

Notes for Statistics 3011
University of Minnesota
Spring 2020

Notes accompanying the 4th Edition of
Statistics: The Art and Science of Learning From Data
by Alan Agresti and Christine Franklin

Contents

CHAPTER 1: INTRODUCTION	1
1.1 The Basics	1
1.2 The Role of Computers in Statistics	4
1.2.1 Introduction to R	4
1.2.2 Getting Started with R	4
1.2.3 Entering Data in R	5
1.2.4 Getting Help	7
CHAPTER 2: EXPLORING DATA	8
2.1 Types of Data	8
2.2 Graphical Summaries of Data	9
2.2.1 Graphical Summaries for Categorical Variables	9
2.2.2 Graphical Summaries for Quantitative Variables	12
2.3 Numerical Summaries of Quantitative Data	16
2.3.1 Measures of Center	16
2.3.2 Measures of Spread	19
CHAPTER 4: GATHERING DATA	32
4.1 Types of Studies	32
4.1.1 Experiment	32
4.1.2 Observational Study	32
4.1.3 Experiments vs. Observational Studies	32
4.2 Good and Poor Ways to Sample	33
CHAPTER 5: PROBABILITY	35
5.1 Randomness	35
5.2 Probability Models	36
5.3 Conditional Probability	44
CHAPTER 6: PROBABILITY DISTRIBUTIONS	48
6.1 Discrete Random Variables	48
6.1.1 Probability Distribution of A Discrete Random Variable	49
6.1.2 Center and Spread of A Probability Distribution	49
6.2 Continuous Random Variables	51
6.2.1 Density Curves	51

6.2.2 The Normal Distribution	55
6.2.3 The Normal Distribution in R	69
CHAPTER 7: SAMPLING DISTRIBUTIONS	70
7.1 The Sampling Distribution of A Sample Mean	82
7.2 The Sampling Distribution of A Sample Proportion	88
CHAPTER 8: CONFIDENCE INTERVALS	91
8.1 Point Estimation	91
8.2 Interval Estimation	93
8.2.1 Confidence Intervals for A Population Proportion, p	94
8.2.2 Confidence Intervals for A Population Mean, μ	103
CHAPTER 9: HYPOTHESIS TESTS	111
9.1 Elements of A Hypothesis Test	111
9.2 Normal Hypothesis Test for Population Proportion p	115
9.3 The t -Test: Hypothesis Testing for Population Mean μ	118
9.4 Possible Errors in Hypothesis Testing	123
9.5 Limitations and Common Misinterpretations of Hypothesis Testing	124
CHAPTER 10: COMPARING TWO GROUPS	126
10.1 Comparing Two Proportions	126
10.1.1 Point Estimation for $p_1 - p_2$	127
10.1.2 Confidence Intervals for $p_1 - p_2$	127
10.1.3 Hypothesis Tests for Comparing p_1 and p_2	129
10.2 Comparing Two Means - Matched Pairs	133
10.3 Comparing Two Means - Independent Samples	136
10.3.1 Point Estimation for $\mu_1 - \mu_2$	136
10.3.2 Confidence Intervals for $\mu_1 - \mu_2$	137
10.3.3 The Two-Sample t -Test for Comparing μ_1 and μ_2	138
CHAPTER 14: ANALYSIS OF VARIANCE	142
14.1 One-Way ANOVA	143
14.2 Follow-Up to the ANOVA F -test	149
CHAPTER 3: TWO-VARIABLE ASSOCIATIONS	151
CHAPTER 11: ASSOCIATION BETWEEN TWO CATEGORICAL VARIABLES	153

