." Thus when designing an experiment I consider the fancy tricks of the trade only when they are needed.

Information and simplicity

1. Do background research. At a minimum, you should

   - Determine what is already *known* about your problem. Researchers know things that have been discovered by experiment and verified by repeated experiments. You may wish to repeat a "known" experiment if you are trying to verify it, extend it to a new population, or learn an experimental technique, but more often you will be looking at new hypotheses.

   - Determine what other researchers *suspect* about your problem. Many experiments are follow-up experiments on vague indications from earlier research. For example, a preliminary experiment may have indicated the possibility that a particular drug was effective against breast cancer, but the sample size was too small to be conclusive.

   - Determine what background or extraneous factors (for example, environmental factors) might affect the outcome of your experiment. Here we are looking ahead to the possibility that blocking might be needed, so we identify the sources of extraneous variation on which we may need to block.

   - Find out what related experiments have been done, what types of designs were used, and what kinds of problems were encountered. There is always room for innovation, particularly if earlier experiments encountered problems, but experimental designs that work well are worth imitating.

   - Determine the cost or availability of experimental material such as animals, equipment, and chemical stocks; determine your time and monetary budgets. Time and money are major constraints on experimentation. Determine these constraints early.

This research takes time, but it will save you time later.

2. Decide which question to address next, and clearly state your question. This process should include:

   - A list of hypotheses to be tested or effects to be estimated.

   - An ordering of these hypotheses or effects by importance.

Draft of Nov 18, 1999

- An ordering of these hypotheses or effects by logical or time sequence if some should be examined before others.

Your experiment is part of the research enterprise, so choose your hypotheses to address your current objectives. Knowing if some hypotheses are more important than others will matter for designs such as split plots, which are more precise for split-plot factors than for whole-plot factors.

Remember, science is sequential, with new results building on old results. Unless you have an overwhelming argument to the contrary, plan for a sequence of hypotheses and experiments and *don't try to do everything in a single experiment!*

3. Determine the treatments to be studied, experimental units to be used, and responses to be measured. These depend on the hypotheses being addressed and the population about which you wish to make inferences. Choice of treatments includes the consideration of controls (probably needed) and/or placebo treatments.

The type of experimental units you use will determine the population about which you can make inferences and usually the size of your experimental errors. Homogeneous units generally lead to smaller experimental errors and thus shorter confidence intervals and more powerful tests. On the other hand, homogeneous units often represent a narrow subset of all potential units, and it can be difficult to argue that conclusions reached about a homogeneous subset of a population hold for the entire population. If you need to work with a heterogeneous population of units, you will probably need to consider blocking the experiment.

The response or responses to be measured are usually determined by the hypotheses, but you must still determine how they will be measured, what the measurement units are, and whether blinding will be needed.

4. Design the current experiment. Try simple designs first; if upon inspection the simple design won't do the job for some reason, you can design a fancier experiment. But at least contemplate the simple experiment first. Keep the qualities of a good design in mind—design to avoid systematic error, to be precise, to allow esimation of error, and to have broad validity.

5. Inspect the design for scientific adequacy and practicality.

- Are there any systematic problems that would invalidate your results or reduce their range of generalization? For example, does your design have confounding that biases your comparisons?

- Are there treatments or factor-level combinations that are impractical or simply cannot be used? For example, you may have several factors that involve time, and the overall time may be impractical when all factors are at the high level; or perhaps some treatments are "a little too exothermic" (as my chemistry T.A. described one of our proposed experiments).

- Do you have the time and resources to carry out the experiment?

If there are problems in any of these areas, you will need to go back to step 4 and revise your design. For example, the simple design was a full factorial, but it was too big, so we could move to a fancier design such as a fractional factorial.

6. Inspect the design for statistical adequacy and practicality.

- Do you know how to analyze the results?

- Will your experiment satisfy the statistical or model assumptions implicit in the statistical analysis?

- Do you have enough degrees of freedom for error for all terms of interest?

- Will you have adequate power or precision?

- Will the analysis be easy to interpret?

- Can you account for aliasing?

If you answer any of these in the negative, you will need to go back to step 4 and revise your design. For example, you might need to add blocking to reduce variability, or you might decide that a design with an unbalanced mixed-effects model was simply too difficult to analyze. Study the design carefully for oversights or mistakes. For example, I have seen split-plot designs with no degrees of freedom for error at the whole-plot level. (The investigator had intended to use an interaction for a surrogate error, but all interactions were at the split-plot level.)

7. Run the experiment.

8. Analyze the results. Pay close attention to where model or distributional assumptions might fail, and take corrective action if necessary. For example,

- Do factors assumed to be additive actually interact, or do treatments act differently in different blocks?

- Is the error variance nonconstant?

- Are there outliers in the data?

- Do the random errors follow the normal distribution?

- Are there unmodeled dependencies in the data (for example, time dependencies)?

Consider whether the experiment as run answers the questions, or if some further observations are needed. For example, you might want to rerun suspected outlier points, or you might need another fraction of a factorial to disentangle some aliases.

9. Draw conclusions, giving estimates of error or reliability. Assess this experiment in relation to similar experiments. Reporting is crucial, and it is only a slight exaggeration to say that an experiment not reported is an experiment not conducted. I like to begin reports with a short "executive summary"

giving the conclusions, and then add sections on the experimental design and analysis (many journals call such sections "Materials and Methods" and "Results").

10. Consider what needs to be studied next. Research is ongoing and sequential, and one completed experiment leads to the design of the next.

It is clear that a carefully planned experiment requires a great deal of effort. Many of the steps in planning an experiment are nonstatistical and require considerable background knowledge in the subject being studied, while other steps require substantial statistical knowledge. Thus experimental design is often a team effort, with subject matter experts and statistical experts working together. One goal of this book has been to make the statistical part of the planning a little easier.

20.3 Final Project

Design an experiment, run the experiment, analyze the results, and report your findings.

This is not an overnight homework problem, but a project with several stages. Stage one is the project proposal, which should include a description of your hypotheses and proposed experimental design. This proposal should be sufficiently complete that anyone could replicate your experiment given just your proposal. Submit your proposal to your instructor for approval before conducting the experiment.

Stage two is running the experiment. Here you are on your own.

Stage three is analysis and reporting. Your report will typically be in the five to ten page range and should include a summary giving the conclusions, an introduction to the problem stating the background and hypothesis to be tested, a description of the experimental design (similar to stage one), and a description of the analysis. The description of the analysis should not be a batch of unannotated computer output. It should say what you are doing, why you are doing it, and what it tells you. Output and figures can be intermixed or appended separately.

The subject of the experiment is up to you and your instructor. Those of you in graduate school or at work in a research area may be able to adapt your own ongoing work to this project. Or just try something fun—food experiments (particularly desserts!) are always attractive, as are the experiments of youth such as rolling balls down inclined planes.

Bibliography

Adcock, C. J. (1988). A bayesian approach to calculating sample sizes. *Journal of the Royal Statistical Society, Series D (The Statistician) 37*, 433–439.

Addelman, S. (1962). Orthogonal main-effects plans for asymmetrical factorial experiments. *Technometrics 4*, 21–46.

Al-Darrab, I. A., Z. A. Khan, and S. I. Ishrat (2009). An experimental study on the effect of mobile phone conversation on drivers reaction time in braking response. *Journal of Safety Research 40*, 185–189.

Ali, M. M. (1984). An approximation to the null distribution and power of the Durbin-Watson statistic. *Biometrika 71*, 253–261.

Alley, M. C., C. B. Uhl, and M. M. Lieber (1982). Improved detection of drug cytotoxicity in the soft agar colony formation assay through use of a metabolizable tetrazolium salt. *Life Sciences 31*, 3071–3078.

Amoah, F. M., K. Osei-Bonsu, and F. K. Oppong (1997). Response of improved robusta coffee to location and management practices in Ghana. *Experimental Agriculture 33*, 103–111.

Anderson, R. L. (1954). The problem of autocorrelation in regression analysis. *Journal of the American Statistical Association 49*, 113–129.

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey (1972). *Robust Estimates of Location: Survey and Advances*. Princeton, NJ: Princeton University Press.

Aniche, N. G. and N. Okafor (1989). Studies on the effect of germination time and temperature on malting of rice. *Journal of the Institute of Brewing 95*, 165–167.

Annadurai, G., R.-S. Juang, and D.-J. Lee (2002). Factorial design analysis for adsorption of dye on activated carbon beads incorporated with calcium alginate. *Advances in Environmental Research 6*, 191–198.

Atkinson, A. C. (1985). *Plots, Transformations, and Regression*. Oxford, U.K.: Oxford University Press.

Baker, M. (2016). 1,500 scientists lift the lid of reproducibility. *Nature News 533*, 452.

Barnett, V. and T. Lewis (1994). *Outliers in Statistical Data* (Third ed.). New York: Wiley.

Basso, D. and L. Salmaso (2006). A discussion of permutation tests conditional to observed responses in unreplicated 2^m full factorial designs. *Communications in Statistics - Theory and Methods 35*, 83–97.

Beckman, R. J. and R. D. Cook (1983). Outlier . . . . s. *Technometrics 25*, 119–149.

Bellavance, F. and S. Tardif (1995). A nonparametric approach to the analysis of three-treatment three-period crossover designs. *Biometrika 82*, 865–875.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B 57*, 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29*, 1165–1188.

Berger, J. O. and R. L. Wolpert (1988). *The Likelihood Principle, 2nd Edition*. Hayward, CA: Institute of Mathematical Statistics.

Bergman, B. and A. Hynén (1997). Dispersion effects from unreplicated designs in the 2^{k-p} series. *Technometrics 39*, 191–198.

Bernhardson, C. S. (1975). Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. *Biometrics 31*, 229–332.

Berrama, T., N. Benaouag, F. Kaouah, and Z. Bendjama (2013). Application of full factorial design to study the simultaneous removal of copper and zinc from aqueous solution by liquid-liquid extraction. *Desalination and Water Treatment 51*, 2135–2145.

Berry, D. A. (1988). Multiple comparisons, multiple tests, and data dredging: a bayesian perspective (with discussion). In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics*, pp. 79–94. Oxford, England: Oxford University Press.

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery 5*, 27–36.

Berry, D. A. and Y. Hochberg (1999). Bayesian perspecitves on multiple comparisons. *Journal of Statistical Planning and Inference 82*, 215–227.

Bezjak, Z. and B. Knez (1995). Workplace design and loadings in the process of sewing garments. *International Journal of Clothing Science and Technology 7*, 89–101.

Bicking, C. A. (1958). Experiences and needs for design in ordnance experimentation. In V. Chew (Ed.), *Experimental Designs in Industry*, pp. 247–252. New York: Wiley.

Bin Jantan, I., A. S. Bin Ahmad, and A. R. Bin Ahmad (1987). Tapping of oleo-resin from *Dipterocarpus kerrii*. *The Malaysian Forester 50*, 343–353.

Boehner, A. W. (1975). *The Effect of Three Species of Logging Slash on the Properties of Aspen Planer Shavings Particleboard*. Ph. D. thesis, University of Minnesota, St. Paul, MN.

Bose, R. C., W. H. Clatworthy, and S. S. Shrikhande (1954). Tables of partially balanced designs with two associate classes. Technical Bulletin 107, North Carolina Agricultural Experiment Station.

Bose, R. C. and K. R. Nair (1939). Partially balanced incomplete block designs. *Sankhya 4*, 337–372.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics 25*, 290–302.

Box, G. E. P. (1988). Signal-to-noise ratios, performance criteria, and transformations. *Technometrics 30*, 1–17.

Box, G. E. P. and S. L. Andersen (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumptions. *Journal of the Royal Statistical Society, Series B 17*, 1–34.

Box, G. E. P., S. Bisgaard, and C. A. Fung (1988). An explanation and critique of Taguchi's contributions to quality engineering. *Quality and Reliability Engineering International 4*, 123–131.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B 26*, 211–243.

Box, G. E. P. and N. R. Draper (1987). *Empirical Model-Building with Response Surfaces*. New York: Wiley.

Box, G. E. P. and J. S. Hunter (1957). Multi-factor experimental designs for exploring response surfaces. *Annals of Mathematical Statistics 28*, 195–241.

Box, G. E. P., W. G. Hunter, and J. S. Hunter (1978). *Statistics for Experimenters*. New York: Wiley.

Bréfort, H., J. X. Guinard, and M. J. Lewis (1989). The contribution of dextrins of beer sensory properties, part II. Aftertaste. *Journal of the Institute of Brewing 95*, 431–435.

Brown, M. B. (1975). Exploring interaction effects in the ANOVA. *Applied Statistics 24*, 288–298.

Brown, M. B. and A. B. Forsythe (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics 16*, 129–132.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software 80*, 1–28.

Bussan, A. J. (1995). Selection for weed competitiveness among soybean genotypes. Master's thesis, University of Minnesota, St. Paul, MN.

Carmer, S. G. and W. M. Walker (1982). Baby bear's dilemma: A statistical tale. *Agronomy Journal 74*, 122–124.

Caro, M. R., E. Zamora, L. Leon, F. Cuello, J. Salinas, D. Megias, M. J. Cubero, and A. Contreras (1990). Isolation and identification of *Listeria monocytogenes* in vegetable byproduct silages containing preservative additives and destined for animal feeding. *Animal Feed Science and Technology 31*, 285–291.

Carroll, M. B. and O. Dykstra (1958). Application of fractional factorials in a food research laboratory. In V. Chew (Ed.), *Experimental Designs in Industry*, pp. 224–234. New York: Wiley.

Casella, G. and R. Berger (2002). *Statistical Inference, second edition*. Pacific Grove, CA: Duxbury–Wadsworth.

CAST Investigators (1989). Effect of encainide and flecanide on mortality in a randomized trial of arrhythmia. *New England Journal of Medicine 312*, 406–412.

Chaloner, K. and I. Verdinelli (1995). Bayesian experimental design: A review. *Statistical Science 10*, 273–304.

Christensen, R. and L. G. Blackwood (1993). Tests for precision and accuracy of multiple measuring devices. *Technometrics 35*, 411–420.

Chu, Y. C. (1970). Comparison of *in vivo* and *vitro* inhibition of AT-Pases by the insecticide Chlordane. Master's thesis, University of Minnesota, St. Paul, MN.

Clatworthy, W. H. (1973). *Tables of Two-Associate Class Partially Balanced Designs*. National Bureau of Standards, Applied Mathematics Series, No. 63.

Cochran, W. G., K. M. Autrey, and C. Y. Cannon (1941). A double change-over design for dairy cattle feeding experiments. *Journal of Dairy Science 24*, 937–951.

Cochran, W. G. and G. M. Cox (1957). *Experimental Designs*. New York: Wiley.

Cole, D. N. (1993). Trampling effects on mountain vegetation in Washington, Colorado, New Hampshire, and North Carolina. Research Paper INT-464, U.S. Forest Service, Intermountain Research Station.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science 349*(6251).

Connolloy, H. M., J. L. Crary, M. D. McGoon, D. D. Hensrud, B. S. Edwards, W. D. Edwards, and H. V. Schaff (1997). Valvular heart disease associated with fenfluramine-phentermine. *New England Journal of Medicine 337*, 581–588.

Connor, W. S. (1958). Experiences with incomplete block designs. In V. Chew (Ed.), *Experimental Designs in Industry*, pp. 193–206. New York: Wiley.

Conover, W. J. (1980). *Practical Nonparametric Statistics* (Second ed.). New York: Wiley.

Conover, W. J., M. E. Johnson, and M. M. Johnson (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics 23*, 351–361.

Cook, R. D. and C. J. Nachtsheim (1980). A comparison of algorithms for constructing exact D-optimal designs. *Technometrics 22*, 315–324.

Cook, R. D. and C. J. Nachtsheim (1989). Computer-aided blocking of factorial and response-surface designs. *Technometrics 31*, 339–346.

Cook, R. D. and S. Weisberg (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.

Cornell, J. A. (1985). Mixture experiments. In S. Kotz and N. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, Volume 5, pp. 569–579. New York: Wiley.

Cornell, J. A. (1990). *Experiments with Mixtures* (Second ed.). New York: Wiley.

Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.

Crainiceanu, C. and D. Ruppert (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B 66*, 165–185.

Cressie, N. A. C. (1991). *Statistics for Spatial Data*. New York: Wiley.

Dale, T. B. (1992). Biological and chemical effects of acidified snowmelt on seasonal wetlands in Minnesota. Master's thesis, University of Minnesota.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two level experiments. *Technometrics 1*, 311–341.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*. New York: Wiley.

de Belleroche, J., A. Dick, and A. Wyrley-Birch (1982). Anticonvulsants and trifluoperazine inhibit the evoked release of GABA from cerebral cortex of rat at different sites. *Life Sciences 31*, 2875–2882.

Dickey, J. M. and B. P. Lientz (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a markov chain. *The Annals of Mathematical Statistics 41*, 214–226.

Draper, N. R., T. P. Davis, L. Pozueta, and D. M. Grove (1994). Isolation of degrees of freedom for Box-Behnken designs. *Technometrics 36*, 283–291.

Draper, N. R., S. M. Lewis, P. W. M. John, P. Prescott, A. M. Dean, and M. G. Tuck (1993). Mixture designs for four components in orthogonal blocks. *Technometrics 35*, 268–276.

Duncan, D. B. (1955). Multiple range and multiple F tests. *Biometrics 11*, 1–42.

Duncan, D. B. (1961). Bayes rule for a common multiple comparisons problem and related student-t problems. *Annals of Mathematical Statistics 32*, 1013–1033.

Dunnett, C. W. (1955). A multiple comparisons procedure for comparing several treatments with a control. *Journal of the American Statistical Association 50*, 1096–1121.

Dunnett, C. W. (1989). Algorithm AS 251: Multivariate normal probability integrals with product correlation structure. *Applied Statistics 38*, 564–579.

Durbin, J. (1960). Estimation of parameters in time-series regression. *Journal of the Royal Statistical Society, Series B 22*, 139–153.

Durbin, J. and G. S. Watson (1950). Testing for serial correlation in least squares regression I. *Biometrika 37*, 409–428.

Durbin, J. and G. S. Watson (1951). Testing for serial correlation in least squares regression II. *Biometrika 38*, 159–178.

Durbin, J. and G. S. Watson (1971). Testing for serial correlation in least squares regression III. *Biometrika 58*, 1–19.

Einot, I. and K. R. Gabriel (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association 70*, 574–583.

Fairfield Smith, H. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science 26*, 1–29.

Federer, W. T. and M. P. Meredith (1992). Covariance analysis for split-plot and split-block designs. *The American Statistician 46*, 155–162.

Fisher, R. A. (1935). *The Design of Experiments*. London: Oliver & Boyd, Ltd.

Fisher, R. A. and F. Yates (1963). *Statistical Tables for Biological, Agricultural, and Medical Research* (Sixth ed.). Edinburgh: Oliver and Boyd.

Franck, C. T., M. Nielsen, Dahlia, and J. A. Osborne (2013). A method for deteching hidden additivity in two-factor unreplicated experiments. *Computational Statistics and Data Analysis 67*, 95–104.

Freedman, D., R. Pisani, R. Purves, and A. Adhikari (1991). *Statistics*. New York: Norton.

Gayen, A. K. (1950). The distribution of the variance ratio in random samples of any size drawn from non-normal universes. *Biometrika 37*, 236–255.

Geary, R. C. (1970). Relative efficiency of count of sign changes for assessing residual autoregression in least squares regression. *Biometrika 57*, 123–127.

Gelman, A. and E. Loken (2014). The statistical crisis in science. *American Scientist 102*, 460–465.

Ghosh, B. K. and P. K. Sen (1991). *Handbook of Sequential Analysis*. New York: Marcel Dekker.

Giguere, V., G. Lefevre, and F. Labrie (1982). Site of calcium requirement for stimulation of ACTH release in rat anterior pituitary cells in culture by synthetic ovine corticotropin-releasing factor. *Life Sciences 31*, 3057–3062.