11.1 Chi-Squared Test for Independence	153
11.2 Measures of Association	159
CHAPTER 12: REGRESSION ANALYSIS	160
12.0 Exploring the Data (A Return to Chapter 3)	161
3.2.1 Graphical Summaries - The Scatterplot	161
3.2.2 Numerical Summaries - Correlation	164
3.2.3 Numerical Summaries - Least Squares Regression	166
12.1 Regression Analysis	174
12.2 Inference About the Population Regression Model	175
12.2.1 Estimating α and β	175
12.2.2 Hypothesis Tests and Confidence Intervals for β	177
12.2.3 Measuring the Strength of the Linear Relationship	179
12.3 Correlation and Regression: A Cautionary Tale	181
CHAPTER 13: MULTIPLE REGRESSION	184
13.1 The Multiple Regression Model	185
13.2 Estimation of the Multiple Regression Model	186
13.3 Inference for the Multiple Regression Model	188
NOTATION	201
TABLE A	202
ADDITIONAL NOTES	204

CHAPTER 1: INTRODUCTION

1.1 The Basics

Definition: statistics

Statistics is the science of collecting, organizing, interpreting, and learning from data.

COURSE GOAL:

Learn how to use statistical methods to translate data into knowledge so that we can investigate questions in an objective manner. Here are some examples of questions we'll have the tools to answer before the end of the semester:

1. How can we estimate the percentage of American citizens who would vote for Hillary Clinton if the presidential election were held today? How certain are we about our estimate?
2. What is the relationship between the amount of time spent studying and the score received on an exam?
3. Is there an association between smoking and divorce? If so, does that mean smoking causes divorce?

THREE ASPECTS OF STATISTICS

1. **Design:** Planning how to obtain data to answer the question of interest.
2. **Description:** Summarizing the data that are obtained.
3. **Inference:** Using sample data to learn about the population: **make decisions and predictions based on the data for answering the statistical question.**

Definition: population

The *population* is a collection of units of interest.

Examples:

- adults in Minneapolis
- polar bears in the Arctic
- shoes from a factory

Definition: subject

Subjects are the individual units of a population (e.g. an adult, a polar bear, a shoe).

NOTE: Very rarely can we observe the *entire* population of interest. The basic goal of statistics is to:

instead, observe a sample and use it to learn about the population.

Definition: sample

A *sample* is a subset of the units of a population.

EXAMPLE 1.1

Suppose we want to know what percentage of Minnesota adults own a firearm. Since it's impossible to ask all adult Minnesotans, we instead take a poll of 1000 Minnesotans by selecting a sample from the phone book.

- What is the population?

All adult Minnesotans.

- What is the sample?

The 1000 Minnesotans selected from the phone book.

- Is this a good sample?

No. Some people don't have phones, have unlisted phone numbers, or have cell phones.

What makes a “good” sample?

It should be representative of the population. This can be obtained by selecting sample subjects randomly (more in Ch 4).

Where do statistical methods come in?

1. Use **design** to obtain an appropriate sample from the population.
2. **Describe** the sample data with graphical and numerical summaries.
3. Perform **statistical inference**.

Definition: statistical inference

The procedure of using a sample to learn about a population is called *statistical inference*.

Definition: parameter

A *parameter* is a number that describes a *population*. It is usually UNKNOWN.

Definition: statistic

A *statistic* is a number that describes a *sample*. It can be computed from data; therefore, it is KNOWN once a sample is obtained.

NOTE: We use a sample statistic to estimate a population parameter!

EXAMPLE 1.2:

We want to know the average height of all students at the U. It is logistically impossible to measure everybody. Instead, we take this class as a sample, measure our heights, and average them.

- population:
all U students
- sample:
this class
- parameter:
average height of all U students
- statistic:
average height of this class

1.2 The Role of Computers in Statistics

Unlike previous generations, it's no longer necessary to perform complex statistical procedures by hand. With this in mind, we will use computers for some of our statistical analysis. This will allow us to focus on the *interpretation* of a statistical analysis instead of getting bogged down by tedious calculations.

In this course we will learn how to use a programming language called "R".

1.2.1 Introduction to R

What is R?

R is free software, which means it is free as in "free beer" (you can download it with no charge) and free as in "free speech" (you can do whatever you want with it except to make it non-free). It is available for download from the Comprehensive R Archive Network (CRAN) at

<http://cran.r-project.org>.

R is the language of choice for research statistics. If it's statistics, you can do it in R.

Why R?

R is the most powerful statistical computing environment in existence, what many applied and research statisticians use. It is best at statistical computing and graphics. So why use R for an introductory statistics course? We could use a graphing calculator, a spreadsheet, or some less sophisticated program. However, we don't want to be limited by simpler tools which are rarely used outside of a classroom.

1.2.2 Getting Started with R

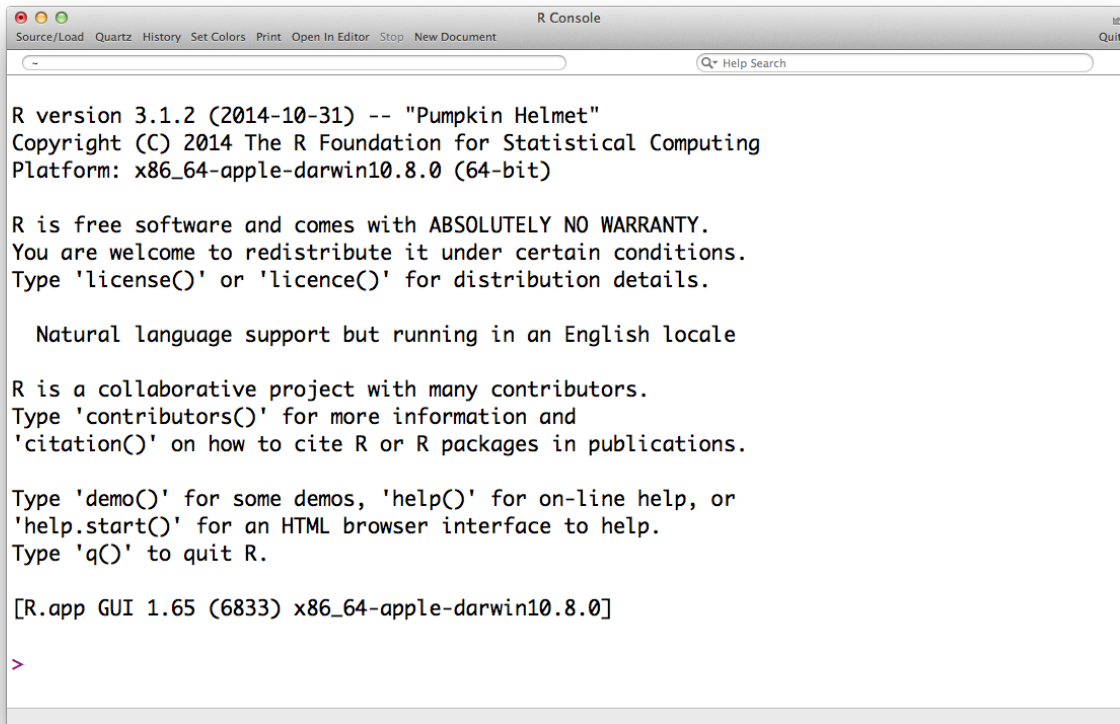
1. The interactive interface

The default graphical user interface (GUI) that comes with R is called the R console, which is a command line interface (like the command prompt on Windows or the Terminal on Mac; see the following figure). In general, command line tools are hard to use for beginners. Better R GUIs, such as the R Commander, RStudio, and Tinn-R, are available online. My favorite R GUI is RStudio, which can be obtained freely from **<http://www.rstudio.com>**. However, we still need to install R before we can use RStudio. To make our experience with R less frustrating, most GUIs provide the following two features.

- **Code auto-completion:** automatic completion of code can be done by pressing the **Tab** key.
- **Retrieving Previous Commands:** recall previous commands using the up and/or down arrow keys.