Golshani, T., E. Jorjani, S. Chelgani S. Chehreh, and N. Y. Heidari (2013). Modeling and process optimization for microbial desulfurization of coal by using a two-level full factorial design. *International Journal of Mining Science and Technology 23*, 261–265.

Greenhouse, S. W. and S. Geisser (1959). On methods in the analysis of profile data. *Psychometrika 24*, 95–112.

Greven, S., C. Crainiceanu, and H. Kuechenhoff (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics 17*, 870–891.

Grondona, M. O. and N. Cressie (1991). Using spatial considerations in the analysis of experiments. *Technometrics 33*, 381–392.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Hanley, J. A. and S. H. Shapiro (1994). Sexual activity and the lifespan of male fruitflies: A dataset that gets attention. *Journal of Statistics Education 2*(1), null.

Hareland, G. A. and M. A. Madson (1989). Barley dormancy and fatty acid composition of lipids isolated from freshly harvested and stored kernels. *Journal of the Institute of Brewing 95*, 437–442.

Hawkins, D. M. (1980). *Identification of Outliers*. London: Chapman and Hall.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Statistics 12*, 61–75.

Hedayat, A. and W. D. Wallis (1978). Hadamard matrices and their applications. *Annals of Statistics 6*, 1184–1238.

Henderson, Jr., C. R. (1982). Analysis of covariance in the mixed model: Higher-level, nonhomogeneous, and random regressions. *Biometrics 38*, 623–640.

Henderson, Jr., C. R. and C. R. Henderson (1979). Analysis of covariance in mixed models with unequal subclass numbers. *Communications in Statistics A 8*, 751–788.

Herr, D. G. (1986). On the history of ANOVA in unbalanced, factorial designs: The first 30 years. *The American Statistician 40*, 265–270.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey (Eds.) (1983). *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey (Eds.) (1991). *Fundamentals of Exploratory Analysis of Variance*. New York: Wiley.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika 75*, 800–802.

Hochberg, Y. and A. C. Tamhane (1987). *Multiple Comparison Procedures*. New York: Wiley.

Hocking, R. R. (1985). *The Analysis of Linear Models*. Monterey, CA: Brooks/Cole.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandanavian Journal of Statistics 6*, 65–70.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Hunt, J. R. and B. J. Larson (1990). Meal protein and zinc levels interact to influence zinc retention by the rat. *Nutrition Research 10*, 697–705.

Huynh, H. and L. S. Feldt (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. *Journal of the American Statistical Association 65*, 1582–1589.

Huynh, H. and L. S. Feldt (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics 1*, 69–82.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine 2.8 e124*, doi:10.1371/journal.pmed.0020124.

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In P. R. Krishnaiah (Ed.), *Handbook of Statistics*, Volume 1, pp. 199–236. Amsterdam: North Holland.

John, J. A. and E. R. Williams (1995). *Cyclic and Computer Generated Designs*. London: Chapman and Hall.

John, P. W. M. (1961). An application of a balanced incomplete block design. *Technometrics 3*, 51–54.

John, P. W. M. (1971). *Statistical Design and Analysis of Experiments*. New York: Macmillan.

Johnson, D. E. and F. A. Graybill (1972). Analysis of two-way model with interaction and no replication. *Journal of the American Statistical Association 67*, 862–868.

Johnson, N. L. and S. Kotz (1970). *Continuous Univariate Distributions*, Volume 1. New York: Wiley.

Joseph, L. and P. Bélisle (1997). Bayesian sample size determination for normal means and differences between normal means. *Journal of the Royal Statistical Society, Series D (The Statistician) 46*, 209–226.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kempthorne, O. (1952). *The Design and Analysis of Experiments*. New York: Wiley.

Kenward, M. G. and J. H. Roger (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics 53*, 983–997.

Keuls, M. (1952). The use of the studentized range in connection with an analysis of variance. *Euphytica 1*, 112–122.

Khuri, A. I. (1992). Response surface models with random block effects. *Technometrics 34*, 26–37.

Kiefer, J. (1958). On the nonrandomized optimality and randomized nonoptimality of symmetrical designs. *Annals of Mathematical Statistics 29*(3), 675–699.

Kim, K. and D. L. DeMets (1987). Design and analysis of group sequential tests based on the type i error spending rate function. *Biometrika 74*(1), 149–154.

Kordkandi, S. A. and M. Forouzesh (2014). Application of full factorial design for methylene blue dye removal using heat-activated persulfate oxidation. *Journal of the Taiwan Institute of Chemical Engineers 45*, 2597–2604.

Kurotori, I. S. (1966). Experiments with mixtures of components having lower bounds. *Industrial Quality Control*, May, 592–596.

Land, C. E. (1972). An evaluation of approximate confidence interval estimation methods for lognormal means. *Technometrics 14*, 145–158.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics 31*, 469–473.

Lindsey, J. K. (1997). Stopping rules and the likelihood function. *Journal of Statistical Planning and Inference 59*, 167–177.

Lohr, S. L. (1995). Hasse diagrams in statistical consulting and teaching. *American Statistician 49*, 376–381.

Low, C. K. and A. R. Bin Mohd. Ali (1985). Experimental tapping of pine oleoresin. *The Malaysian Forester 48*, 248–253.

Lund, R. E. and J. R. Lund (1983). Algorithm AS 190: Probabilities and upper quantiles for the studentized range. *Applied Statistics 32*, 204–210.

Lye, L. M. (2019). *Applications of DEO in Engineering and Science: A Collection of 2^6 Case Studies*. St. John's, Newfoundland: Leonard Lye.

Mandel, J. (1961). Non-additivity in two-way analysis of variance. *Journal of the American Statistical Association 56*, 878–888.

Mathew, T. and B. K. Sinha (1992). Exact and optimum tests in unbalanced split-plot designs under mixed and random models. *Journal of the American Statistical Association 87*, 192–200.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *Annals of Mathematical Statistics 11*, 204–209.

McCall, A. L., W. R. Millington, and R. J. Wurtman (1982). Blood-brain barrier transport of caffeine: dose-related restriction of adenine transport. *Life Sciences 31*, 2709–2715.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (Second ed.). London: Chapman and Hall.

McElhoe, H. B. and W. D. Conner (1986). Remote measurement of sulfur dioxide emissions using an ultraviolet light sensitive video system. *Journal of the Air Pollution Control Association 36*, 42–47.

Michicich, M. K. (1995). Sensory characteristics, consumer acceptance, and consumption of dairy products made from designer fats. Master's thesis, University of Minnesota.

Miller, A. (1997). Strip-plot configurations of fractional factorials. *Technometrics 39*, 153–161.

Miller, Jr., R. G. (1981). *Simultaneous Statistical Inference* (Second ed.). New York: Springer-Verlag.

Moore, D. and G. McCabe (1999). *Introduction to the Practice of Statistics* (Third ed.). New York: Freeman.

Morey, R. D., J. N. Rouder, and T. Jamil (2015). Bayesfactor: Computation of vayes factors for common designs.

Moses, L. E. (1987). Analysis of Bernoulli data from a 2^5 design done in blocks of size four. In C. Mallows (Ed.), *Design, Data, and Analysis*, pp. 275–290. New York: Wiley.

Mosteller, F., R. E. K. Rourke, and G. B. Thomas (1970). *Probability with Statistical Applications*. Reading, MA: Addison-Wesley.

Mosteller, F. and J. W. Tukey (1977). *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison-Wesley.

Myers, R. H., A. I. Khuri, and W. H. Carter, Jr. (1989). Response surface methodology: 1966-1988. *Technometrics 31*, 137–157.

Neath, A. A. and J. E. Cavanaugh (2006). A bayesian approach to the multiple comparisons problem. *Journal of Data Science 4*, 131–146.

Nelson, J. W. (1961). *The Nature of Wheat Lipids and Their Role in Flour Deterioration*. Ph. D. thesis, University of Minnesota, St. Paul, MN.

Nelson, P. R. (1993). Additional uses for the analysis of means and extended tables of critical values. *Technometrics 35*, 61–71.

Nelson, T. S., L. K. Kriby, and Z. B. Johnson (1990). Effect of minerals on the incidence of leg abnormalities in growing broiler chicks. *Nutrition Research 10*, 525–533.

Nelson, W. (1990). *Accelerated Testing*. New York: Wiley.

Newman, D. (1939). The distribution of the range in samples from a normal population, expressed in terms of an independent estimate of the standard deviation. *Biometrika 31*, 20–30.

O'Brien, P. C. and T. R. Fleming (1979). A multiple testing procedure for clinical trials. *Biometrics 35*, 549–556.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician 46*, 27–29.

Oehlert, G. W. (1994). Isolating one-cell interactions. *Technometrics 36*, 403–408.

Orman, B. A. (1986). Maize germination and seedling growth at suboptimal temperatures. Master's thesis, University of Minnesota, St. Paul, MN.

Park, S. H. (1978). Selecting contrasts among parameters in Scheffé's mixture models: Screening components and model reduction. *Technometrics 20*, 273–279.

Paskova, T. and C. Meyer (1997). Low-cycle fatigue of plain and fiber-reinforced concrete. *ACI Materials Journal 94*, 273–285.

Patterson, H. D. and E. R. Williams (1976). A new class of resolvable incomplete block designs. *Biometrika 63*, 83–92.

Patterson, H. D., E. R. Williams, and E. A. Hunter (1978). Block designs for variety trials. *Journal of Agricultural Science 90*, 395–400.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika 23*, 114–133.

Pierce, D. A. (1971). Least squares estimation in the regression model with autoregressive-moving average errors. *Biometrika 58*, 299–312.

Pignatiello, J. J. and J. S. Ramberg (1985). Comments on "Off-line quality control, parameter design, and the Taguchi method," by R. N. Kackar. *Journal of Quality Technology 17*, 198–206.

Pignatiello, J. J. and J. S. Ramberg (1991). Top ten triumphs and tragedies of Genichi Taguchi. *Quality Engineering 4*, 211–225.

Plackett, R. L. and J. P. Burman (1946). The design of optimum multifactorial experiments. *Biometrika 33*, 305–325.

Plummer, M. (2003, 04). Jags: A program for analysis of bayesian graphical models using gibbs sampling. *124*.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika 64*, 191–199.

Prato, H. H. and M. A. Morris (1984). Soil remaining on fabric after laundering as evaluated by response surface methodology. *Textile Research Journal 54*, 637–644.

Quinlan, J. (1985). Product improvement by application of Taguchi methods. In *American Supplier Institute News* (special symposium ed.), pp. 11–16. Dearborn, MI: American Supplier Institute.

Rey, D. K. (1981). *Characterization of the Effect of Solutes on the Water-Binding and Gel Strength Properties of Carrageenan*. Ph. D. thesis, University of Minnesota, St. Paul, MN.

Rey, W. J. J. (1983). *Introduction to Robust and Quasi-Robust Statistical Methods*. New York: Sprinter-Verlag.

Richards, J. A. (1965). Effects of fertilizers and management on three promising tropical grasses in Jamaica. *Expl. Agric. 1*, 281–288.

Ridzuan, N., F. Adam, and Z. Yaacob (2016). Screening of factor influencing wax deposition using full factorial experimental design. *Petroleum Science and Technology 34*, 84–90.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.

Rollag, M. D. (1982). Ability of tryptophan derivatives to mimic melatonin's action upon the Syrian hamster reproductive system. *Life Sciences 31*, 2699–2707.

Rosenbaum, P. R. and D. B. Rubin (1984). Sensitivity of bayes inference with data-dependent stopping rules. *American Statistician 38*, 106–109.

Rouder, J. N. (2014). Optional stopping: No problem for bayesians. *Psychonomic Bulletin and Review 21*, 301–308.

Ryan, T. A. (1960). Significance tests for multiple comparison of proportions, variances and other statistics. *Psychological Bulletin 57*, 318–328.

Sahrmann, H. F., G. F. Piepel, and J. A. Cornell (1987). In search of the optimum Harvey Wallbanger recipe via mixture experiment techniques. *The American Statistician 41*, 190–194.

Sanborn, A. N. and T. T. Hills (2013). The frequentist implications of optional stopping on bayesian hypothesis tests. *Psychonmetric Bulletin and Review 16*, 225–237.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics 2*, 110–114.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika 40*, 87–104.

Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.

Searle, S. R. (1971). Topics in variance component estimation. *Biometrics 27*, 1–76.

Searle, S. R., G. Casella, and C. E. McCulloch (1992). *Variance Components*. New York: Wiley.

Sellke, T., M. J. Bayarri, and J. O. Berger (1999). Calibration of p-values for testing precise null hypotheses. Technical report, Institute of Statistics and Decision Sciences, Duke University, Durham, NC.

Selwyn, M. R. and N. R. Hall (1984). On Bayesian methods for bioequivalence. *Biometrics 40*, 1103–1108.

Shaffer, J. P. (2007). Controlling the false discover rate with constraints: The newman-keuls test revisited. *Biometrical Journal 49*, 136–143.

Shah, M. and S. K. Garg (2014). Application of 2^k full factorial design in optimization for solvent-free microwave extraction of ginger essential oil. *Journal of Engineering* (828606), 5.

Shoemaker, A. C., K.-L. Tsui, and C. F. J. Wu (1991). Economical experimentation methods for robust design. *Technometrics 33*, 415–427.

Silvey, S. D. (1980). *Optimal Design: An Introduction to the Theory for Parameter Estimation*. London: Chapman and Hall.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751–754.

Simmons, J. P., L. D. Nelson, and U. Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science 22*, 1359–1366.

Simpson, J., A. Olsen, and J. C. Eden (1975). A Bayesian analysis of a multiplicative treatment effect in weather modification. *Technometrics 17*, 161–166.

Smith, J. R. and J. M. Beverly (1981). The use and analysis of staggered nested factorial designs. *Journal of Quality Technology 13*, 166–173.

Snedecor, G. W. and W. G. Cochran (1967). *Statistical Methods* (6th ed.). Ames, Iowa: Iowa State University Press.

Stan Development Team (2016). rstanarm: Bayesian applied regression modeling via stan. R package version 2.13.1.

Stan Development Team (2017). The stan core library. Version 2.17.0.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.17.3.

Straus, M. A., D. B. Sugarman, and J. Giles-Sims (1997). Spanking by parents and subsequent antisocial behavior of children. *Archives of Pediatrics and Adolescents Medicine 151*, 761–767.

Sutheerawattananonda, M. (1994). Variation in physical properties and microstructure of extruded wheat flours. Master's thesis, University of Minnesota, St. Paul, MN.

Swallow, W. H. and S. R. Searle (1978). Minimum variance quadratic unbiased estimation (MIVQUE) of variance components. *Technometrics 20*, 265–272.

Swanlund, D. J., M. R. N'Diaye, K. J. Loseth, J. L. Pryor, and B. G. Crabo (1995). Diverse testicular responses to exogenous growth hormone and follicle-stimulating hormone in prepubertal boars. *Biology of Reproduction 53*, 749–757.

Taam, W. and M. Hamada (1993). Detecting spatial effects from factorial experiments: An application from integrated-circuit manufacturing. *Technometrics 35*, 149–160.

Taguchi, G. and Y. Wu (1980). *Introduction to Off-Line Quality Engineering*. Nagoya, Japan: Central Japan Quality Control Association.

Tajima, A. (1987). *Some Aspects of Preserving Chicken Semen: Glycerol Effect, Assay Method, and Application*. Ph. D. thesis, University of Minnesota, St. Paul, MN.

Tjahjadi, C. (1983). *Isolation and Characterization of Adzuki Bean (Vigna angularis) Protein and Starch*. Ph. D. thesis, University of Minnesota, St. Paul, MN.

Tsay, R. (1984). Regression models with time series errors. *Journal of the American Statistical Association 79*, 118–124.

Tukey, J. W. (1952). Allowances for various types of error rates. Unpublished IMS address.

Tukey, J. W. (1956). Variances of variance components: I. Balanced designs. *Annals of Mathematical Statistics 27*, 722–736.

Tukey, J. W. (1957a). On the comparative anatomy of transformations. *Annals of Mathematical Statistics 28*, 602–632.

Tukey, J. W. (1957b). Variances of variance components: II. The unbalanced single classification. *Annals of Mathematical Statistics 28*, 43–56.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science 6*, 100–116.

US FDA (1997). FDA announces withdrawl of fenfluramine and dexfenfluramine. Press release P97-32.

Vangel, M. G. (1992). New methods for one-sided tolerance limits for a one-way balanced random-effects ANOVA model. *Technometrics 34*, 176–185.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing 27*, 1413–1432.

Wagenmakers, E.-J. (2007). Stopping rules and their irrelevance for bayesian inference: Online appendix to "a practical solution to the pervasive problems of *p*=values". `http://www.ejwagenmakers.com/2007/StoppingRuleAppendix.pdf`. Accessed: 2018–3-16.

Wald, A. (1947). *Sequential Analysis*. New York: Dover.

Waller, R. A. and D. B. Duncan (1969). A bayes rule for the symmetric multiple comparisons problem. *Journal of the American Statistical Association 64*, 1484–1503.

Wasserstein, R. L. and N. A. Lazar (2016). The asa's statement on p-values: Context, process, and purpose. *The American Statistician 70*(2), 129–133.

Wedin, D. A. (1990). *Nitrogen Cycling and Competition among Grass Species*. Ph. D. thesis, University of Minnesota, Minneapolis, MN.

Weisberg, S. (1985). *Applied linear regression* (Second ed.). New York: Wiley.

Welch, B. (1996). Effects of humidity on storing big sagebrush seed. Technical Report Research Paper INT-RP-493, USDA Forest Service, Intermountain Research Station.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association 72*, 566–575.

Westlake, W. J. (1974). The use of balanced incomplete block designs in comparative bioavailability trials. *Biometrics 30*, 319–327.

Whiting, K. R. (1990). *Host-Specific Pathogens and the Corn/Soybean Rotation Effect*. Ph. D. thesis, University of Minnesota, St. Paul, MN.

Windels, H. F. (1964). *The Influence of Diet and of Duration of Fast upon Plasma Levels of Free Leucine, Isoleucine, and Valine in the Growing Pig*. Ph. D. thesis, University of Minnesota, St. Paul, MN.

Wood, T. and F. H. Bormann (1974). Effects of an artificial acid mist upon the growth of *Betula alleghaninsis* britt. *Environmental Pollution 7*, 259–268.

Xhonga, R. (1971). Direct gold alloys—part II. *Journal of the American Academy of Gold Foil Operators 14*, 5–15.

Yates, F. (1936a). Incomplete randomized blocks. *Annals of Eugenics 7*, 121–140.

Yates, F. (1936b). A new method of arranging variety trials involving a large number of varieties. *Journal of Agricultural Science 26*, 424–455.

Yates, F. (1939). The recovery of inter-block information in variety trials arranged in three dimensional lattices. *Annals of Eugenics 9*, 136–156.

Yates, F. (1940). Lattice squares. *Journal of Agricultural Science 30*, 672–687.

Appendix A

Linear Models for Fixed Effects

Much of our analysis has used the Analysis of Variance, and we have approached ANOVA in a classical way, with lots of sums over indices i, j, and k. This approach is valid, but does not give insight into why ANOVA works or where the formulae come from. This appendix is meant as a *brief* introduction and survey of the theory of linear models for fixed effects. We can achieve a great deal of simplification and unity in our analysis approach through the use of linear models. Hocking (1985) is a good book-length reference for this material.

A.1   Models

Let $y \in \mathcal{R}^N$ be a vector of length N; y contains the responses in an experiment. A *model M* is a linear subspace of $\mathcal{R}^N$. For example, in a one-factor ANOVA the hypothesis of zero treatment effects corresponds to a model in $\mathcal{R}^N$ where all the vectors in M are constant vectors: $x \in M \leftrightarrow x = \mathbf{1}\beta$, where $\mathbf{1} = (1, 1, \ldots, 1)'$ is a vector of all ones. In a one-factor ANOVA, the hypothesis of k separate treatment means corresponds to a model in $\mathcal{R}^N$ where for any $x \in M$, the elements of x corresponding to the same treatment must all be the same, but the elements corresponding to different treatments can be different. Such a model can also be described as the range of a matrix $X_{N \times k}$, where $X_{i,j}$ is 1 if the ith response was in the jth treatment group, and zero otherwise. This means that $Y \in M$ can be written as $Y = X\beta$ for a k-vector β with elements interpreted $\mu_1, \mu_2, \ldots, \mu_k$. If $k = 3$; the treatment sample sizes were 2, 3, and 5; and the units were in treatment

order; then X could be written

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} .$$

There are many other matrices that span the same space, including:

(a) $\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$, (b) $\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$,

(c) $\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix}$, and (d) $\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -0.4 & -0.6 \\ 1 & -0.4 & -0.6 \\ 1 & -0.4 & -0.6 \\ 1 & -0.4 & -0.6 \\ 1 & -0.4 & -0.6 \end{bmatrix}$.

These matrices are shown because they illustrate the use of restrictions. For matrix (a), $Y \in M$ if $Y = X\beta$, where β is a 4-vector with elements interpreted $(\mu, \alpha_1, \alpha_2, \alpha_3)$. Recall that the separate means model is overparameterized if we don't put some kind of restrictions on the α_i's. This is what happens with matrix (a); if we add 100 to μ and subtract 100 from the α_i's, we get the same Y. Note that matrix (a) has 4 columns but only spans a subspace of dimension 3; matrix (a) is rank deficient.

To make the parameters unique, we need some restrictions. Some statistics programs assume that α_1 is zero and use μ, $\mu + \alpha_2$, and $\mu + \alpha_3$ as the treatment means. Thus α_2 is the difference in means between groups 2 and 1. Matrix (b) reflects this parameterization if we interpret the coefficients β as $(\mu, \alpha_2, \alpha_3)$.

One standard set of restrictions is that the treatment effects sum to 0, or equivalently, that $\alpha_g = -\sum_{i=1}^{g-1}\alpha_i$. Thus we may replace the last α_g with minus the sum of the others. Matrix (c) reflects this parameterization. For matrix (c), $Y \in M$ if $Y = X\beta$, where β is a 3-vector with elements interpreted $(\mu, \alpha_1, \alpha_2)$. The mean in the last treatment is $\mu - \alpha_1 - \alpha_2 = \mu + \alpha_3$.

Finally, a fourth possible set of restrictions is that the weighted sum of the treatment effects is 0, or equivalently, that $\alpha_g = -\sum_{i=1}^{g-1} n_i\alpha_i/n_g$. Matrix (d) reflects this parameterization. For matrix (d), $Y \in M$ if $Y = X\beta$, where β is a 3-vector with elements interpreted $(\mu, \alpha_1, \alpha_2)$. The mean in the last treatment is $\mu - n_1\alpha_1/n_3 - n_2\alpha_2/n_3 = \mu + \alpha_3$. Notice that the last two columns of matrix (d) are orthogonal to the first. This orthogonality is what makes the weighted-sum restrictions easier for hand work.

We arrange models in a lattice. A *lattice* is a partially ordered set in which every pair has a union and an intersection. For a lattice of models, the intersection is the largest submodel contained in both models (the intersection of the two model subspaces), and the union is the smallest (or simplest) model containing both submodels (the subspace spanned by the two models). The role of lattices in linear models is that it is easy to compare models up and down a lattice, but difficult to compare models if one model is not a subset of the other. Here is a sample lattice for a two-factor factorial:

Zero mean
|
Single mean

Row effects   Column effects

Additive model
|
Interactive model
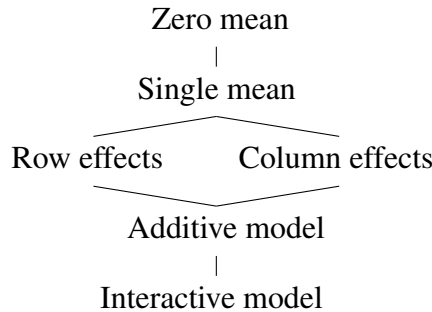
We can easily compare the "no row effects" model with the "interactive model," but it is more difficult to compare the "no row effects" model with the "no column effects" model. It should also be rather clear that lattice representations of several models and Hasse diagrams are related.

A.2   Least Squares

Suppose that we have a model M which is spanned by a matrix $X_{N\times r}$; thus $M = \mathcal{C}(X)$, where $\mathcal{C}(X)$ is the column space of X. We want to fit the model M to the data $\boldsymbol{y} \in \mathcal{R}^N$. This means we want to find the $Y \in M$ that is closest to $\boldsymbol{y}$. We measure closeness by the sum of squared errors: $(\boldsymbol{y} - Y)'(\boldsymbol{y} - Y)$. This is the same as finding the least squares regression of $\boldsymbol{y}$ on the r independent variables given by the columns of X. The minimum

occurs when

$$X'Xb = X'\boldsymbol{y} \ ,$$

(the normal equations), or when

$$X'(\boldsymbol{y} - Xb) = 0 \ .$$

The latter says that the residuals $(\boldsymbol{y} - Xb)$ are orthogonal to X, or equivalently, to $\mathcal{C}(X)$. The observations are then decomposed into the sum of fitted values Y and residuals $\boldsymbol{y} - Y$. This may be formalized as a theorem.

Theorem A.1 *For any $\boldsymbol{y} \in \mathcal{R}^N$ and any model $\boldsymbol{M} = \mathcal{C}(X_{N \times r})$, there exists a unique $Y \in \mathcal{C}(X)$ such that $\boldsymbol{y} - Y \perp \mathcal{C}(X)$. This Y is the least squares fit of the model $\boldsymbol{M}$ to $\boldsymbol{y}$. Y may be written as Xb for any b that solves the normal equations. If X has full rank, then b is unique and $b = (X'X)^{-1}X'\boldsymbol{y}$. If $\boldsymbol{M}$ is reparameterized to $\boldsymbol{M} = \mathcal{C}(X^\star)$ where $\mathcal{C}(X) = \mathcal{C}(X^\star)$, then Y remains the same, though the parameter estimates b may change.*

Look at Figure A.1; the triangle formed by Y_0, Y, and $\boldsymbol{y}$ will be a right triangle for any Y_0 in $\mathcal{C}(X)$, so the Pythagorean Theorem gives us the following for any $Y_0 \in \mathcal{C}(X)$:

$$(\boldsymbol{y} - Y_0)'(\boldsymbol{y} - Y_0) = (Y - Y_0)'(Y - Y_0) + (\boldsymbol{y} - Y)'(\boldsymbol{y} - Y) \ .$$

In particular, if we take Y_0 to be zero, this tells us that we may decompose the (uncorrected) total sum of squares in $\boldsymbol{y}$ into a model sum of squares $(Y - Y_0)'(Y - Y_0)$ and a residual sum of squares $(\boldsymbol{y} - Y)'(\boldsymbol{y} - Y)$. If the vector $\boldsymbol{1}$ lies in $\boldsymbol{M}$, then we may decompose the corrected total sum of squares in $\boldsymbol{y}$ into a model sum of squares around the overall mean $(Y - \overline{y}\boldsymbol{1})'(Y - \overline{y}\boldsymbol{1})$ and a residual sum of squares $(\boldsymbol{y} - Y)'(\boldsymbol{y} - Y)$.

We may revise the usual ANOVA terminology to reflect this geometric perspective. A source of variation is a model subspace. Variation of a certain type is variation that lies in a particular subspace. The degrees of freedom for a source or model is merely the dimension of the subspace. The sum of squares for a model (source) is the squared length of the part of $\boldsymbol{y}$ that lies in that subspace. The ANOVA table becomes (assuming that the model subspace has dimension r)

Source	DF	SS
Model subspace	Dimension of subspace	Squared length in subspace
Model $\boldsymbol{M}$	r	$Y'Y$
Deviations $\boldsymbol{M}^\perp$	$N - r$	$(\boldsymbol{y} - Y)'(\boldsymbol{y} - Y)$
Total $\mathcal{R}^N$	N	$\boldsymbol{y}'\boldsymbol{y}$
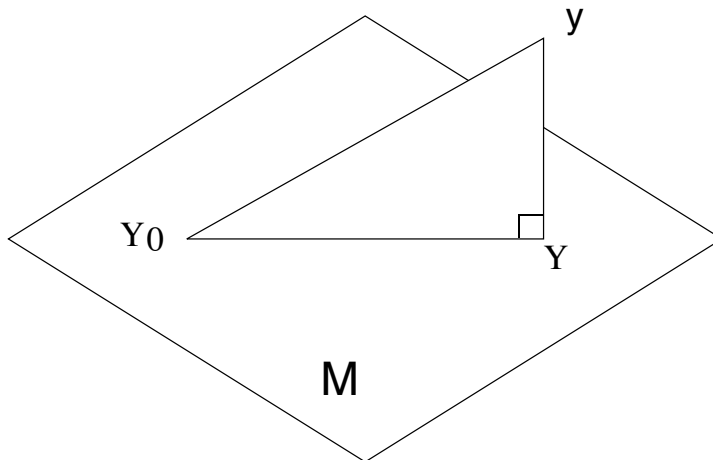
Figure A.1: Fitting a model.

We can also construct an ANOVA table for observations corrected for the grand mean, assuming that $1 \in M$, as is usually the case.

Source Subspace	DF Dimension	SS Squared length
Model corrected for grand mean $M \cap \mathbf{1}^{\perp}$	$r - 1$	$(Y - \bar{y}\mathbf{1})'(Y - \bar{y}\mathbf{1})$
Deviations $M^{\perp}$	$N - r$	$(\boldsymbol{y} - Y)'(\boldsymbol{y} - Y)$
Corrected total $\mathcal{R}^N \cap \mathbf{1}^{\perp}$	$N - 1$	$(\boldsymbol{y} - \bar{y}\mathbf{1})'(\boldsymbol{y} - \bar{y}\mathbf{1})$

A.3   Comparison of Models

Suppose that we have two models with $M_1 \cap M_2 = M_1$. Thus M_1 is above M_2 in the model lattice. If we have $M_1 = \mathcal{C}(X_1)$ and $M_2 = \mathcal{C}(X_2)$, then $M_1 \cap M_2 = M_1$ is equivalent to $\mathcal{C}(X_1) \subset \mathcal{C}(X_2)$. Let $\mathcal{C}(X_1)$ have dimension r_1, and let $\mathcal{C}(X_2)$ have dimension r_2. Y_1 is the fit of M_1 to $\boldsymbol{y}$, and Y_2 is the fit of M_2 to $\boldsymbol{y}$.

Look at Figure A.2. Not only is Y_1 the fit of M_1 to $\boldsymbol{y}$, Y_1 is the fit of M_1 to Y_2. We have right triangles everywhere we look.

Right angle	Right triangle
$(\boldsymbol{y} - Y_2) \perp M_2$	$(0, Y_2, \boldsymbol{y})$
$(\boldsymbol{y} - Y_1) \perp M_1$	$(0, Y_1, \boldsymbol{y})$
$(Y_2 - Y_1) \perp M_1$	$(0, Y_1, Y_2)$

Using these right triangles and the Pythagorean Theorem, we can make a variety of squared-length decompositions.

$$\boldsymbol{y}'\boldsymbol{y} \;=\; Y_2' Y_2 + (\boldsymbol{y} - Y_2)'(\boldsymbol{y} - Y_2)$$

$$\boldsymbol{y}'\boldsymbol{y} \;=\; Y_1' Y_1 + (\boldsymbol{y} - Y_1)'(\boldsymbol{y} - Y_1)$$

$$Y_2' Y_2 \;=\; Y_1' Y_1 + (Y_2 - Y_1)'(Y_2 - Y_1)$$

$$\boldsymbol{y}'\boldsymbol{y} \;=\; Y_1' Y_1 + (Y_2 - Y_1)'(Y_2 - Y_1) + (\boldsymbol{y} - Y_2)'(\boldsymbol{y} - Y_2)$$

$$(\boldsymbol{y} - Y_1)'(\boldsymbol{y} - Y_1) \;=\; (Y_2 - Y_1)'(Y_2 - Y_1) + (\boldsymbol{y} - Y_2)'(\boldsymbol{y} - Y_2)$$

In an Analysis of Variance, these squared-length decompositions are usually arranged as follows:
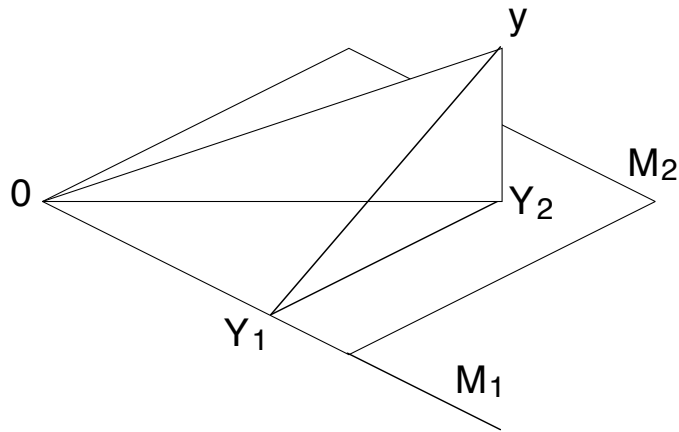
Figure A.2: Comparing two model fits.

Source Subspace	DF Dimension	SS Squared length
Model 1 M_1	r_1	$Y_1' Y_1$
Improvement of model 2 over model 1 $M_2 \cap M_1^\perp$	$r_2 - r_1$	$(Y_2 - Y_1)'(Y_2 - Y_1)$
Deviations $M_2^\perp$	$N - r_2$	$(y - Y_2)'(y - Y_2)$
Total $\mathcal{R}^N$	N	$y'y$

For example, consider model 1 to be the model of common means, $M_1 = \mathcal{C}(\mathbf{1})$, and model 2 to be the model of separate treatment means in a one-factor ANOVA. Then $M_1 \subset M_2$, because the separate treatment means could all be equal. We have $r_1 = 1$, and $r_2 = g$; thus the improvement in going from model 1 to model 2 is a $g - 1$ dimensional improvement. In the ANOVA, model 1 is usually called the constant or grand mean, and the improvement sum of squares going from model 1 to model 2 is called the between treatments sum of squares.

The parameterization in matrix (d) above is easier for hand work. It arises when we want to compute the sum of squares for the improvement of model 2 (g group means) over model 1 (common mean). This is the sum of squares for the orthogonal complement of model 1 in model 2. However, for matrix (d), the orthogonal complement of model 1 in model 2 is spanned by the last two columns of matrix (d). The orthogonality is built in.

We can, of course, extend model comparison to a series of three (or more) nested models: $M_1 \subset M_2 \subset M_3$. This gives an ANOVA table as follows:

Source Subspace	DF Dimension	SS Squared length
Model 1 M_1	r_1	$Y_1' Y_1$
Improvement of model 2 over model 1 $M_2 \cap M_1^{\perp}$	$r_2 - r_1$	$(Y_2 - Y_1)'(Y_2 - Y_1)$
Improvement of model 3 over model 2 $M_3 \cap M_2^{\perp}$	$r_3 - r_2$	$(Y_3 - Y_2)'(Y_3 - Y_2)$
Deviations $M_3^{\perp}$	$N - r_3$	$(\boldsymbol{y} - Y_3)'(\boldsymbol{y} - Y_3)$
Total $\mathcal{R}^N$	N	$\boldsymbol{y}'\boldsymbol{y}$

A.4 Projections

The *sum* of two subspaces U_1 and U_2 of a vector space V is $U_1 + U_2 = \{u_1 + u_2 : u_1 \in U_1, u_2 \in U_2\}$; $U_1 + U_2$ is also a subspace of V. If $U_1 \cap U_2 = \{0\}$, the sum is called *direct* and is written $U_1 \dot{+} U_2$. If V is the direct sum of U_1 and U_2, then $v \in V$ may be written uniquely as $v = u_1 + u_2$, where $u_1 \in U_1$ and $u_2 \in U_2$.

If V is the direct sum of U_1 and U_2 with $v \in V$ written as $v = u_1 + u_2$ ($u_1 \in U_1$, $u_2 \in U_2$), then the *projection of V onto U_1 parallel to U_2* is the
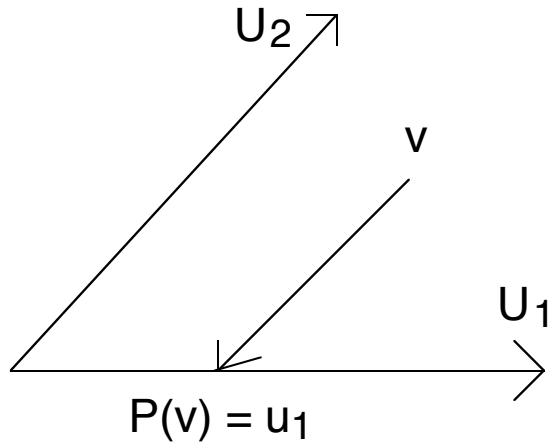
Figure A.3: Projection onto U_1 parallel to U_2.

linear map $P\colon V \to U_1$ given by $P(v) = u_1$. See Figure A.3. A linear mapping is a projection if and only if $P^2 = P$.

If two subspaces are orthogonal ($U_1 \perp U_2$), we write their direct sum as $U_1 \oplus U_2$ to emphasize their orthogonality. If $V = U_1 \oplus U_2$, then the projection of V onto U_1 is called an *orthogonal projection*.

Suppose we have a space $V = U_1 \oplus U_2$, with P_i being the orthogonal projection onto U_i. Then $P_1 P_2 = 0$. (Figure out why!) Furthermore, we have that since $v = u_1 + u_2$, then $v = P_1 v + P_2 v$, so that $(I - P_1) = P_2$.

Linear maps from $\mathcal{R}^N$ to $\mathcal{R}^N$ can be written as N by N matrices. Thus, we can express projections in $\mathcal{R}^N$ as matrices. The N by N matrix P is an orthogonal projection onto $U \in \mathcal{R}^N$ if and only if P is symmetric, idempotent (that is, $P^2 = P$), and $\mathcal{C}(P) = U$. If $U = \mathcal{C}(X)$ and X has full rank, then $P = X(X'\,X)^{-1}X'$.

What does all this have to do with linear models? If M is a model and P is the orthogonal projection onto M, then the fitted values for fitting M to y are Py. Least-squares fitting of models to data is simply the use of the orthogonal projection onto the model subspace.

Suppose we have two models M_1 and M_2, along with their union $M_{12} = M_1 \dot{+} M_2$. When does the sum of squares for M_{12} equal the sum of

squares for M_1 plus the sum of squares for M_2? By Pythagorean Theorem, the sum of squares for M_{12} is the sum of the sum of squares for M_1 and the sum of squares for $M_{12} \cap M_1^{\perp}$. This second model is M_2 if and only if model 2 is orthogonal to model 1, so the sums of squares add up if and only if the two original models are orthogonal.

How do we use this in ANOVA? We will have sums of squares that add up properly if we break $\mathcal{R}^N$ up into orthogonal subspaces. Our model lattices are hierarchical, with higher models including lower models. Thus to get orthogonal subspaces, we must look at the orthogonal complement of the smaller subspace in the larger subspace. This is the improvement in going from the smaller subspace to the larger subspace.

In the usual two-factor balanced ANOVA, the model of separate column means (M_C) is not orthogonal to the model of separate row means (M_R); these models have the constant-mean model as intersection. However, the model "improvement going from constant mean to separate column means" ($M_C \cap \mathbf{1}^{\perp}$) is orthogonal to the model "improvement going from constant mean to separate row means" ($M_R \cap \mathbf{1}^{\perp}$). This orthogonality is not present in the general unbalanced case.

When we have two nonorthogonal models, we will get different sums of squares if we decompose M_{12} as $M_1 \oplus M_{12} \cap M_1^{\perp}$ or $M_2 \oplus M_{12} \cap M_2^{\perp}$. The first corresponds to fitting model 1, and then getting the improvement going to M_{12}, and the second corresponds to fitting model 2, and then getting the improvement going to M_{12}. These have different projections in different orders. See Figure A.4. These changing subspaces are why sequential sums of squares (Type I) depend on order. Thus the sum of squares for B will not equal the sum of squares for B after A unless B and A represent orthogonal subspaces. The same applies for A and A after B.