A detailed description of these two features for RStudio can be found at

<https://support.rstudio.com/hc/en-us/articles/200404846-Working-in-the-Console>.



```

R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.65 (6833) x86_64-apple-darwin10.8.0]
>

```

2. Using R as a calculator

Type the following in the R console. Anything following a `#` is just a note that you do not need to type in. You can submit them one at a time or all at once.

```

3*9 + 10
20/2
2^3          # ^ denotes power. This returns 2 cubed.
sqrt(16)     # square root of 16
pi*4^2       # area of a circle with radius 4
log(15)      # natural logarithm (base e) of 15
log10(100)   # logarithm of 100 using base 10

```

The results screen reprints the command along with your result:

```

> 3*9 + 10
[1] 37

```

1.2.3 Entering Data in R

- **By hand**

First, come up with a name for your variable (`y`, `Y`, `weight`, `Age`).

R is case sensitive, thus “weight” and “Weight” refer to different variables.

Also, no spaces can be used in the variable names. If necessary, separate words using periods or capitalization: “RainAmount” or “rain.amount”.

Example:

We record the weights of 5 people: 120, 160, 135, 190, and 210

```
weight <- c(120, 160, 135, 190, 210)
```

where “<-” is the assignment operator in R and “c” denotes a column of data.

- **From a URL**

Entering data from a URL into R

Data can be imported into R using the “`read.table`” or “`read.csv`” commands. For example, if we wanted to import the Australian crime data set into R and work with the variables inside this data set, we would choose a name for it (say “`au_crime`”) and type the following:

```
> au_crime <- read.table(
>     "http://www.stat.umn.edu/~wuxxx725/data/crime.txt",
>     header = TRUE
> )
> names(au_crime)
[1] "Year"           "firearm.suicide"   "firearm.homicide"
[4] "non.firearm.suicide" "non.firearm.homicide"
```

where “`header = TRUE`” tells R that the file contains column labels (or headers).

For csv (comma-separated values) file, you can use `read.csv` command. Type `?read.csv` in console to learn more.

```
> flies<-read.csv("http://stat2.org/datasets/FruitFlies.csv")
> View(flies)
```

Once the dataset is imported properly, you should see the name of the data set in ‘Environment’ window in RStudio (upper right window). To view the dataset, click on the name of the dataset or type `View(NameOfYourDataset)` command in console.

- **From a file on your computer**

Entering data from a file into R

Importing data from a file in your computer into R is similar to importing data from a URL. However, instead of typing the *web location* into the “`read.table`” command, you would type

the *location of the file in your computer*.

You can also use `file.choose()` command to locate the data file in your computer.

```
> dat<-read.csv(file.choose())
```

After running the command above, there will be a pop-up window asking you to locate the csv file. You may use `read.table` or `read.delim` instead of `read.csv` for different types of dataset.

1.2.4 Getting Help

You can get help on any function in R by typing “`help(function name)`”. For instance, if we want to learn more about the “`hist`” function, type “`help(hist)`”.

It is also easy to find help online and a more complete guide to programming in R can be accessed from <http://cran.r-project.org/doc/manuals/R-intro.pdf> (this may or may not be useful for this class).

CHAPTER 2: EXPLORING DATA

Motivating Example:

Suppose we have IQ test scores for 60 randomly chosen fifth-grade students:

145, 101, 123, 106, 117, 102, 139, 142, 94, 124, 90, 108, 126, 134, 100,
115, 103, 110, 122, 124, 136, 133, 114, 128, 125, 112, 109, 116, 139, 114,
130, 109, 131, 102, 101, 112, 96, 134, 117, 127, 122, 114, 110, 113, 110,
117, 105, 102, 118, 81, 127, 109, 97, 82, 118, 113, 124, 137, 89, 101

What does this data tell you about the IQ of a fifth grader? It's hard to understand much about the data by just looking at a pile of numbers!

The first step of any statistical analysis is getting to know the data you're working with. In Chapter 2 we discuss exploring **sample** data using both graphical and numerical summaries.

2.1 Types of Data

Definition: variable

A *variable* is any characteristic of a subject in a population.

Examples: height, IQ, income, # of hot dogs eaten last year, gender, eye color

TWO TYPES OF VARIABLES:

1. **Categorical (Qualitative) Variable:**

Classifies subjects as belonging to a certain group/category.

ex: gender, eye color, car make, race, major, area code

2. **Quantitative Variable:** Takes on numerical values that represent different magnitudes.

ex: height, income, weight

A quantitative variable can either be

- (a) Discrete: The possible values of a discrete quantitative variable form a set of separate numbers (i.e. can be listed).

ex: # of hot dogs eaten, # of t.v.'s, # of accidents/day

- (b) Continuous: The possible values of a continuous quantitative variable form an interval. That is, there is an infinite continuum of possible values.

ex: height, blood pressure, amount of rainfall

A Quick Summary:

2.2 Graphical Summaries of Data

2.2.1 Graphical Summaries for Categorical Variables

Graphical summaries of categorical variables help us visualize the distribution of the data among the separate categories. Before constructing the graphical summary, we first organize the categorical data into a *frequency table*.

Definition: frequency table

A *frequency table* is a listing of possible values for a variable, together with the number of observations for each value. (Note that we can also construct frequency tables for quantitative variables.)

Definition: proportion

A *proportion* of observations that fall in a certain category is the count of observations in that category divided by the total number of observations.

(NOTE: percent = $100 \times$ proportion)

EXAMPLE 2.1

In an online poll from a few years ago, 351 people weighed in on the following question:

On which issue are political candidates most helped/hurt by their stance?

Their answers are summarized in the following frequency table (source: BuzzDash.com):

Issue	Frequency	Proportion	Percent
Abortion	33	.094	9.4%
Gay Marriage	24	.068	6.8%
Religion	40	.114	11.4%
War	179	.510	51.0%
Economy	75	.214	21.4%
Total	351	1	100%

1. Categorical variable:
2. What percentage of those surveyed thought either the war or the economy was the most influential issue?
3. What proportion of respondents did *not* think religion was the most influential issue?

Two Graphical Summaries for Categorical Variables:

1. Pie Chart

A circle is drawn with a “slice of pie” representing each category’s % of observations.

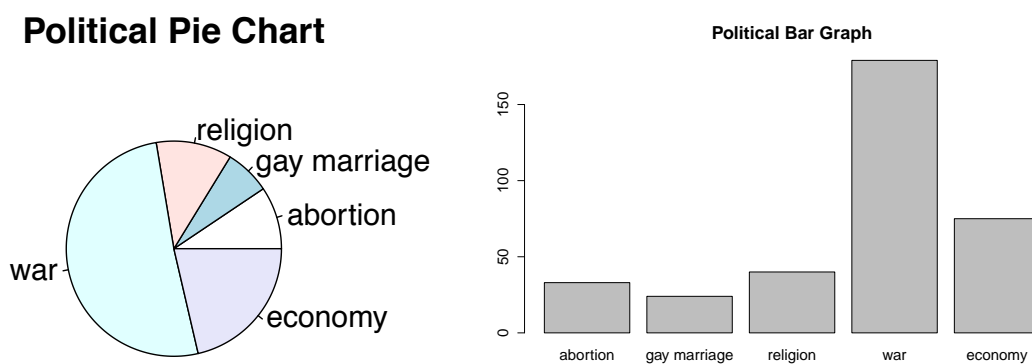
2. Bar Graph

A bar is drawn for each category with the bar’s height representing the % or count of observations.

EXAMPLE 2.1 CONTINUED

351 web users surveyed about what they believe to be the most influential issue in political elections.

```
# R code for drawing a pie chart and bar graph:
> issue <- c("abortion", "gay marriage", "religion", "war", "economy")
> count <- c(33,24,40,179,75)
> pie(count, issue, main="Political Pie Chart")           #creates a pie chart titled
                                                         #"Political Pie Chart"
> barplot(count, names=issue, main="Political Bar Graph") #creates a bar graph titled
                                                         #"Political Bar Graph"
```



Observations:

War seems to be the overriding issue that people base their vote on. Social issues aren't as influential as national issues.