A.5   Random Variation

So far, the linear models computations have not included any random variation, but we add that in. Our observations $y \in \mathcal{R}^N$ will have a normal distribution with mean μ and variance matrix Σ. The mean μ will lie in some model M. We usually assume that $\Sigma = \sigma^2 I$, where I is the N by N identity matrix. If y has the above distribution, then Cy (where C is a p by N matrix of constants) has a normal distribution with mean $C\mu$ and variance matrix $C\Sigma C'$.

Let's assume that $y \sim N(\mu, \sigma^2 I)$, where $\mu \in M$, and $M = \mathcal{C}(X)$ has dimension r. We can thus find a β (possibly infinitely many β's) such that $\mu = X\beta$. Let P be the orthogonal projection onto M; $(I - P)$ is thus the orthogonal projection onto $M^{\perp}$. The fitted values Y have the distribution

$$
\begin{aligned}
Y \;=\; Py \;\sim\;& N(P\mu,\ \sigma^2 PP') \\
=\;& N(\mu,\ \sigma^2 P) \\
=\;& N(X\beta,\ \sigma^2 P) \ .
\end{aligned}
$$
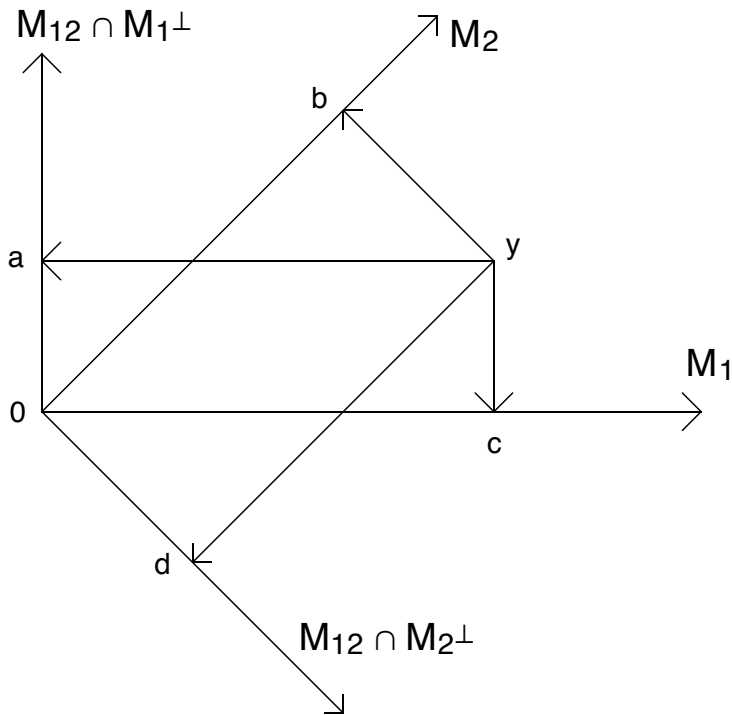
Figure A.4: Projecting in different orders.

The residuals have the distribution

$$\begin{aligned}
\boldsymbol{y} - Y \;=\; (I - P)\boldsymbol{y} \;\;\sim\;\; & N((I - P)\mu,\; \sigma^2(I - P)(I - P)') \\
=\;\; & N(0,\; \sigma^2(I - P)) \;\;.
\end{aligned}$$

These derivations give us the distributions of the fitted values and the residuals: they are both normal. However, we need to know their joint distribution. To discover this, we use a little trick and look at two copies of $\boldsymbol{y}$ just stacked into a vector of length $2N$, and we do separate projections on the

two copies.

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{y} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix}, \sigma^2 \begin{pmatrix} I & I \\ I & I \end{pmatrix} \right)$$

$$\begin{pmatrix} P & 0 \\ 0 & (I-P) \end{pmatrix} \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{y} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu} \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} P & P - P^2 \\ P - P^2 & I - P \end{pmatrix} \right)$$

$$\sim N\left(\begin{pmatrix} \boldsymbol{\mu} \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} P & 0 \\ 0 & I - P \end{pmatrix} \right)$$

This shows that the residuals and fitted values are uncorrelated. Because they are normally distributed, they are also independent.

How are the sums of squares distributed? Sums of squares are squared lengths, or quadratic forms, of normally distributed vectors. Normal vectors are easier to work with if they have a diagonal variance matrix, so let's work towards a diagonal variance matrix.

Let H_1 (N by r) be an orthonormal basis for $\boldsymbol{M}$; then $H_1'H_1$ is the r by r identity matrix. Let H_2 (N by $N-r$) be an orthonormal basis for $\boldsymbol{M}^\perp$; then $H_2'H_2$ is the $N-r$ by $N-r$ identity matrix. Furthermore, both $H_1'H_2$ and $H_2'H_1$ are 0. (The two matrices have columns that are bases for orthogonal subspaces; their columns must be orthogonal.) Now let H be the N by N matrix formed by joining H_1 and H_2 by $H = (H_1 : H_2)$. H is an orthogonal matrix, meaning that $H'H = HH' = I$.

The squared length of z and $H'z$ is the same for any $z \in \mathcal{R}^N$, because

$$z'z = z'Iz = zHH'z = (H'z)'(H'z)$$

So for sums of squares calculations, we may premultiply by H' before taking the squared length without changing the value or distribution.

Let's look at the residual sum of squares by looking at $H'(I-P)\boldsymbol{y}$.

$$H'(I-P)\boldsymbol{y} \sim N\left(\begin{pmatrix} H_1' \\ H_2' \end{pmatrix}(I-P)\boldsymbol{\mu},\, \sigma^2 \begin{pmatrix} H_1' \\ H_2' \end{pmatrix}(I-P)(H_1, H_2) \right)$$

$$\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix},\, \sigma^2 \begin{pmatrix} H_1' \\ H_2' \end{pmatrix}(0, H_2) \right)$$

$$\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix},\, \sigma^2 \begin{pmatrix} 0 & 0 \\ 0 & I_{N-r} \end{pmatrix} \right)$$

Thus the distribution of the sum of squared residuals is the same as the distribution of the sum of $N-r$ independent normals with mean 0 and variance σ^2. This is, of course, σ^2 times a chi-square distribution with $N-r$ degrees of freedom. The expected sum of squared errors is just $(N-r)\sigma^2$.

What about the model sum of squares? Look at $H'P\boldsymbol{y}$.

$$H'P\boldsymbol{y} \sim N\left(\left(\begin{array}{c} H_1' \\ H_2' \end{array}\right)P\boldsymbol{\mu},\ \sigma^2\left(\begin{array}{c} H_1' \\ H_2' \end{array}\right)P(H_1, H_2)\right)$$

$$\sim N\left(\left(\begin{array}{c} H_1'\boldsymbol{\mu} \\ 0 \end{array}\right),\ \sigma^2\left(\begin{array}{c} H_1' \\ H_2' \end{array}\right)(H_1, 0)\right)$$

$$\sim N\left(\left(\begin{array}{c} H_1'\boldsymbol{\mu} \\ 0 \end{array}\right),\ \sigma^2\left(\begin{array}{cc} I_r & 0 \\ 0 & 0 \end{array}\right)\right)$$

Thus the distribution of the model sum of squares is σ^2 times a noncentral chi-square with noncentrality parameter $\boldsymbol{\mu}'H_1H_1'\boldsymbol{\mu}/\sigma^2$ and r degrees of freedom. The noncentrality parameter $\boldsymbol{\mu}'H_1H_1'\boldsymbol{\mu}/\sigma^2$ also equals $\boldsymbol{\mu}'\boldsymbol{\mu}/\sigma^2$, so the expected model sum of squares is $\boldsymbol{\mu}'\boldsymbol{\mu} + r\sigma^2$. We may test the null hypothesis $H_0 : \boldsymbol{\mu} = 0$ against the alternative $H_a : \boldsymbol{\mu} \neq 0$ by taking the ratio of the model mean square to the error mean square; this ratio has an F-distribution under the null hypothesis and a noncentral F-distribution under the alternative.

We can generalize these distributional results to a sequence of models. Consider models $\boldsymbol{M}_1 = \mathcal{C}(X_1)$ and $\boldsymbol{M}_2 = \mathcal{C}(X_2)$ with $\boldsymbol{M}_1 \subset \boldsymbol{M}_2$. Let P_1 and P_2 be the orthogonal projections onto $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$. As usual, $\boldsymbol{\mu} \in \boldsymbol{M}_2$ is the expected value of $\boldsymbol{y}$; decompose $\boldsymbol{\mu}$ into $P_1\boldsymbol{\mu}$ and $(P_2 - P_1)\boldsymbol{\mu}$. These are the parts of the mean that lie in $\boldsymbol{M}_1$ and that are orthogonal to $\boldsymbol{M}_1$. Work with a pile of three copies of $\boldsymbol{y}$.

$$\left[\begin{array}{c} \boldsymbol{y} \\ \boldsymbol{y} \\ \boldsymbol{y} \end{array}\right] \sim N\left(\left[\begin{array}{c} \boldsymbol{\mu} \\ \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{array}\right],\ \sigma^2\left[\begin{array}{ccc} I & I & I \\ I & I & I \\ I & I & I \end{array}\right]\right)$$

$$\left[\begin{array}{ccc} P_1 & 0 & 0 \\ 0 & P_2 - P_1 & 0 \\ 0 & 0 & I - P_2 \end{array}\right]\left[\begin{array}{c} \boldsymbol{y} \\ \boldsymbol{y} \\ \boldsymbol{y} \end{array}\right] \sim N\left(\left[\begin{array}{c} P_1\boldsymbol{\mu} \\ (P_2 - P_1)\boldsymbol{\mu} \\ 0 \end{array}\right],\right.$$

$$\left. \sigma^2\left[\begin{array}{ccc} P_1 & 0 & 0 \\ 0 & P_2 - P_1 & 0 \\ 0 & 0 & I - P_2 \end{array}\right]\right)$$

Thus the fitted values Y_1, the difference in fitted values between the two models $Y_2 - Y_1$, and the residuals are all independent. The sum of squares for error is a multiple of chi-square with $N - r_2$ degrees of freedom. The improvement sum of squares going from the smaller to the larger model is a multiple of a chi-square with $r_2 - r_1$ degrees of freedom if the null is true $((P_2 - P_1)\boldsymbol{\mu} = 0)$; otherwise it is a multiple of a noncentral chi-square.

A.6   Estimable Functions

Assume that $\boldsymbol{y} = \boldsymbol{\mu} + \epsilon$, where $\boldsymbol{\mu} \in \boldsymbol{M} = \mathcal{C}(X)$ and $\epsilon \sim N(0, \sigma^2 I)$. Since $\boldsymbol{\mu} \in \mathcal{C}(X)$, we have that $\boldsymbol{\mu} = X\beta$ for some β. Let $Y = Xb$ be the projection of $\boldsymbol{y}$ onto $\boldsymbol{M}$.

A linear combination of the β's given by $h'\beta$ is *estimable* if there exists a vector $\boldsymbol{t} \in \mathcal{R}^N$ such that

$$E(\boldsymbol{t}'\boldsymbol{y}) = h'\beta,$$

for all values of β. Note that estimability is defined in terms of a particular set of parameters, so estimability depends on the matrix X, not just the model space $\boldsymbol{M}$. For $h'\beta$ to be estimable, we must have

$$h'\beta = E(\boldsymbol{t}'\boldsymbol{y}) = \boldsymbol{t}'E(\boldsymbol{y}) = \boldsymbol{t}'X\beta$$

for all β, so that

$$h = X'\boldsymbol{t} \ .$$

Thus $h'\beta$ is estimable if and only if $h = X'\boldsymbol{t}$, or in other words, if h is a linear combination of the rows of X.

We estimate $h'\beta$ by $h'b$, where b is any solution of the normal equations. There may be many solutions to the normal equations; is $h'b$ unique? Yes, it is unique because

$$h'b = \boldsymbol{t}'Xb = \boldsymbol{t}'Y \ ,$$

so the estimable function only depends on the fitted value Y. Note that $\boldsymbol{t}'y$ has the same expectation as $h'b$, but we will see below that $\boldsymbol{t}'y$ can have a larger variance.

What are the mean and variance of an estimable function? Let $\boldsymbol{t}^\star$ be the projection of $\boldsymbol{t}$ onto $\boldsymbol{M}$, and let $\boldsymbol{t} = \boldsymbol{t}^\star + \boldsymbol{t}_r$. Then

$$
\begin{aligned}
E(h'b) &= E(\boldsymbol{t}'\boldsymbol{y}) \\
&= E(\boldsymbol{t}^{\star\prime}\boldsymbol{y} + \boldsymbol{t}_r'\boldsymbol{y}) \\
&= \boldsymbol{t}^{\star\prime}X\beta + \boldsymbol{t}_r'X\beta \\
&= \boldsymbol{t}^{\star\prime}X\beta + 0\beta \\
&= \boldsymbol{t}^{\star\prime}X\beta
\end{aligned}
$$

So the expected value of $\boldsymbol{t}'\boldsymbol{y}$ only depends on the part of $\boldsymbol{t}$ that lies in $\boldsymbol{M}$. Variance is a bit trickier. If we directly attack $h'b$ we get

$$\mathrm{Var}(h'b) = \mathrm{Var}(\boldsymbol{t}'Y) = \sigma^2 \boldsymbol{t}'P\boldsymbol{t} = \sigma^2 \boldsymbol{t}^{\star\prime}\boldsymbol{t}^\star \ .$$

On the other hand, if we look at $\boldsymbol{t}'\boldsymbol{y}$, we find

$$\mathrm{Var}(\boldsymbol{t}'\boldsymbol{y}) = \sigma^2 \boldsymbol{t}'\boldsymbol{t} = \sigma^2(\boldsymbol{t}^{\star\prime}\boldsymbol{t}^\star + \boldsymbol{t}_r'\boldsymbol{t}_r) \ .$$

In the second version we only get minimum variance if $\boldsymbol{t}_r$ is 0. Because $\boldsymbol{t}_r$ does not affect expected value, we may restrict our attention to $\boldsymbol{t}$'s that lie

entirely in M; these will give us minimum variance no matter which way we use them.

Consider a one-factor model with g treatments, parameterized by μ and α_i, for $i = 1, 2, \ldots, g$. The ith treatment group has n_i observations and mean $\mu + \alpha_i$. The X matrix looks like

$$
\begin{array}{cccc}
\left.\begin{array}{cc} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{array}\right\} n_1 & \begin{array}{c} 0 \\ 0 \\ \vdots \\ 0 \end{array} & \cdots & \begin{array}{c} 0 \\ 0 \\ \vdots \\ 0 \end{array} \\
\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{array} & \left.\begin{array}{c} 1 \\ 1 \\ \vdots \\ 1 \end{array}\right\} n_2 & \cdots & \begin{array}{c} 0 \\ 0 \\ \vdots \\ 0 \end{array} \\
\begin{array}{cc} \vdots & \vdots \\ \end{array} & \begin{array}{c} \vdots \end{array} & \cdots & \begin{array}{c} \vdots \end{array} \\
\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{array} & \begin{array}{c} 0 \\ 0 \\ \vdots \\ 0 \end{array} & \cdots & \left.\begin{array}{c} 1 \\ 1 \\ \vdots \\ 1 \end{array}\right\} n_g
\end{array}
$$

For an estimable function given by a vector $t \in M$, the first n_1 elements of t are the same, the next n_2 are the same, and so on. Call these g unique values $s_1, s_2, \cdots, s_g$. An estimable h is of the form $h = X't$, and with this X, $X't$ leads to

$$
\begin{aligned}
h_\mu &= \sum_{i=1}^{g} n_i s_i \\
h_{\alpha_1} &= n_1 s_1 \\
h_{\alpha_2} &= n_2 s_2 \\
&\vdots \\
h_{\alpha_g} &= n_g s_g
\end{aligned}
$$

Thus for $h'\beta$ to be estimable, we only need to have that

$$
h_\mu = h_{\alpha_1} + h_{\alpha_2} + \cdots + h_{\alpha_g} \ .
$$

A.7   Contrasts

An estimable function $h'\beta$ for which the associated $t \in M$ satisfies $t'\mathbf{1} = 0$ is called a *contrast*. A contrast thus describes a direction $t \in M$ that is orthogonal to the grand mean. For the one-factor ANOVA problem, an estimable function is a contrast if

$$
0 = h_\mu = \sum_{i=1}^{g} n_i s_i = \sum_{i=1}^{g} h_{\alpha_i} \ .
$$

For contrasts, the overall mean must have a 0 coefficient, so we usually don't bother with a coefficient for μ at all, and denote the h_{α_i} by w_i.

Two contrasts are *orthogonal* if their corresponding t vectors are orthogonal:

$$
t \perp t^\star \Leftrightarrow 0 = \sum_{i=1}^{n} t_i t_i^\star = \sum_{i=1}^{g} n_i s_i s_i^\star = \sum_{i=1}^{g} \frac{w_i w_i^\star}{n_i} \ .
$$

M has r dimensions, so $M \cap \mathbf{1}^\perp$ has $r - 1$ dimensions. All contrasts lie in $M \cap \mathbf{1}^\perp$, so we can have at most $r - 1$ mutually orthogonal contrasts in a collection. These contrasts form an orthogonal basis for $M \cap \mathbf{1}^\perp$, and of course there are many such bases.

Every contrast determines a model $\mathcal{C}(t)$, and we may compute a sum of squares for this model via

$$
\mathrm{SS}(t) = \frac{(t'Y)^2}{t't} \ .
$$

We may do F-tests on this sum of squares exactly as we would on any model sum of squares. For a complete set of orthogonal contrasts $t_{(k)}$, we have

$$
M \cap \mathbf{1}^\perp = \mathcal{C}(t_{(1)}) \oplus \mathcal{C}(t_{(2)}) \oplus \cdots \oplus \mathcal{C}(t_{(r-1)})
$$

so that

$$
\mathrm{SS}(M \cap \mathbf{1}^\perp) = \mathrm{SS}(t_{(1)}) + \mathrm{SS}(t_{(2)}) + \cdots + \mathrm{SS}(t_{(r-1)}) \ .
$$

Alternatively, $t'y = h'b \sim N(h'\beta, \sigma^2 t't)$, so we may use t-style inference with the error mean square estimating σ^2. If $t't^\star = 0$, then $t'y$ and $t^{\star\prime}y$ are independent.

A.8   The Scheffé Method

How large can the sum of squares for a contrast be? The sum of squares for a contrast is the sum of squares for $\mathcal{C}(t)$, the model subspace spanned by the contrast. All contrast subspaces lie in $M \cap \mathbf{1}^\perp$, so we can make the decomposition

$$\mathrm{SS}(M \cap \mathbf{1}^\perp) = \mathrm{SS}(t) + \mathrm{SS}(M \cap \mathbf{1}^\perp \cap t^\perp) \ .$$

Thus the maximum that $\mathrm{SS}(t)$ could possibly be is $\mathrm{SS}(M \cap \mathbf{1}^\perp)$, which equals $(Y - \overline{Y}\mathbf{1})'(Y - \overline{Y}\mathbf{1})$. We can achieve this maximum by taking $t = (Y - \overline{Y}\mathbf{1})$:

$$
\begin{aligned}
\frac{(t'Y)^2}{t't} &= \frac{((Y - \overline{Y}\mathbf{1})'Y)^2}{(Y - \overline{Y}\mathbf{1})'(Y - \overline{Y}\mathbf{1})} \\
&= \frac{((Y - \overline{Y}\mathbf{1})'(Y - \overline{Y}\mathbf{1}))^2}{(Y - \overline{Y}\mathbf{1})'(Y - \overline{Y}\mathbf{1})} \\
&= (Y - \overline{Y}\mathbf{1})'(Y - \overline{Y}\mathbf{1}) \ .
\end{aligned}
$$

In a one-factor ANOVA, the maximum sum of squares for a contrast is the between groups sum of squares. Under the null hypothesis of no treatment differences, this sum of squares is distributed as σ^2 times a chi-square with $g - 1$ degrees of freedom. We do inference by comparing the F-ratio to the F distribution. Notice, however, that the maximal contrast sum of squares is equal to the treatment sum of squares. Thus we can do inference on arbitrarily many contrasts by treating them as if they were the maximal contrast. This is the basis for the Scheffé method of multiple comparisons.

A.9   Problems

Let y be an N by 1 random vector with $E\,y = X\beta$, and $Var(y) = \sigma^2 I_N$, where X is N by p and β is p by 1. Let $Y = Py$, where P is a projection (not necessarily orthogonal) onto the range of X. (a) Find the mean and (co)variance of Y and $y - Y$. (b) Prove that $\mathrm{Cov}(Y, y - Y)$ is 0 if and only if P is an orthogonal projection.   **Question A.1**

Let $y = X\beta + \epsilon$, where ϵ is $iid\ N(0, \sigma^2)$; y is N by 1, X is N by p, and β is p by 1. Let g be any N by 1 vector. What is the distribution of $(g'y)^2$? What, if anything, changes when $g'X$ is zero?   **Question A.2**

Consider a linear model $M = C(X)$ with parameters μ, β_1, β_2, and β_3, where X is as follows: **Question A.3**

$$
\begin{array}{cccc}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1
\end{array}
$$

Which of the following are estimable (give a brief reason): (a) μ, (b) β_1, (c) $\beta_2 - \beta_3$, (d) $\mu + (\beta_1 + \beta_2 + \beta_3)/3$, (e) $\beta_1 + \beta_2 - \beta_3$.

Consider a two by three factorial with proportional balance: $n_{ij} = n_{i\bullet} n_{\bullet j}/n_{\bullet\bullet}$. **Question A.4** Show that contrasts in factor A are orthogonal to contrasts in factor B.

Consider the following X matrices parameterizing models 1 and 2. **Question A.5**

$X1$		$X2$	
1	0	1	0
1	0	0	1
1	0	-1	-1
0	1	1	0
0	1	0	1
0	1	-1	-1
-1	-1	1	0
-1	-1	0	1
-1	-1	-1	-1

Let model 3 be the union of the models spanned by these two matrices. Will the sum of squares for model 3 be the sum of the sums of squares for models 1 and 2? Why or why not?

In the one-way ANOVA problem, show that the three restrictions $\sum \alpha_i = 0$, $\sum n_i \alpha_i = 0$, and $\alpha_1 = 0$ lead to the same values of $\alpha_1 - \alpha_2$. Interpret this result in terms of estimable functions.

Question A.6

Consider a one-factor model parameterized by the following matrix:

Question A.7

$$
\begin{array}{ccc}
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 0 & 1 \\
1 & -1 & -1 \\
1 & -1 & -1
\end{array}
$$

The parameters are μ, α_1, and α_2. Which of the following are estimable: (a) μ, (b) $\mu + \alpha_1$, (c) $\alpha_1 + \alpha_2$, (d) $\mu - \alpha_1$, and (e) $\alpha_1 - \alpha_2$?

Consider a completely randomized design with twelve treatments and 24 units (all $n_i = 2$). The twelve treatments have a three by four factorial structure.

Question A.8

(a) Find the variance/covariance matrix for the estimated factor A effects.

(b) Find the variance/covariance matrix for the estimated interaction effects.

(c) Show that the t-test for testing the equality of two factor A main effects can be found by treating the two estimated main effects as means of independent samples of size eight.

(d) Show that the t-test for testing the equality of two interaction effects can *not* be found by treating the two estimated interaction effects as means of independent samples of size two.

Consider the one-way ANOVA model with g groups. The sample sizes are n_i are not all equal. The treatments correspond to the levels of a quantitative factor; the level for treatment i is z_i, and the z_i are not equally spaced. We may compute linear, quadratic (adjusted for linear), and cubic (adjusted for linear and quadratic) sums of squares by linear regression. We may also compute these sums of squares via contrasts in the treatment means, but we need to find the contrast coefficients. Describe how to find the contrast coefficients for linear and quadratic (adjusted for linear). (Hint: use the t and s_i formulation in Sections A.6 and A.7, and remember your linear regression.)

Question A.9

Suppose that $Y_{N\times 1}$ is multivariate normal with mean μ and variance $\sigma^2 I$, and that we have models M_1 and M_2 with M_1 contained in M_2; M_1 has dimension r_1, M_2 has dimension r_2, and P_1 and P_2 are the orthogonal projections onto M_1 and M_2. **Question A.10**

(a) Find the distribution of $(P_2 - P_1)Y$.

(b) What can you say in addition about the distribution of $(P_2 - P_1)Y$ when μ lies in M_1?

Consider a proportionally balanced two-factor model with n_{ij} units in the ijth factor-level combination. Let M_A be the model of factor A effects ($Ey_{ijk} = \mu + \alpha_i$) and let M_B be the model of factor B effects ($Ey_{ijk} = \mu + \beta_j$). Show that $M_A \cap 1^\perp$ is orthogonal to $M_B \cap 1^\perp$. **Question A.11**

If X and $X^\star$ are n by p matrices and X has rank p, show that the range of X equals the range of $X^\star$ if and only if there exists a p by p nonsingular matrix Q such that $X^\star = XQ$. **Question A.12**

Appendix B

Experimental Design Plans

B.1   Latin Squares

The plans are presented in two groups. First we present sets of standard squares for several values of g. These sets are complete for $g = 3, 4$ and are incomplete for larger g. Next we present sets of up to four orthogonal Latin Squares (there are at most $g - 1$ orthogonal squares for any g). Graeco-Latin squares (and hyper-Latin squares) may be constructed by combining two (or more) orthogonal Latin Squares. All plans come from Fisher and Yates (1963).

B.1.1   Standard Latin Squares

3 × 3

```
A  B  C
B  C  A
C  A  B
```

4 × 4

```
A  B  C  D      A  B  C  D      A  B  C  D      A  B  C  D
B  A  D  C      B  C  D  A      B  D  A  C      B  A  D  C
C  D  B  A      C  D  A  B      C  A  D  B      C  D  A  B
D  C  A  B      D  A  B  C      D  C  B  A      D  C  B  A
```

5 × 5

```
A B C D E    A B C D E    A B C D E    A B C D E
B A E C D    B C E A D    B D A E C    B E A C D
C D A E B    C D B E A    C E D B A    C A D E B
D E B A C    D E A C B    D C E A B    D C E B A
E C D B A    E A D B C    E A B C D    E D B A C
```

6 × 6

```
A B C D E F    A B C D E F    A B C D E F
B C A F D E    B A E F C D    B A E C F D
C A B E F D    C F A B D E    C F B A D E
D F E B A C    D E B A F C    D E F B C A
E D F A C B    E D F C B A    E D A F B C
F E D C B A    F C D E A B    F C D E A B
```

7 × 7

```
A B C D E F G    A B C D E F G    A B C D E F G
B E A G F D C    B F E G C A D    B C D E F G A
C F G B D A E    C D A E B G F    C D E F G A B
D G E F B C A    D C G A F E B    D E F G A B C
E D B C A G F    E G B F A D C    E F G A B C D
F C D A G E B    F A D C G B E    F G A B C D E
G A F E C B D    G E F B D C A    G A B C D E F
```

B.1.2   Orthogonal Latin Squares

3 × 3

```
A B C    A B C
B C A    C A B
C A B    B C A
```

4 × 4

```
A B C D    A B C D    A B C D
B A D C    C D A B    D C B A
C D A B    D C B A    B A D C
D C B A    B A D C    C D A B
```

5 × 5

```
A  B  C  D  E      A  B  C  D  E      A  B  C  D  E      A  B  C  D  E
B  C  D  E  A      C  D  E  A  B      D  E  A  B  C      E  A  B  C  D
C  D  E  A  B      E  A  B  C  D      B  C  D  E  A      D  E  A  B  C
D  E  A  B  C      B  C  D  E  A      E  A  B  C  D      C  D  E  A  B
E  A  B  C  D      D  E  A  B  C      C  D  E  A  B      B  C  D  E  A
```

7 × 7

```
A  B  C  D  E  F  G      A  B  C  D  E  F  G      A  B  C  D  E  F  G
E  F  G  A  B  C  D      F  G  A  B  C  D  E      G  A  B  C  D  E  F
B  C  D  E  F  G  A      D  E  F  G  A  B  C      F  G  A  B  C  D  E
F  G  A  B  C  D  E      B  C  D  E  F  G  A      E  F  G  A  B  C  D
C  D  E  F  G  A  B      G  A  B  C  D  E  F      D  E  F  G  A  B  C
G  A  B  C  D  E  F      E  F  G  A  B  C  D      C  D  E  F  G  A  B
D  E  F  G  A  B  C      C  D  E  F  G  A  B      B  C  D  E  F  G  A
```

B.2 Balanced Incomplete Block Designs

The plans are sorted first by number of treatments g, then by size of block k. The number of blocks is b; the replication for any treatment is r; any pair of treatments occurs together in $\lambda = r(k-1)/(g-1)$ blocks; and the efficiency is $E = g(k-1)/[(g-1)k]$. Designs that can be arranged as Youden Squares are marked with YS and shown as Youden Squares. Designs involving all combinations of g treatments taken k at a time that cannot be arranged as Youden Squares are simply labeled *unreduced*. Some designs are generated as complements of other designs, that is, by including in one block all those treatments not appearing in the corresponding block of the other design. Additional plans can be found in Cochran and Cox (1957), who even include some plans with 91 treatments. Fisher and Yates (1963) describe methods for generating BIBD designs. BIBD plans given here were generated using the instructions in Fisher and Yates or de novo and then arranged in Youden Squares when feasible.

BIBD 1 g = 3, k = 2, b = 3, r = 2, λ = 1, E = .75, YS

$$\begin{array}{|c|c|c|} 1 & 2 & 3 \\ 2 & 3 & 1 \end{array}$$

BIBD 2 g = 4, k = 2, b = 6, r = 3, λ = 1, E = .67

Unreduced

BIBD 3 g = 4, k = 3, b = 4, r = 3, λ = 2, E = .89, YS

$$\begin{array}{|c|c|c|c|} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \end{array}$$

BIBD 4 g = 5, k = 2, b = 10, r = 4, λ = 1, E = .63, YS

1	1	4	5	2	5	3	3	4	2
2	3	1	1	4	2	4	5	5	3

BIBD 5 g = 5, k = 3, b = 10, r = 6, λ = 3, E = .83, YS

1	2	5	1	3	4	2	5	4	3
2	4	1	3	1	5	3	2	5	4
3	1	2	4	5	1	4	3	2	5

BIBD 6 g = 5, k = 4, b = 5, r = 4, λ = 3, E = .94, YS

1	2	3	4	5
2	3	4	5	1
3	4	5	1	2
4	5	1	2	3

BIBD 7 g = 6, k = 2, b = 15, r = 5, λ = 1, E = .6

Unreduced

BIBD 8 g = 6, k = 3, b = 10, r = 5, λ = 2, E = .8

1	2	3	5	5	6	4	1	5	6
4	4	4	6	6	1	1	2	2	3
5	6	5	1	2	3	2	3	3	4

BIBD 9 g = 6, k = 4, b = 15, r = 10, λ = 6, E = .9

Unreduced

BIBD 10 g = 6, k = 5, b = 6, r = 5, λ = 4, E = .96, YS

1	2	3	4	5	6
2	3	4	5	6	1
3	4	5	6	1	2
4	5	6	1	2	3
5	6	1	2	3	4

BIBD 11 g = 7, k = 2, b = 21, r = 6, λ = 1, E = .58, YS

1	1	1	5	6	7	3	4	2	2	2
2	3	4	1	1	1	2	2	5	6	7

3	3	6	7	5	4	4	5	7	6
4	5	3	3	4	6	7	6	5	7

BIBD 12 g = 7, k = 3, b = 7, r = 3, λ = 1, E = .78, YS

1	3	7	5	4	2	6
2	1	4	3	6	7	5
5	6	1	4	2	3	7

BIBD 13 g = 7, k = 4, b = 7, r = 4, λ = 2, E = .88, YS

3	1	2	7	6	5	4
4	2	7	1	5	6	3
6	7	4	5	3	1	2
7	6	5	3	2	4	1

BIBD 14 g = 7, k = 5, b = 21, r = 15, λ = 10, E = .93, YS

1	6	4	3	2	1	5	7	2	6	1	4	7	3	5
2	1	7	5	3	2	1	4	6	5	6	1	3	7	4
3	2	1	6	5	3	2	1	4	7	4	7	1	5	6
4	3	2	1	7	6	4	5	1	2	3	5	6	1	7
5	4	3	2	1	7	6	2	7	1	5	3	4	6	1

2	7	6	5	4	3
3	2	7	6	5	4
4	3	2	7	6	5
5	4	3	2	7	6
6	5	4	3	2	7

BIBD 15 g = 7, k = 6, b = 7, r = 6, λ = 5, E = .97, YS

1	2	3	4	5	6	7
2	3	4	5	6	7	1
3	4	5	6	7	1	2
4	5	6	7	1	2	3
5	6	7	1	2	3	4
6	7	1	2	3	4	5

BIBD 16 g = 8, k = 2, b = 28, r = 7, λ = 1, E = .57

Unreduced

BIBD 17 g = 8, k = 3, b = 56, r = 21, λ = 6, E = .76, YS

```
1 | 4 | 2 | 1 | 7 | 2 | 3 | 5 | 1 | 3 | 8 | 1 | 6 | 4 | 1
2 | 1 | 5 | 2 | 1 | 8 | 1 | 3 | 6 | 1 | 3 | 4 | 1 | 7 | 4
3 | 2 | 1 | 6 | 2 | 1 | 4 | 1 | 3 | 7 | 1 | 5 | 4 | 1 | 8

6 | 5 | 1 | 1 | 8 | 7 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 2 | 3
1 | 7 | 5 | 8 | 1 | 6 | 3 | 4 | 5 | 6 | 7 | 8 | 2 | 4 | 5
5 | 1 | 8 | 6 | 6 | 1 | 7 | 8 | 2 | 3 | 4 | 5 | 6 | 8 | 2

4 | 5 | 6 | 7 | 8 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 2 | 3 | 4
6 | 7 | 8 | 2 | 3 | 3 | 4 | 5 | 6 | 7 | 8 | 2 | 5 | 6 | 8
3 | 4 | 5 | 6 | 7 | 5 | 6 | 7 | 8 | 2 | 3 | 4 | 7 | 8 | 2

5 | 6 | 7 | 8 | 2 | 3 | 4 | 5 | 6 | 7 | 8
2 | 3 | 4 | 5 | 6 | 7 | 8 | 2 | 3 | 4 | 5
3 | 4 | 5 | 6 | 8 | 2 | 3 | 4 | 5 | 6 | 7
```

BIBD 18 g = 8, k = 4, b = 14, r = 7, λ = 3, E = .86

```
1 | 5 | 1 | 3 | 1 | 2 | 1 | 2 | 1 | 3 | 1 | 2 | 1 | 2
2 | 6 | 2 | 4 | 3 | 4 | 4 | 3 | 2 | 4 | 3 | 4 | 4 | 3
3 | 7 | 7 | 5 | 6 | 5 | 6 | 5 | 5 | 7 | 5 | 6 | 5 | 6
4 | 8 | 8 | 6 | 8 | 7 | 7 | 8 | 6 | 8 | 7 | 8 | 8 | 7
```

BIBD 19 g = 8, k = 5, b = 56, r = 35, λ = 20, E = .91, YS

```
1 | 6 | 4 | 3 | 2 | 1 | 5 | 7 | 2 | 6 | 1 | 4 | 7 | 3 | 5
2 | 1 | 7 | 5 | 3 | 2 | 1 | 4 | 6 | 5 | 6 | 1 | 3 | 7 | 4
3 | 2 | 1 | 6 | 5 | 3 | 2 | 1 | 4 | 7 | 4 | 7 | 1 | 5 | 6
4 | 3 | 2 | 1 | 7 | 6 | 4 | 5 | 1 | 2 | 3 | 5 | 6 | 1 | 7
5 | 4 | 3 | 2 | 1 | 7 | 6 | 2 | 7 | 1 | 5 | 3 | 4 | 6 | 1

2 | 7 | 6 | 5 | 4 | 3 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 1 | 2
3 | 2 | 7 | 6 | 5 | 4 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 8
4 | 3 | 2 | 7 | 6 | 5 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3
5 | 4 | 3 | 2 | 7 | 6 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4
6 | 5 | 4 | 3 | 2 | 7 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 5 | 6

3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 3 | 3
8 | 8 | 8 | 8 | 8 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4
4 | 5 | 6 | 7 | 1 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 4 | 5 | 6
5 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 8 | 8 | 8
7 | 1 | 2 | 3 | 4 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 5 | 6 | 7

4 | 5 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 6 | 7
5 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1
7 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3
8 | 8 | 8 | 8 | 6 | 7 | 1 | 2 | 3 | 4 | 5
1 | 2 | 3 | 4 | 8 | 8 | 8 | 8 | 8 | 8 | 8
```

BIBD 20 g = 8, k = 6, b = 28, r = 21, λ = 15, E = .95

Unreduced

BIBD 21 g = 8, k = 7, b = 8, r = 7, λ = 6, E = .98, YS

1	2	3	4	5	6	7	8
2	3	4	5	6	7	8	1
3	4	5	6	7	8	1	2
4	5	6	7	8	1	2	3
5	6	7	8	1	2	3	4
6	7	8	1	2	3	4	5
7	8	1	2	3	4	5	6

BIBD 22 g = 9, k = 2, b = 36, r = 8, λ = 1, E = .56, YS

1	1	1	1	6	7	8	9	3	4	5	2	2	2	2	7	8	9
2	3	4	5	1	1	1	1	2	2	2	6	7	8	9	8	7	9

3	3	3	7	8	9	5	6	4	4	4	5	5	8	9	7	6	6
4	5	6	3	3	3	4	4	7	8	9	6	7	5	5	6	8	9

BIBD 23 g = 9, k = 3, b = 12, r = 4, λ = 1, E = .75

1	4	7	1	2	3	1	2	3	1	2	3
2	5	8	4	5	6	6	4	5	5	6	4
3	6	9	7	8	9	8	9	7	9	7	8

BIBD 24 g = 9, k = 4, b = 18, r = 8, λ = 3, E = .84, YS

1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
2	3	4	5	6	7	8	9	1	4	5	6	7	8	9	1	2	3
3	4	5	6	7	8	9	1	2	6	7	8	9	1	2	3	4	5
5	6	7	8	9	1	2	3	4	9	1	2	3	4	5	6	7	8

BIBD 25 g = 9, k = 5, b = 18, r = 10, λ = 5, E = .9, YS

4	5	6	7	8	9	1	2	3	2	3	4	5	6	7	8	9	1
6	7	8	9	1	2	3	4	5	3	4	5	6	7	8	9	1	2
7	8	9	1	2	3	4	5	6	5	6	7	8	9	1	2	3	4
8	9	1	2	3	4	5	6	7	7	8	9	1	2	3	4	5	6
9	1	2	3	4	5	6	7	8	8	9	1	2	3	4	5	6	7

BIBD 26 g = 9, k = 6, b = 12, r = 8, λ = 5, E = .94

4	1	1	2	1	1	2	1	1	2	1	1
5	2	2	3	3	2	3	3	2	3	3	2
6	3	3	5	4	4	4	5	4	4	4	5
7	7	4	6	6	5	5	6	6	6	5	6
8	8	5	8	7	7	7	7	8	7	8	7
9	9	6	9	9	8	9	8	9	8	9	9

BIBD 27 g = 9, k = 7, b = 36, r = 28, λ = 21, E = .96, YS

3	4	5	6	7	8	9	1	2	2	3	4	5	6	7	8	9	1
4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3
5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4
6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5
7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6
8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7
9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8

2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1
3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2
5	6	7	8	9	1	2	3	4	4	5	6	7	8	9	1	2	3
6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5
7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6
8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7
9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8

BIBD 28 g = 9, k = 8, b = 9, r = 8, λ = 7, E = .98, YS

1	2	3	4	5	6	7	8	9
2	3	4	5	6	7	8	9	1
3	4	5	6	7	8	9	1	2
4	5	6	7	8	9	1	2	3
5	6	7	8	9	1	2	3	4
6	7	8	9	1	2	3	4	5
7	8	9	1	2	3	4	5	6
8	9	1	2	3	4	5	6	7

B.3 Efficient Cyclic Designs

Using this table you can generate an incomplete block design for g treatments in $b = mg$ blocks of size k with each treatment appearing $r = mk$ times. The design will be the union of m individual cyclic patterns, with these m patterns determined by the first m rows of this table for a given k. See John and Williams (1995).

k	r	First $k-1$ treatments	kth treatment, $g=$ 6	7	8	9	10	11	12	13	14	15
2	2	1	2	2	2	2	2	2	2	2	2	2
	4	1	3	4	4	4	4	4	4	6	5	5
	6	1	4	3	3	3	3	6	6	3	7	3
	8	1	6	5	5	5	5	3	3	5	4	8
	10	1	5	6	6	6	6	5	5	4	6	6
3	3	1 2	4	4	4	4	5	5	5	5	5	5
	6	1 3	2	4	8	7	8	8	6	8	8	9
	9	1 2	4	4	5	6	4	4	7	5	7	6
4	4	1 2 4	3	7	8	8	7	8	8	10	8	8
	8	1 2 5	3	7	8	9	3	7	7	7	7	7

k	r	First $k-1$ treatments	kth treatment, $g=$ 6	7	8	9	10	11	12	13	14	15
5	5	1 2 3 5	6	6	8	8	8	8	8	8	10	11
	10	1 3 4 5	6	6	8	9	10	9				
	10	1 3 4 7							8	12	13	11
6	6	1 2 3 4 7		6	6	6	6	11	11	11	11	11
7	7	1 2 3 4 5 8			6	6	10	10	10	10	10	11
8	8	1 2 3 4 5 7 9				6	10	10	10	10	12	12
9	9	1 2 3 4 5 6 8 10					9	9	9	11	11	11
10	10	1 2 3 4 5 6 7 10 11						8	8	8	13	13

B.4 Alpha Designs

Alpha Designs are resolvable block designs for $g = mk$ treatments in $b = mr$ blocks of size k. These tables give the initial alpha arrays for $5 \leq m \leq 15$, block sizes from 4 up to the minimum of m and $100/m$, and up to four replications. These tables are adapted from Table 2 of Patterson, Williams, and Hunter (1978).