Pie Charts vs. Bar Graphs

1. Pie charts emphasize a category's relation to the whole, but make it difficult to compare categories to each other!
2. Bar graphs compare the sizes of each group of a categorical variable (not in relation to the whole).
3. Bar graphs are easier to read and more flexible than pie charts.

2.2.2 Graphical Summaries for Quantitative Variables

Definition: distribution

A *distribution* of data shows the values a variable takes and how often they occur.

Graphical summaries help us visualize the following features of distributions for quantitative variables:

- 1.
- 2.
- 3.

Two Graphical Summaries for Quantitative Variables:

1. Stem-and-Leaf Plot
2. Histogram

Constructing a Stem-and-Leaf Plot

Each observation is represented by a “stem” and a “leaf”...

1. Order the data from smallest to largest.
2. Select one or more leading digits to be the *stem*. The final digit is the *leaf*.
3. Place stems in a column from smallest to largest.
Do not skip a stem even if it has no observations.
4. Draw a vertical line to the right of the stems.
5. Write down the value of each leaf in the row to the right of its stem (in increasing order).

Examples:

1. 10, 20, 23, 27, 27, 27, 40, 49, 55, 56

2. 40133, 40598, 41532, 41808, 41875, 42200

EXAMPLE 2.2

Basketball-reference.com reported on NBA team salaries for the 2016–17 season. This data can also be found at <http://www.stat.umn.edu/~wuxxx725/data/NBASalary.txt>, where each team is listed with their team salary and a label indicating their conference (“E” for Eastern and “W” for Western):

Team	Conference	Salary
Atlanta Hawks	E	99374470
Boston Celtics	E	94533435
.	.	.
.	.	.

We can visualize this *quantitative* data by constructing a stem-and-leaf plot in R:

```
> dat <- read.table("http://users.stat.umn.edu/~wuxxx725/data/NBASalary.txt", sep = '\t',
+                   header = TRUE)
> attach(dat)
> stem(Salary, scale = 0.5)
```

The decimal point is 7 digit(s) to the right of the |

```
7 | 267
8 | 01348
9 | 22566899
10 | 013555899
11 | 23568
```

NOTES:

1. “sep = ‘\t’” tells R that the field separator of the data file is the tab character.
2. “scale = 0.5” tells R to make the stem half as long as the default.

Observations:

Shape: The lower “tail” extends further than the upper “tail”.

Outlier: No team appears to have a salary much smaller or larger than the other teams.

Center: 100–110 million.

Spread: The salaries range from 72 to 118 million.

Constructing a Histogram

Histograms break up the range of values of a variable into classes and display the count (or percent) of the observations that fall into each class.

1. Divide the range of the data into intervals of equal width.

NOTE: We need to choose a width that gives us a good picture of the distribution of the data. The number of intervals should not be too many or too few (see the example below).

2. Count the number of observations that fall into each interval.

3. On the horizontal axis, mark the scale of the variable.

On the vertical axis, mark the scale for counts (or percents).

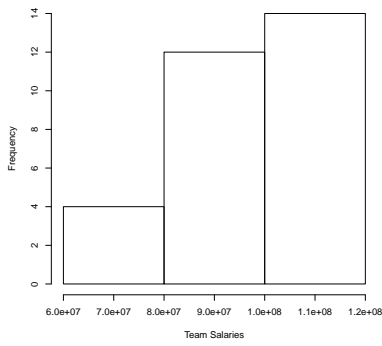
4. Above each interval, draw a bar whose height is either the corresponding count or percent for that interval.