$m = 5$			
$4 \leq k \leq 5$			
1	1	1	1
1	2	5	3
1	3	4	5
1	4	3	2
1	5	2	4

$m = 6$			
$4 \leq k \leq 6$			
1	1	1	1
1	2	6	5
1	4	3	6
1	3	4	2
1	5	2	3
1	6	2	4

$m = 7$			
$4 \leq k \leq 7$			
1	1	1	1
1	2	4	3
1	3	7	5
1	5	6	2
1	4	3	7
1	6	2	4
1	7	5	6

$m = 8$			
$4 \leq k \leq 8$			
1	1	1	1
1	2	3	7
1	4	8	2
1	6	4	5
1	3	6	4
1	5	2	7
1	7	1	3
1	8	7	6

$m = 9$			
$4 \leq k \leq 9$			
1	1	1	1
1	2	9	8
1	4	7	5
1	8	3	4
1	3	4	6
1	5	2	7
1	6	8	3
1	7	6	2
1	9	5	8

$m = 10$			
$4 \leq k \leq 10$			
1	1	1	1
1	2	10	6
1	4	7	10
1	6	8	3
1	5	6	7
1	7	4	2
1	8	3	5
1	9	5	8
1	10	9	3
1	3	7	4

$m = 11$			
$4 \leq k \leq 9$			
1	1	1	1
1	2	7	8
1	5	9	2
1	10	8	6
1	3	4	7
1	6	2	4
1	7	6	11
1	4	10	5
1	8	5	2

$m = 12$			
$4 \leq k \leq 8$			
1	1	1	1
1	2	3	4
1	8	6	2
1	10	7	5
1	5	12	9
1	12	4	11
1	11	5	8
1	6	2	7

$m = 13$			
$4 \leq k \leq 7$			
1	1	1	1
1	2	5	11
1	4	9	12
1	10	3	2
1	13	11	7
1	9	6	13
1	7	8	9

$m = 14$			
$4 \leq k \leq 7$			
1	1	1	1
1	2	9	11
1	10	11	8
1	12	14	3
1	3	7	2
1	6	12	13
1	4	2	12

$m = 15$			
$4 \leq k \leq 6$			
1	1	1	1
1	2	9	8
1	4	13	15
1	8	3	6
1	11	14	12
1	15	4	9

B.5   Two-Series Confounding and Fractioning Plans

The table gives suggested defining contrasts for confounding a 2^k design into 2^p blocks. It also gives the generalized interactions that are confounded. When only a particular block of the design is run, the resulting 2^{k-p} fractional factorial has aliases of I the same as the defining contrasts and their interactions. Other fractions have the same basic aliases, though the signs differ.

k	2^p	Defining contrasts	Generalized interactions
3	2	ABC	
	4	AB, BC	AC
4	2	ABCD	
	4	ABC, AD	BCD
	8	AB, BC, CD	AC, AD, BD, ABCD

k	2^p	Defining contrasts	Generalized interactions
5	2	ABCDE	
	4	ABCD, BCE	ADE
	8	ABC, BD, AE	ACD, BCE, ABDE, CDE
	16	AB, BC, CD, DE	AC, ABCD, BD, AD, ABDE, BCDE, ACDE, CE, ABCE, BE, AE
6	2	ABCDEF	
	4	BCDE, ABDF	ACEF
	8	ABCD, BCE, ACF	ADE, BDF, ABEF, CDEF
	16	CD, ACE, BCF, ABC	ADE, BDF, ABEF, ABCDEF, ABD, BE, BCDE, AF, ACDF, CEF, DEF
	32	AB, BC, CD, DE, EF	All other two-factor interactions, plus all four-factor and six-factor interactions
7	2	ABCDEFG	
	4	ADEF, ABCDG	BCEFG
	8	BCDE, ACDF, ABCG	ABEF, ADEG, BDFG, CEFG
	16	ABCD, BCE, ACF, ABG	ADE, BDF, ABEF, CDEF, CDG, ACEG, BDEG, BCFG, ADFG, EFG, ABCDEFG
	32	ADG, ACG, ABG, ABF, CEF	CD, BD, BC, ABCDG, BDFG, BCFG, ABCDF, FG, ADF, ACF, CDFG, ACDEFG, AEFG, DEF, ABCEFG, BCDEF, BEF, ABDEFG, ABCE, BCDEG, BEG, ABDE, CEG, ACDE, AE, DEG
	64	AB, BC, CD, DE, EF, FG	All other two-factor interactions, plus all four-factor and six-factor interactions
8	2	ABCDEFGH	
	4	ABDFG, BCDEH	ACEFGH
	8	BCEG, BCDH, ACDEF	DEGH, ABDFG, ABEFH, ACFGH
	16	BCDE, ACDF, ABDG, ABCH	ABEF, ACEG, BCFG, DEFG, ADEH, BDFH, CEFH, CDGH, BEGH, AFGH, ABCDEFGH

k	2^p	Defining contrasts	Generalized interactions
8	32	ABD, ACE, BCF, ABCG, ABCH	BCDE, ACDF, ABEF, DEF, CDG, BEG, ADEG, AFG, BDFG, CEFG, ABCDEFG, CDH, BEH, ADEH, AFH, BDFH, CEFH, ABCDEFH, GH, ABDGH, ACEGH, BCDEGH, BCFGH, ACDFGH, ABEFGH, DEFGH
	64	AG, BF, BCE, AEF, BDG, ADH	ABFG, ABCEG, CEF, ACEFG, EFG, ABE, BEG, ABCF, BCFG, AC, CG, ABD, DFG, ADF, CDEG, ACDE, BCDEFG, ABCDEF, ABDEFG, BDEF, ADEG, DE, ACDFG, CDF, ABCDG, BCD, DGH, ABDFH, BDFGH, ABCDEH, BCDEGH, ACDEFH, CDEFGH, DEFH, ADEFGH, BDEH, ABDEGH, BCDFH, ABCDFGH, CDH, ACDGH, ABGH, BH, AFGH, FH, ACEGH, CEH, ABCEFGH, BCEFH, BEFGH, ABEFH, EGH, AEH, CFGH, ACFH, BCGH, ABCH

Appendix C

Tables

All table values were computed in MacAnova.

Table C.1: Random digits.

```
68094  23539  18913  86955  39327  02225  69423  06689  99791  76722
01909  10889  72439  61293  21529  36388  14555  95914  25254  38422
81253  33731  00873  30545  50227  94749  07761  77740  19743  21724
20501  57876  10081  07431  91817  25296  52198  75278  45922  19728
30557  32116  68368  18292  37433  27636  92360  74374  00155  19623
91740  24671  12987  73192  97251  12516  38695  12790  63529  58111
08388  48988  91806  24777  61809  84551  29619  26471  87362  05818
76006  06178  10765  76938  42086  66950  90720  88483  66611  19710
72600  85770  88793  66291  41081  61031  60104  02545  86041  62345
32209  77328  41324  68614  57322  94583  07415  27313  26322  93218
38420  57120  12268  15017  44456  90919  73640  69974  61200  82209
49690  34002  11553  49387  44354  92179  79960  61804  70374  71782
85210  59681  38002  41958  90125  02819  78165  44800  17792  96272
35229  78839  46776  00944  67288  59471  23715  05753  87214  06758
78568  94584  71728  81741  38433  59390  57344  27554  90465  95245
00679  26121  29667  83237  67154  10246  33005  72851  34876  29007
15398  98457  22406  30927  90111  14065  51246  18592  85397  92122
89014  44909  62227  24503  59774  69233  29556  14126  26810  67044
84538  98456  19149  54714  36332  89999  02248  26089  77989  98072
33618  91123  84227  34110  74523  73244  27365  89167  02035  90366
48194  17487  33892  64522  69065  98755  49765  90609  57786  31991
54929  29666  72716  59146  86232  38765  33335  35127  71464  69505
13639  16775  89564  73978  73321  63868  65447  15689  37789  22178
28420  16687  25081  99131  15641  59055  11472  31110  58669  49621
57905  96871  07126  01978  06563  18504  80138  96710  51019  13183
36490  13154  96356  90278  47401  47783  14283  47107  43874  73050
15852  60522  54438  97802  18869  06219  62244  67309  21556  62034
28614  54310  58953  24393  09880  69588  34399  19114  17086  19286
92594  10130  04030  12348  62118  35368  11032  28513  38832  49642
10119  22185  14692  59461  98941  51851  82728  60066  75060  48027
27970  68214  84216  82761  54280  98276  48123  50611  11562  44945
83423  24025  55539  30343  44943  79061  54400  09157  08448  81417
91821  56637  02232  65331  24585  58902  70981  84902  30673  66372
56385  90995  94482  90187  15461  78394  38276  07567  17556  42504
45081  92518  67475  26920  36524  67476  11973  65938  74470  80782
87655  77363  79749  74171  35109  51652  32671  47315  50862  24683
77287  08196  64511  04557  45941  87701  00805  64707  43178  32760
60633  66288  95791  18232  14346  80974  50836  21944  24407  95112
03089  42195  14802  55732  92821  48338  27293  61239  70050  83121
10570  71691  04943  33707  35118  06278  28534  79418  85857  52665
```

Table C.1: Random digits, continued.

30263	25135	17075	56131	64430	43573	77506	09510	65985	17159
13811	98464	48063	98483	60748	07379	89540	07699	60560	93391
80280	46665	54480	90895	94555	77376	55074	69674	22124	86546
96302	09821	31198	06423	69016	71408	48673	22035	92401	40242
34922	65539	17012	69492	97661	66351	94296	00451	99255	98999
81090	48413	74876	24165	42912	58517	51494	80415	28758	96355
67224	24891	38160	78489	73226	95368	19123	78424	47010	44371
63204	25405	51831	00562	23640	97596	73613	31668	81299	13975
39678	79440	84900	06251	93120	57470	68970	82673	88484	93689
30374	19502	99804	25596	07763	02914	05334	52321	74595	47068
06813	76019	12479	03459	51078	44527	02086	01367	26591	69118
57097	14846	92151	95357	73479	53708	04442	30282	82320	99043
09521	48055	19823	82346	38890	31327	98995	37520	73670	48277
77991	19227	65802	92645	13378	06593	52303	15173	98557	43631
47605	33709	36996	22976	78611	39221	95962	06137	72056	44395
29969	01292	47429	28477	72881	83330	57842	96953	66190	29761
26978	10916	24087	68880	42657	93404	74540	22069	56907	53591
43115	41945	85148	43539	19452	69583	88827	22232	52494	19895
51493	62141	57091	26829	61899	03433	04983	85869	31376	31307
57731	27002	19954	12314	10234	99589	59101	28150	65083	85057
37816	75263	68459	32095	15844	20352	46919	82419	59487	78779
65009	90859	76655	46234	24073	93183	85770	60190	69870	44997
89443	17030	30366	18026	64815	64790	24439	24153	75360	85068
19978	11146	54195	18001	39458	50082	47801	79655	11199	00978
69137	35105	62192	60958	32109	00787	79202	74700	27231	39559
00102	19753	27900	16409	42548	81604	16881	03009	62624	94651
86465	06647	56974	45774	38612	54604	35113	14259	08609	86134
74692	64914	61361	55581	79265	85121	94402	66705	02455	63518
25531	67924	61704	95032	48824	40759	83063	89562	74811	42721
87057	63223	84910	27744	36979	00578	63738	47473	66356	59676
22723	61335	89609	98968	78238	94353	11790	62264	78866	86637
61837	60095	22904	83603	57362	85576	24298	25868	08558	17143
07208	30664	53006	15714	92246	91157	97898	43295	26162	85001
09265	97806	06556	70909	24791	81907	92463	80405	32493	57985
60079	09778	70500	69276	16192	39024	42519	69661	59750	15740
11620	30055	59498	63231	90667	12729	99405	17906	20684	65483
20210	31650	23408	32631	87779	62148	03322	98071	41217	03952
91935	61772	67324	44921	75176	32383	21611	23145	51109	13168
15449	91085	09246	06833	93677	60567	20180	59763	01650	41798
33759	00216	03782	18185	98508	07890	02365	50624	55194	85954
59706	03210	55372	71993	55247	40554	12783	36287	19884	58491

Table C.2: Tail areas for the standard normal distribution.

Table entries are $\mathcal{E} = P(Z > z_\mathcal{E}) = 1 - \Phi(z_\mathcal{E})$.

$z_\mathcal{E}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024

Table C.3: Percent points for the Student t distribution.

Table entries are $t_{\mathcal{E},\nu}$ where $P_\nu(t > t_{\mathcal{E},\nu}) = \mathcal{E}$.

ν	.2	.1	.05	.025	$\mathcal{E}$ .01	.005	.001	.0005	.0001
1	1.376	3.078	6.314	12.71	31.82	63.66	318.3	636.6	3183
2	1.061	1.886	2.920	4.303	6.965	9.925	22.33	31.60	70.70
3	.978	1.638	2.353	3.182	4.541	5.841	10.22	12.92	22.20
4	.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610	13.03
5	.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869	9.678
6	.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959	8.025
7	.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7.063
8	.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041	6.442
9	.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781	6.010
10	.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587	5.694
11	.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437	5.453
12	.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318	5.263
13	.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221	5.111
14	.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140	4.985
15	.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073	4.880
16	.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015	4.791
17	.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965	4.714
18	.862	1.330	1.734	2.101	2.552	2.878	3.610	3.922	4.648
19	.861	1.328	1.729	2.093	2.539	2.861	3.579	3.883	4.590
20	.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850	4.539
21	.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819	4.493
22	.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792	4.452
23	.858	1.319	1.714	2.069	2.500	2.807	3.485	3.768	4.415
24	.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745	4.382
25	.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725	4.352
26	.856	1.315	1.706	2.056	2.479	2.779	3.435	3.707	4.324
27	.855	1.314	1.703	2.052	2.473	2.771	3.421	3.690	4.299
28	.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674	4.275
29	.854	1.311	1.699	2.045	2.462	2.756	3.396	3.659	4.254
30	.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646	4.234
35	.852	1.306	1.690	2.030	2.438	2.724	3.340	3.591	4.153
40	.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551	4.094
45	.850	1.301	1.679	2.014	2.412	2.690	3.281	3.520	4.049
50	.849	1.299	1.676	2.009	2.403	2.678	3.261	3.496	4.014
60	.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460	3.962

Table C.4: Percent points for the chi-square distribution.

Table entries are $\chi^2_{\mathcal{E},\nu}$ where $P_\nu(\chi^2 > \chi^2_{\mathcal{E},\nu}) = \mathcal{E}$.

ν	.995	.99	.975	$\mathcal{E}$ .95	.05	.025	.01	.005
1	.000039	.00016	.0010	.0039	3.841	5.024	6.635	7.879
2	.0100	.0201	.0506	.1026	5.991	7.378	9.210	10.60
3	.0717	.1148	.2158	.3518	7.815	9.348	11.34	12.84
4	.2070	.2971	.4844	.7107	9.488	11.14	13.28	14.86
5	.4117	.5543	.8312	1.145	11.07	12.83	15.09	16.75
6	.6757	.8721	1.237	1.635	12.59	14.45	16.81	18.55
7	.9893	1.239	1.690	2.167	14.07	16.01	18.48	20.28
8	1.344	1.646	2.180	2.733	15.51	17.53	20.09	21.95
9	1.735	2.088	2.700	3.325	16.92	19.02	21.67	23.59
10	2.156	2.558	3.247	3.940	18.31	20.48	23.21	25.19
11	2.603	3.053	3.816	4.575	19.68	21.92	24.72	26.76
12	3.074	3.571	4.404	5.226	21.03	23.34	26.22	28.30
13	3.565	4.107	5.009	5.892	22.36	24.74	27.69	29.82
14	4.075	4.660	5.629	6.571	23.68	26.12	29.14	31.32
15	4.601	5.229	6.262	7.261	25.00	27.49	30.58	32.80
16	5.142	5.812	6.908	7.962	26.30	28.85	32.00	34.27
17	5.697	6.408	7.564	8.672	27.59	30.19	33.41	35.72
18	6.265	7.015	8.231	9.390	28.87	31.53	34.81	37.16
19	6.844	7.633	8.907	10.12	30.14	32.85	36.19	38.58
20	7.434	8.260	9.591	10.85	31.41	34.17	37.57	40.00
21	8.034	8.897	10.28	11.59	32.67	35.48	38.93	41.40
22	8.643	9.542	10.98	12.34	33.92	36.78	40.29	42.80
23	9.260	10.20	11.69	13.09	35.17	38.08	41.64	44.18
24	9.886	10.86	12.40	13.85	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	43.77	46.98	50.89	53.67
35	17.19	18.51	20.57	22.47	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	79.08	83.30	88.38	91.95

Table C.5: Percent points for the F distribution.

Table entries are $F_{.05,\nu_1,\nu_2}$ where $P_{\nu_1,\nu_2}(F > F_{.05,\nu_1,\nu_2}) = .05$.

ν_2	ν_1 1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.62	8.59
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.50	4.46
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.81	3.77
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.38	3.34
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.04
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.83
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.70	2.66
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.57	2.53
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.47	2.43
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.38	2.34
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.31	2.27
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.25	2.20
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.19	2.15
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.15	2.10
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.11	2.06
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	2.04	1.99
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.02	1.98	1.94
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.96	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.97	1.94	1.89
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.84	1.79
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.74	1.69
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.87	1.78	1.73	1.69	1.63
75	3.97	3.12	2.73	2.49	2.34	2.22	2.13	2.06	2.01	1.96	1.88	1.80	1.71	1.65	1.61	1.55
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.77	1.68	1.62	1.57	1.52
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.72	1.62	1.56	1.52	1.46
∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.51	1.46	1.40

Table C.5: Percent points for the F distribution, continued.

Table entries are $F_{.01,\nu_1,\nu_2}$ where $P_{\nu_1,\nu_2}(F > F_{.01,\nu_1,\nu_2}) = .01$.

ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.45	9.38	9.29
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.30	7.23	7.14
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.06	5.99	5.91
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.26	5.20	5.12
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.71	4.65	4.57
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.31	4.25	4.17
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.01	3.94	3.86
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.76	3.70	3.62
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.57	3.51	3.43
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.41	3.35	3.27
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.28	3.21	3.13
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.16	3.10	3.02
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.07	3.00	2.92
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.98	2.92	2.84
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.91	2.84	2.76
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.84	2.78	2.69
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.79	2.72	2.64
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.73	2.67	2.58
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.69	2.62	2.54
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.64	2.58	2.49
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.60	2.54	2.45
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.45	2.39	2.30
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.27	2.20	2.11
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.42	2.27	2.17	2.10	2.01
75	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57	2.43	2.29	2.13	2.03	1.96	1.87
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.22	2.07	1.97	1.89	1.80
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.27	2.13	1.97	1.87	1.79	1.69
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.77	1.70	1.59

Table C.5: Percent points for the F distribution, continued.

Table entries are $F_{.001,\nu_1,\nu_2}$ where $P_{\nu_1,\nu_2}(F > F_{.001,\nu_1,\nu_2}) = .001$.

								ν_1								
ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40
2	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999	999
3	167	149	141	137	135	133	132	131	130	129	128	127	126	126	125	125
4	74.1	61.2	56.2	53.4	51.7	50.5	49.7	49.0	48.5	48.1	47.4	46.8	46.1	45.7	45.4	45.1
5	47.2	37.1	33.2	31.1	29.8	28.8	28.2	27.6	27.2	26.9	26.4	25.9	25.4	25.1	24.9	24.6
6	35.5	27.0	23.7	21.9	20.8	20.0	19.5	19.0	18.7	18.4	18.0	17.6	17.1	16.9	16.7	16.4
7	29.2	21.7	18.8	17.2	16.2	15.5	15.0	14.6	14.3	14.1	13.7	13.3	12.9	12.7	12.5	12.3
8	25.4	18.5	15.8	14.4	13.5	12.9	12.4	12.0	11.8	11.5	11.2	10.8	10.5	10.3	10.1	9.92
9	22.9	16.4	13.9	12.6	11.7	11.1	10.7	10.4	10.1	9.89	9.57	9.24	8.90	8.69	8.55	8.37
10	21.0	14.9	12.6	11.3	10.5	9.93	9.52	9.20	8.96	8.75	8.45	8.13	7.80	7.60	7.47	7.30
11	19.7	13.8	11.6	10.3	9.58	9.05	8.66	8.35	8.12	7.92	7.63	7.32	7.01	6.81	6.68	6.52
12	18.6	13.0	10.8	9.63	8.89	8.38	8.00	7.71	7.48	7.29	7.00	6.71	6.40	6.22	6.09	5.93
13	17.8	12.3	10.2	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.52	6.23	5.93	5.75	5.63	5.47
14	17.1	11.8	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	6.13	5.85	5.56	5.38	5.25	5.10
15	16.6	11.3	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.81	5.54	5.25	5.07	4.95	4.80
16	16.1	11.0	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81	5.55	5.27	4.99	4.82	4.70	4.54
17	15.7	10.7	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58	5.32	5.05	4.78	4.60	4.48	4.33
18	15.4	10.4	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39	5.13	4.87	4.59	4.42	4.30	4.15
19	15.1	10.2	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22	4.97	4.70	4.43	4.26	4.14	3.99
20	14.8	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.82	4.56	4.29	4.12	4.00	3.86
21	14.6	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95	4.70	4.44	4.17	4.00	3.88	3.74
22	14.4	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.58	4.33	4.06	3.89	3.78	3.63
23	14.2	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73	4.48	4.23	3.96	3.79	3.68	3.53
24	14.0	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64	4.39	4.14	3.87	3.71	3.59	3.45
25	13.9	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56	4.31	4.06	3.79	3.63	3.52	3.37
30	13.3	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	4.00	3.75	3.49	3.33	3.22	3.07
40	12.6	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.64	3.40	3.14	2.98	2.87	2.73
50	12.2	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67	3.44	3.20	2.95	2.79	2.68	2.53
75	11.7	7.58	6.01	5.16	4.62	4.24	3.96	3.74	3.56	3.42	3.19	2.96	2.71	2.55	2.44	2.29
100	11.5	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30	3.07	2.84	2.59	2.43	2.32	2.17
200	11.2	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26	3.12	2.90	2.67	2.42	2.26	2.15	2.00
∞	10.8	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96	2.74	2.51	2.27	2.10	1.99	1.84

Table C.6: Coefficients of orthogonal polynomial contrasts.

g	Order	Coefficients						
		1	2	3	4	5	6	7
3	1	-1	0	1				
	2	1	-2	1				
4	1	-3	-1	1	3			
	2	1	-1	-1	1			
	3	-1	3	-3	1			
5	1	-2	-1	0	1	2		
	2	2	-1	-2	-1	2		
	3	-1	2	0	-2	1		
	4	1	-4	6	-4	1		
6	1	-5	-3	-1	1	3	5	
	2	5	-1	-4	-4	-1	5	
	3	-5	7	4	-4	-7	5	
	4	1	-3	2	2	-3	1	
	5	-1	5	-10	10	-5	1	
7	1	-3	-2	-1	0	1	2	3
	2	5	0	-3	-4	-3	0	5
	3	-1	1	1	0	-1	-1	1
	4	3	-7	1	6	1	-7	3
	5	-1	4	-5	0	5	-4	1
	6	1	-6	15	-20	15	-6	1

Table C.7: Critical values for the two-sided Bonferroni t statistic.

Table entries are $t_{\mathcal{E},\nu}$ where $P_\nu(t > t_{\mathcal{E},\nu}) = \mathcal{E}$ and $\mathcal{E} = .05/2/K$.

ν	2	3	4	5	6	7	8	9	10	15	20	30	50
1	25.5	38.2	50.9	63.7	76.4	89.1	102	115	127	191	255	382	637
2	6.21	7.65	8.86	9.92	10.9	11.8	12.6	13.4	14.1	17.3	20.0	24.5	31.6
3	4.18	4.86	5.39	5.84	6.23	6.58	6.90	7.18	7.45	8.58	9.46	10.9	12.9
4	3.50	3.96	4.31	4.60	4.85	5.07	5.26	5.44	5.60	6.25	6.76	7.53	8.61
5	3.16	3.53	3.81	4.03	4.22	4.38	4.53	4.66	4.77	5.25	5.60	6.14	6.87
6	2.97	3.29	3.52	3.71	3.86	4.00	4.12	4.22	4.32	4.70	4.98	5.40	5.96
7	2.84	3.13	3.34	3.50	3.64	3.75	3.86	3.95	4.03	4.36	4.59	4.94	5.41
8	2.75	3.02	3.21	3.36	3.48	3.58	3.68	3.76	3.83	4.12	4.33	4.64	5.04
9	2.69	2.93	3.11	3.25	3.36	3.46	3.55	3.62	3.69	3.95	4.15	4.42	4.78
10	2.63	2.87	3.04	3.17	3.28	3.37	3.45	3.52	3.58	3.83	4.00	4.26	4.59
11	2.59	2.82	2.98	3.11	3.21	3.29	3.37	3.44	3.50	3.73	3.89	4.13	4.44
12	2.56	2.78	2.93	3.05	3.15	3.24	3.31	3.37	3.43	3.65	3.81	4.03	4.32
13	2.53	2.75	2.90	3.01	3.11	3.19	3.26	3.32	3.37	3.58	3.73	3.95	4.22
14	2.51	2.72	2.86	2.98	3.07	3.15	3.21	3.27	3.33	3.53	3.67	3.88	4.14
15	2.49	2.69	2.84	2.95	3.04	3.11	3.18	3.23	3.29	3.48	3.62	3.82	4.07
16	2.47	2.67	2.81	2.92	3.01	3.08	3.15	3.20	3.25	3.44	3.58	3.77	4.01
17	2.46	2.65	2.79	2.90	2.98	3.06	3.12	3.17	3.22	3.41	3.54	3.73	3.97
18	2.45	2.64	2.77	2.88	2.96	3.03	3.09	3.15	3.20	3.38	3.51	3.69	3.92
19	2.43	2.63	2.76	2.86	2.94	3.01	3.07	3.13	3.17	3.35	3.48	3.66	3.88
20	2.42	2.61	2.74	2.85	2.93	3.00	3.06	3.11	3.15	3.33	3.46	3.63	3.85
21	2.41	2.60	2.73	2.83	2.91	2.98	3.04	3.09	3.14	3.31	3.43	3.60	3.82
22	2.41	2.59	2.72	2.82	2.90	2.97	3.02	3.07	3.12	3.29	3.41	3.58	3.79
23	2.40	2.58	2.71	2.81	2.89	2.95	3.01	3.06	3.10	3.27	3.39	3.56	3.77
24	2.39	2.57	2.70	2.80	2.88	2.94	3.00	3.05	3.09	3.26	3.38	3.54	3.75
25	2.38	2.57	2.69	2.79	2.86	2.93	2.99	3.03	3.08	3.24	3.36	3.52	3.73
26	2.38	2.56	2.68	2.78	2.86	2.92	2.98	3.02	3.07	3.23	3.35	3.51	3.71
27	2.37	2.55	2.68	2.77	2.85	2.91	2.97	3.01	3.06	3.22	3.33	3.49	3.69
28	2.37	2.55	2.67	2.76	2.84	2.90	2.96	3.00	3.05	3.21	3.32	3.48	3.67
29	2.36	2.54	2.66	2.76	2.83	2.89	2.95	3.00	3.04	3.20	3.31	3.47	3.66
30	2.36	2.54	2.66	2.75	2.82	2.89	2.94	2.99	3.03	3.19	3.30	3.45	3.65
35	2.34	2.51	2.63	2.72	2.80	2.86	2.91	2.96	3.00	3.15	3.26	3.41	3.59
40	2.33	2.50	2.62	2.70	2.78	2.84	2.89	2.93	2.97	3.12	3.23	3.37	3.55
45	2.32	2.49	2.60	2.69	2.76	2.82	2.87	2.91	2.95	3.10	3.20	3.35	3.52
50	2.31	2.48	2.59	2.68	2.75	2.81	2.85	2.90	2.94	3.08	3.18	3.32	3.50
100	2.28	2.43	2.54	2.63	2.69	2.75	2.79	2.83	2.87	3.01	3.10	3.23	3.39
∞	2.24	2.39	2.50	2.58	2.64	2.69	2.73	2.77	2.81	2.94	3.02	3.14	3.29

Table C.7: Critical values for the two-sided Bonferroni t statistic, continued.

Table entries are $t_{\mathcal{E},\nu}$ where $P_\nu(t > t_{\mathcal{E},\nu}) = \mathcal{E}$ and $\mathcal{E} = .01/2/K$.

ν	2	3	4	5	6	7	8	9	10	15	20	30	50
K													
1	127	191	255	318	382	446	509	573	637	955	1273	1910	3183
2	14.1	17.3	20.0	22.3	24.5	26.4	28.3	30.0	31.6	38.7	44.7	54.8	70.7
3	7.45	8.58	9.46	10.2	10.9	11.5	12.0	12.5	12.9	14.8	16.3	18.7	22.2
4	5.60	6.25	6.76	7.17	7.53	7.84	8.12	8.38	8.61	9.57	10.3	11.4	13.0
5	4.77	5.25	5.60	5.89	6.14	6.35	6.54	6.71	6.87	7.50	7.98	8.69	9.68
6	4.32	4.70	4.98	5.21	5.40	5.56	5.71	5.84	5.96	6.43	6.79	7.31	8.02
7	4.03	4.36	4.59	4.79	4.94	5.08	5.20	5.31	5.41	5.80	6.08	6.50	7.06
8	3.83	4.12	4.33	4.50	4.64	4.76	4.86	4.96	5.04	5.37	5.62	5.97	6.44
9	3.69	3.95	4.15	4.30	4.42	4.53	4.62	4.71	4.78	5.08	5.29	5.60	6.01
10	3.58	3.83	4.00	4.14	4.26	4.36	4.44	4.52	4.59	4.85	5.05	5.33	5.69
11	3.50	3.73	3.89	4.02	4.13	4.22	4.30	4.37	4.44	4.68	4.86	5.12	5.45
12	3.43	3.65	3.81	3.93	4.03	4.12	4.19	4.26	4.32	4.55	4.72	4.96	5.26
13	3.37	3.58	3.73	3.85	3.95	4.03	4.10	4.16	4.22	4.44	4.60	4.82	5.11
14	3.33	3.53	3.67	3.79	3.88	3.96	4.03	4.09	4.14	4.35	4.50	4.71	4.99
15	3.29	3.48	3.62	3.73	3.82	3.90	3.96	4.02	4.07	4.27	4.42	4.62	4.88
16	3.25	3.44	3.58	3.69	3.77	3.85	3.91	3.96	4.01	4.21	4.35	4.54	4.79
17	3.22	3.41	3.54	3.65	3.73	3.80	3.86	3.92	3.97	4.15	4.29	4.47	4.71
18	3.20	3.38	3.51	3.61	3.69	3.76	3.82	3.87	3.92	4.10	4.23	4.42	4.65
19	3.17	3.35	3.48	3.58	3.66	3.73	3.79	3.84	3.88	4.06	4.19	4.36	4.59
20	3.15	3.33	3.46	3.55	3.63	3.70	3.75	3.80	3.85	4.02	4.15	4.32	4.54
21	3.14	3.31	3.43	3.53	3.60	3.67	3.73	3.78	3.82	3.99	4.11	4.28	4.49
22	3.12	3.29	3.41	3.50	3.58	3.64	3.70	3.75	3.79	3.96	4.08	4.24	4.45
23	3.10	3.27	3.39	3.48	3.56	3.62	3.68	3.72	3.77	3.93	4.05	4.21	4.42
24	3.09	3.26	3.38	3.47	3.54	3.60	3.66	3.70	3.75	3.91	4.02	4.18	4.38
25	3.08	3.24	3.36	3.45	3.52	3.58	3.64	3.68	3.73	3.88	4.00	4.15	4.35
26	3.07	3.23	3.35	3.43	3.51	3.57	3.62	3.67	3.71	3.86	3.97	4.13	4.32
27	3.06	3.22	3.33	3.42	3.49	3.55	3.60	3.65	3.69	3.84	3.95	4.11	4.30
28	3.05	3.21	3.32	3.41	3.48	3.54	3.59	3.63	3.67	3.83	3.94	4.09	4.28
29	3.04	3.20	3.31	3.40	3.47	3.52	3.58	3.62	3.66	3.81	3.92	4.07	4.25
30	3.03	3.19	3.30	3.39	3.45	3.51	3.56	3.61	3.65	3.80	3.90	4.05	4.23
35	3.00	3.15	3.26	3.34	3.41	3.46	3.51	3.55	3.59	3.74	3.84	3.98	4.15
40	2.97	3.12	3.23	3.31	3.37	3.43	3.47	3.51	3.55	3.69	3.79	3.92	4.09
45	2.95	3.10	3.20	3.28	3.35	3.40	3.44	3.48	3.52	3.66	3.75	3.88	4.05
50	2.94	3.08	3.18	3.26	3.32	3.38	3.42	3.46	3.50	3.63	3.72	3.85	4.01
120	2.87	3.01	3.1	3.17	3.23	3.28	3.32	3.36	3.39	3.51	3.60	3.72	3.86
∞	2.81	2.94	3.02	3.09	3.14	3.19	3.23	3.26	3.29	3.40	3.48	3.59	3.72

Table C.8: Percent points for the Studentized range.

Table entries are $q_{.05}(K, \nu)$.

ν	2	3	4	5	6	7	8	9	10	15	20	30	50
1	18.0	27.0	32.8	37.1	40.4	43.1	45.4	47.4	49.1	55.4	59.6	65.1	71.7
2	6.09	8.33	9.80	10.9	11.7	12.4	13.0	13.5	14.0	15.7	16.8	18.3	20.0
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	10.5	11.2	12.2	13.4
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.66	9.23	10.0	10.9
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.72	8.21	8.87	9.67
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	7.14	7.59	8.19	8.91
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.76	7.17	7.73	8.40
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.48	6.87	7.40	8.03
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	6.28	6.64	7.14	7.75
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	6.11	6.47	6.95	7.53
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.98	6.33	6.79	7.35
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.88	6.21	6.66	7.21
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.79	6.11	6.55	7.08
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.71	6.03	6.46	6.98
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.65	5.96	6.38	6.89
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.59	5.90	6.31	6.81
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.54	5.84	6.25	6.74
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.50	5.79	6.20	6.68
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.46	5.75	6.15	6.63
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.43	5.71	6.10	6.58
21	2.94	3.56	3.94	4.21	4.42	4.60	4.74	4.87	4.98	5.40	5.68	6.07	6.53
22	2.93	3.55	3.93	4.20	4.41	4.58	4.72	4.85	4.96	5.37	5.65	6.03	6.49
23	2.93	3.54	3.91	4.18	4.39	4.56	4.70	4.83	4.94	5.34	5.62	6.00	6.45
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.32	5.59	5.97	6.42
25	2.91	3.52	3.89	4.15	4.36	4.53	4.67	4.79	4.90	5.30	5.57	5.94	6.39
26	2.91	3.51	3.88	4.14	4.35	4.51	4.65	4.77	4.88	5.28	5.55	5.92	6.36
27	2.90	3.51	3.87	4.13	4.33	4.50	4.64	4.76	4.86	5.26	5.53	5.89	6.34
28	2.90	3.50	3.86	4.12	4.32	4.49	4.62	4.74	4.85	5.24	5.51	5.87	6.31
29	2.89	3.49	3.85	4.11	4.31	4.47	4.61	4.73	4.84	5.23	5.49	5.85	6.29
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	5.21	5.47	5.83	6.27
35	2.87	3.46	3.81	4.07	4.26	4.42	4.56	4.67	4.77	5.15	5.41	5.76	6.18
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	5.11	5.36	5.70	6.11
45	2.85	3.43	3.77	4.02	4.21	4.36	4.49	4.61	4.70	5.07	5.32	5.66	6.06
50	2.84	3.42	3.76	4.00	4.19	4.34	4.47	4.58	4.68	5.04	5.29	5.62	6.02
100	2.81	3.36	3.70	3.93	4.11	4.26	4.38	4.48	4.58	4.92	5.15	5.46	5.83
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.80	5.01	5.30	5.65

Table C.8: Percent points for the Studentized range, continued.

Table entries are $q_{.01}(K, \nu)$.

ν	2	3	4	5	6	7	8	9	10	15	20	30	50
1	90.2	135	164	186	202	216	227	237	246	277	298	326	359
2	14.0	19.0	22.3	24.7	26.6	28.2	29.5	30.7	31.7	35.4	38.0	41.3	45.3
3	8.27	10.6	12.2	13.3	14.2	15.0	15.6	16.2	16.7	18.5	19.8	21.4	23.4
4	6.51	8.12	9.17	9.96	10.6	11.1	11.5	11.9	12.3	13.5	14.4	15.6	17.0
5	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.2	11.2	11.9	12.9	14.0
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.95	10.5	11.3	12.3
7	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	9.12	9.65	10.4	11.2
8	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.55	9.03	9.68	10.5
9	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49	8.13	8.57	9.18	9.91
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.81	8.23	8.79	9.49
11	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.56	7.95	8.49	9.15
12	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	7.36	7.73	8.25	8.87
13	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	7.19	7.55	8.04	8.65
14	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	7.05	7.39	7.87	8.46
15	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.93	7.26	7.73	8.29
16	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.82	7.15	7.60	8.15
17	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.73	7.05	7.49	8.03
18	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.65	6.97	7.40	7.92
19	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.58	6.89	7.31	7.83
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.52	6.82	7.24	7.74
21	4.00	4.61	4.99	5.26	5.47	5.65	5.79	5.92	6.04	6.47	6.76	7.17	7.67
22	3.99	4.59	4.96	5.22	5.43	5.61	5.75	5.88	5.99	6.42	6.71	7.11	7.60
23	3.97	4.57	4.93	5.20	5.40	5.57	5.72	5.84	5.95	6.37	6.66	7.05	7.53
24	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.33	6.61	7.00	7.48
25	3.94	4.53	4.89	5.14	5.35	5.51	5.65	5.78	5.89	6.29	6.57	6.95	7.42
26	3.93	4.51	4.87	5.12	5.32	5.49	5.63	5.75	5.86	6.26	6.53	6.91	7.37
27	3.92	4.49	4.85	5.10	5.30	5.46	5.60	5.72	5.83	6.22	6.50	6.87	7.33
28	3.91	4.48	4.83	5.08	5.28	5.44	5.58	5.70	5.80	6.20	6.47	6.84	7.29
29	3.90	4.47	4.81	5.06	5.26	5.42	5.56	5.67	5.78	6.17	6.44	6.80	7.25
30	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	6.14	6.41	6.77	7.21
35	3.85	4.40	4.74	4.98	5.17	5.32	5.45	5.57	5.67	6.04	6.29	6.64	7.07
40	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.96	6.21	6.55	6.96
45	3.80	4.34	4.66	4.89	5.07	5.22	5.34	5.45	5.55	5.90	6.14	6.47	6.88
60	3.79	4.32	4.63	4.86	5.04	5.19	5.31	5.41	5.51	5.85	6.09	6.42	6.81
120	3.71	4.22	4.52	4.73	4.90	5.03	5.14	5.24	5.33	5.65	5.86	6.16	6.51
∞	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.45	5.65	5.91	6.23

Table C.9: Critical values for one-sided Dunnett's t.

Entries are $d'_{.05}(K, \nu)$ where $P(\max_{j=1}^{K} t_{0j} > d'_{.05}(K, \nu)) = .05$.

							K						
ν	2	3	4	5	6	7	8	9	10	15	20	30	40
1	9.51	11.6	13.1	14.3	15.2	16.0	16.7	17.3	17.9	19.9	21.3	23.2	24.5
2	3.80	4.34	4.71	5.00	5.24	5.43	5.60	5.75	5.88	6.38	6.72	7.18	7.50
3	2.94	3.28	3.52	3.70	3.85	3.97	4.08	4.17	4.25	4.56	4.78	5.07	5.27
4	2.61	2.88	3.08	3.22	3.34	3.44	3.52	3.59	3.66	3.90	4.07	4.30	4.46
5	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30	3.36	3.57	3.71	3.92	4.05
6	2.34	2.56	2.71	2.83	2.92	3.00	3.06	3.12	3.17	3.37	3.50	3.68	3.81
7	2.27	2.48	2.62	2.73	2.81	2.89	2.95	3.00	3.05	3.23	3.36	3.53	3.64
8	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92	2.96	3.14	3.25	3.41	3.52
9	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86	2.90	3.06	3.18	3.33	3.44
10	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81	2.85	3.01	3.12	3.27	3.37
11	2.13	2.31	2.43	2.53	2.60	2.67	2.72	2.77	2.81	2.96	3.07	3.21	3.31
12	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74	2.78	2.93	3.03	3.17	3.27
13	2.09	2.27	2.39	2.48	2.55	2.61	2.66	2.71	2.75	2.90	3.00	3.14	3.23
14	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69	2.73	2.87	2.97	3.11	3.20
15	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67	2.71	2.85	2.95	3.08	3.17
16	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65	2.69	2.83	2.93	3.06	3.15
17	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64	2.67	2.81	2.91	3.04	3.13
18	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62	2.66	2.80	2.89	3.02	3.11
19	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61	2.65	2.79	2.88	3.01	3.10
20	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60	2.64	2.77	2.87	2.99	3.08
21	2.02	2.19	2.30	2.38	2.45	2.50	2.55	2.59	2.63	2.76	2.86	2.98	3.07
22	2.02	2.18	2.29	2.37	2.44	2.50	2.54	2.58	2.62	2.75	2.85	2.97	3.06
23	2.01	2.17	2.28	2.37	2.43	2.49	2.54	2.58	2.61	2.75	2.84	2.96	3.05
24	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57	2.60	2.74	2.83	2.95	3.04
25	2.00	2.17	2.27	2.36	2.42	2.48	2.52	2.56	2.60	2.73	2.82	2.94	3.03
26	2.00	2.16	2.27	2.35	2.42	2.47	2.52	2.56	2.59	2.72	2.81	2.94	3.02
27	2.00	2.16	2.27	2.35	2.41	2.47	2.51	2.55	2.59	2.72	2.81	2.93	3.01
28	1.99	2.15	2.26	2.34	2.41	2.46	2.51	2.55	2.58	2.71	2.80	2.92	3.01
29	1.99	2.15	2.26	2.34	2.40	2.46	2.50	2.54	2.58	2.71	2.80	2.92	3.00
30	1.99	2.15	2.25	2.34	2.40	2.45	2.50	2.54	2.57	2.70	2.79	2.91	2.99
35	1.98	2.13	2.24	2.32	2.38	2.44	2.48	2.52	2.55	2.68	2.77	2.89	2.97
40	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51	2.54	2.67	2.75	2.87	2.95
45	1.96	2.12	2.22	2.30	2.36	2.41	2.46	2.50	2.53	2.66	2.74	2.86	2.94
50	1.96	2.11	2.22	2.29	2.36	2.41	2.45	2.49	2.52	2.65	2.73	2.85	2.93
100	1.94	2.09	2.19	2.26	2.32	2.37	2.42	2.45	2.48	2.61	2.69	2.80	2.88
∞	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42	2.45	2.57	2.65	2.75	2.83

Table C.9: Critical values for one-sided Dunnett's t, continued.

Entries are $d'_{.01}(K, \nu)$ where $P(\max_{j=1}^{K} t_{0j} > d'_{.01}(K, \nu)) = .01$.

ν	2	3	4	5	6	7	8	9	10	15	20	30	40
1	47.7	58.1	65.6	71.5	76.3	80.3	83.8	86.8	89.5	99.6	107	116	122
2	8.88	10.0	10.9	11.5	12.0	12.5	12.8	13.2	13.5	14.6	15.3	16.4	17.1
3	5.48	6.04	6.44	6.74	6.99	7.20	7.38	7.54	7.67	8.20	8.56	9.06	9.41
4	4.41	4.80	5.07	5.28	5.45	5.59	5.72	5.82	5.92	6.28	6.53	6.87	7.11
5	3.90	4.21	4.43	4.60	4.73	4.85	4.94	5.03	5.11	5.39	5.59	5.87	6.06
6	3.61	3.88	4.06	4.21	4.32	4.42	4.51	4.58	4.64	4.89	5.06	5.30	5.46
7	3.42	3.66	3.83	3.96	4.06	4.15	4.22	4.29	4.35	4.57	4.72	4.93	5.08
8	3.29	3.51	3.66	3.78	3.88	3.96	4.03	4.09	4.14	4.35	4.49	4.68	4.81
9	3.19	3.40	3.54	3.66	3.75	3.82	3.89	3.94	3.99	4.18	4.31	4.49	4.62
10	3.11	3.31	3.45	3.56	3.64	3.72	3.78	3.83	3.88	4.06	4.18	4.35	4.47
11	3.06	3.25	3.38	3.48	3.56	3.63	3.69	3.74	3.79	3.96	4.08	4.24	4.35
12	3.01	3.19	3.32	3.42	3.50	3.56	3.62	3.67	3.71	3.88	3.99	4.15	4.26
13	2.97	3.15	3.27	3.37	3.44	3.51	3.56	3.61	3.65	3.81	3.92	4.08	4.18
14	2.94	3.11	3.23	3.33	3.40	3.46	3.52	3.56	3.60	3.76	3.87	4.01	4.12
15	2.91	3.08	3.20	3.29	3.36	3.42	3.47	3.52	3.56	3.71	3.82	3.96	4.06
16	2.88	3.05	3.17	3.26	3.33	3.39	3.44	3.48	3.52	3.67	3.78	3.92	4.01
17	2.86	3.03	3.14	3.23	3.30	3.36	3.41	3.45	3.49	3.64	3.74	3.88	3.97
18	2.84	3.01	3.12	3.21	3.28	3.33	3.38	3.43	3.46	3.61	3.71	3.84	3.94
19	2.83	2.99	3.10	3.18	3.25	3.31	3.36	3.40	3.44	3.58	3.68	3.81	3.90
20	2.81	2.97	3.08	3.17	3.23	3.29	3.34	3.38	3.42	3.56	3.65	3.78	3.88
21	2.80	2.96	3.07	3.15	3.22	3.27	3.32	3.36	3.40	3.53	3.63	3.76	3.85
22	2.79	2.94	3.05	3.13	3.20	3.25	3.30	3.34	3.38	3.51	3.61	3.74	3.83
23	2.78	2.93	3.04	3.12	3.18	3.24	3.28	3.33	3.36	3.50	3.59	3.72	3.81
24	2.77	2.92	3.03	3.11	3.17	3.22	3.27	3.31	3.35	3.48	3.57	3.70	3.79
25	2.76	2.91	3.02	3.10	3.16	3.21	3.26	3.30	3.33	3.47	3.56	3.68	3.77
26	2.75	2.90	3.01	3.08	3.15	3.20	3.25	3.29	3.32	3.45	3.54	3.67	3.75
27	2.74	2.89	3.00	3.07	3.14	3.19	3.24	3.27	3.31	3.44	3.53	3.65	3.74
28	2.74	2.88	2.99	3.07	3.13	3.18	3.22	3.26	3.30	3.43	3.52	3.64	3.72
29	2.73	2.88	2.98	3.06	3.12	3.17	3.22	3.25	3.29	3.42	3.51	3.63	3.71
30	2.72	2.87	2.97	3.05	3.11	3.16	3.21	3.25	3.28	3.41	3.50	3.62	3.70
35	2.70	2.84	2.94	3.02	3.08	3.13	3.17	3.21	3.24	3.37	3.45	3.57	3.65
40	2.68	2.82	2.92	2.99	3.05	3.10	3.14	3.18	3.21	3.34	3.42	3.54	3.62
45	2.67	2.81	2.90	2.98	3.03	3.08	3.12	3.16	3.19	3.31	3.40	3.51	3.59
50	2.65	2.79	2.89	2.96	3.02	3.07	3.11	3.14	3.18	3.30	3.38	3.49	3.57
60	2.61	2.74	2.83	2.90	2.95	3.00	3.04	3.07	3.10	3.22	3.29	3.40	3.47
∞	2.56	2.69	2.77	2.84	2.89	2.93	2.97	3.00	3.03	3.14	3.21	3.31	3.38

Table C.9: Critical values for two-sided Dunnett's t, continued.

Entries are $d_{.05}(K, \nu)$ where $P(\max_{j=1}^{K} t_{0j} > d_{.05}(K, \nu)) = .05$.

ν	2	3	4	5	6	7	8	9	10	15	20	30	40
							K						
1	17.4	20.0	21.9	23.2	24.3	25.2	25.9	26.6	27.1	29.3	30.7	32.6	33.9
2	5.42	6.06	6.51	6.85	7.12	7.35	7.54	7.71	7.85	8.40	8.77	9.28	9.62
3	3.87	4.26	4.54	4.75	4.92	5.06	5.18	5.28	5.37	5.72	5.95	6.27	6.49
4	3.31	3.62	3.83	3.99	4.13	4.23	4.33	4.41	4.48	4.75	4.94	5.19	5.36
5	3.03	3.29	3.48	3.62	3.73	3.82	3.90	3.97	4.03	4.26	4.42	4.64	4.79
6	2.86	3.10	3.26	3.39	3.49	3.57	3.64	3.71	3.76	3.97	4.11	4.31	4.45
7	2.75	2.97	3.12	3.24	3.33	3.41	3.47	3.53	3.58	3.78	3.91	4.09	4.22
8	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41	3.46	3.64	3.76	3.93	4.05
9	2.61	2.81	2.95	3.05	3.14	3.20	3.26	3.32	3.36	3.53	3.65	3.82	3.93
10	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24	3.29	3.45	3.57	3.72	3.83
11	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19	3.23	3.39	3.50	3.65	3.76
12	2.50	2.68	2.81	2.90	2.98	3.04	3.09	3.14	3.18	3.34	3.45	3.59	3.69
13	2.48	2.65	2.78	2.87	2.94	3.00	3.06	3.10	3.14	3.29	3.40	3.54	3.64
14	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07	3.11	3.26	3.36	3.50	3.60
15	2.44	2.61	2.73	2.82	2.89	2.95	3.00	3.04	3.08	3.23	3.33	3.47	3.56
16	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02	3.06	3.20	3.30	3.43	3.53
17	2.41	2.58	2.69	2.78	2.85	2.90	2.95	3.00	3.03	3.18	3.27	3.41	3.50
18	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98	3.01	3.16	3.25	3.38	3.48
19	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96	3.00	3.14	3.23	3.36	3.45
20	2.38	2.54	2.65	2.73	2.80	2.86	2.90	2.95	2.98	3.12	3.22	3.34	3.43
21	2.37	2.53	2.64	2.72	2.79	2.84	2.89	2.93	2.97	3.11	3.20	3.33	3.42
22	2.36	2.52	2.63	2.71	2.78	2.83	2.88	2.92	2.96	3.09	3.19	3.31	3.40
23	2.36	2.51	2.62	2.70	2.77	2.82	2.87	2.91	2.95	3.08	3.17	3.30	3.38
24	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90	2.94	3.07	3.16	3.29	3.37
25	2.34	2.50	2.61	2.69	2.75	2.81	2.85	2.89	2.93	3.06	3.15	3.27	3.36
26	2.34	2.49	2.60	2.68	2.74	2.80	2.84	2.88	2.92	3.05	3.14	3.26	3.35
27	2.33	2.49	2.59	2.67	2.74	2.79	2.84	2.88	2.91	3.04	3.13	3.25	3.34
28	2.33	2.48	2.59	2.67	2.73	2.78	2.83	2.87	2.90	3.03	3.12	3.24	3.33
29	2.32	2.48	2.58	2.66	2.73	2.78	2.82	2.86	2.90	3.03	3.11	3.24	3.32
30	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86	2.89	3.02	3.11	3.23	3.31
35	2.30	2.46	2.56	2.64	2.70	2.75	2.79	2.83	2.86	2.99	3.08	3.20	3.28
40	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81	2.84	2.97	3.05	3.17	3.25
45	2.28	2.43	2.53	2.61	2.67	2.72	2.76	2.80	2.83	2.95	3.04	3.15	3.23
50	2.28	2.42	2.52	2.60	2.66	2.71	2.75	2.79	2.82	2.94	3.02	3.14	3.22
100	2.24	2.39	2.48	2.55	2.61	2.66	2.70	2.74	2.77	2.88	2.96	3.07	3.15
∞	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69	2.72	2.83	2.91	3.01	3.08

Table C.9: Critical values for two-sided Dunnett's t, continued.

Entries are $d_{.01}(K, \nu)$ where $P(\max_{j=1}^{K} t_{0j} > d_{.01}(K, \nu)) = .01$.

ν	2	3	4	5	6	7	8	9	10	15	20	30	40
1	87.0	100	109	116	122	126	130	133	136	146	154	163	169
2	12.4	13.8	14.8	15.6	16.2	16.7	17.1	17.5	17.8	19.1	19.9	21.0	21.8
3	6.97	7.64	8.10	8.46	8.75	8.99	9.19	9.37	9.53	10.1	10.5	11.1	11.5
4	5.36	5.81	6.12	6.36	6.55	6.72	6.85	6.98	7.08	7.49	7.77	8.15	8.41
5	4.63	4.97	5.22	5.41	5.56	5.68	5.79	5.89	5.97	6.29	6.51	6.81	7.02
6	4.21	4.51	4.71	4.87	5.00	5.10	5.20	5.28	5.35	5.62	5.80	6.06	6.24
7	3.95	4.21	4.39	4.53	4.64	4.74	4.82	4.89	4.95	5.19	5.35	5.58	5.74
8	3.77	4.00	4.17	4.29	4.40	4.48	4.56	4.62	4.68	4.90	5.05	5.25	5.40
9	3.63	3.85	4.01	4.12	4.22	4.30	4.37	4.43	4.48	4.68	4.82	5.01	5.15
10	3.53	3.74	3.88	3.99	4.08	4.16	4.22	4.28	4.33	4.52	4.65	4.83	4.96
11	3.45	3.65	3.79	3.89	3.98	4.05	4.11	4.16	4.21	4.39	4.52	4.69	4.81
12	3.39	3.58	3.71	3.81	3.89	3.96	4.02	4.07	4.12	4.29	4.41	4.57	4.69
13	3.33	3.52	3.65	3.74	3.82	3.89	3.94	3.99	4.04	4.20	4.32	4.48	4.59
14	3.29	3.47	3.59	3.69	3.76	3.83	3.88	3.93	3.97	4.13	4.24	4.40	4.50
15	3.25	3.43	3.55	3.64	3.71	3.78	3.83	3.88	3.92	4.07	4.18	4.33	4.43
16	3.22	3.39	3.51	3.60	3.67	3.73	3.78	3.83	3.87	4.02	4.13	4.27	4.37
17	3.19	3.36	3.47	3.56	3.63	3.69	3.74	3.79	3.83	3.98	4.08	4.22	4.32
18	3.17	3.33	3.45	3.53	3.60	3.66	3.71	3.75	3.79	3.94	4.04	4.18	4.28
19	3.15	3.31	3.42	3.50	3.57	3.63	3.68	3.72	3.76	3.90	4.00	4.14	4.24
20	3.13	3.29	3.40	3.48	3.55	3.60	3.65	3.69	3.73	3.87	3.97	4.11	4.20
21	3.11	3.27	3.37	3.46	3.52	3.58	3.63	3.67	3.71	3.85	3.94	4.08	4.17
22	3.09	3.25	3.36	3.44	3.50	3.56	3.61	3.65	3.68	3.82	3.92	4.05	4.14
23	3.08	3.23	3.34	3.42	3.48	3.54	3.59	3.63	3.66	3.80	3.89	4.02	4.11
24	3.07	3.22	3.32	3.40	3.47	3.52	3.57	3.61	3.64	3.78	3.87	4.00	4.09
25	3.05	3.21	3.31	3.39	3.45	3.51	3.55	3.59	3.63	3.76	3.85	3.98	4.07
26	3.04	3.19	3.30	3.37	3.44	3.49	3.54	3.58	3.61	3.74	3.83	3.96	4.05
27	3.03	3.18	3.28	3.36	3.42	3.48	3.52	3.56	3.60	3.73	3.82	3.94	4.03
28	3.03	3.17	3.27	3.35	3.41	3.46	3.51	3.55	3.58	3.71	3.80	3.93	4.01
29	3.02	3.16	3.26	3.34	3.40	3.45	3.50	3.54	3.57	3.70	3.79	3.91	3.99
30	3.01	3.15	3.25	3.33	3.39	3.44	3.49	3.52	3.56	3.69	3.77	3.90	3.98
35	2.98	3.12	3.22	3.29	3.35	3.40	3.44	3.48	3.51	3.64	3.72	3.84	3.92
40	2.95	3.09	3.19	3.26	3.32	3.37	3.41	3.44	3.48	3.60	3.68	3.80	3.88
45	2.93	3.07	3.16	3.24	3.29	3.34	3.38	3.42	3.45	3.57	3.65	3.76	3.84
50	2.92	3.05	3.15	3.22	3.27	3.32	3.36	3.40	3.43	3.55	3.63	3.74	3.82
60	2.86	2.98	3.07	3.14	3.19	3.24	3.27	3.31	3.34	3.45	3.52	3.63	3.70
∞	2.79	2.92	3.00	3.06	3.11	3.15	3.19	3.22	3.25	3.35	3.42	3.52	3.59

foo