EXAMPLE 2.2 CONTINUED

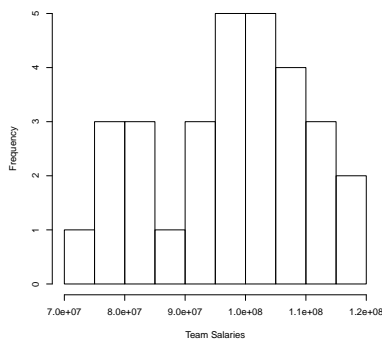
Use R to construct a histogram for the 2016–17 NBA salary data.

```
> hist(Salary, xlab="Team Salaries", main="", breaks=2)
> hist(Salary, xlab="Team Salaries", main="", breaks=10)
> hist(Salary, xlab="Team Salaries", main="", breaks=20)
```

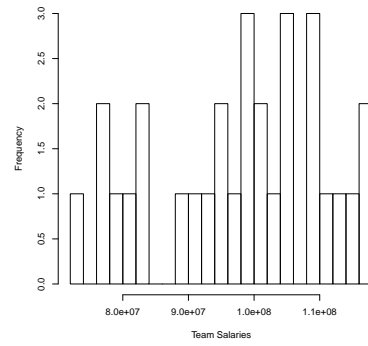
- “xlab” gives a label for the horizontal axis
- “main” supplies the title
- “break” specifies the desired number of breaks or bars you want (it doesn’t always give you exactly what you want).



(a)

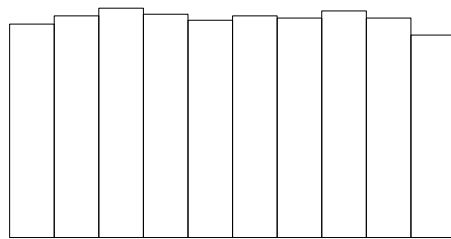
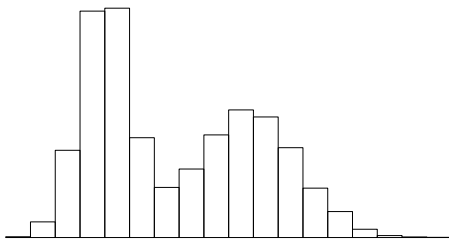
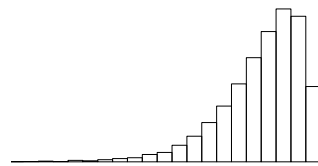
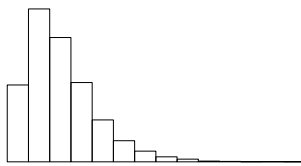
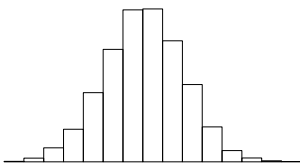


(b)



(c)

Common Distribution Shapes



Histograms vs Stem-and-Leaf Plots

1. A stem-and-leaf plot allows us to see the value of each individual observation. We lose this detail with a histogram.
2. Stem-and-leaf plots become unwieldy for large data sets.
3. Histograms are more versatile.

2.3 Numerical Summaries of Quantitative Data

Graphical summaries give us a good idea of the shape of a distribution as well as a rough idea of its center and spread. However, numerical summaries provide more precise descriptions of the characteristics of a distribution (specifically, its center and spread).

Notation:

1. n = the number of observations in a sample
2. x_i = the i th observation of a sample (so the list of observations is x_1, x_2, \dots, x_n)
3. \sum = summation

$$\sum x_i = x_1 + x_2 + \dots + x_n$$

2.3.1 Measures of Center

Two measures of center:

1. **mean** (\bar{x}) = the average of all observations

$$\bar{x} =$$

2. **median** (M) = the middle number when measurements are ordered from smallest to largest

When n is odd, $M =$

When n is even, $M =$

Examples: Calculate mean and median for the following samples.

1. 6, 5, 9, 5, 1001 (Sorted: 5, 5, 6, 9, 1001) 2. 6, 5, 9, 5 (Sorted: 5, 5, 6, 9)

1. $\bar{x} =$

sorted data: 5, 5, 6, 9, 1001

$M =$

2. $\bar{x} =$

sorted data: 5, 5, 6, 9

$M =$

Mean vs Median (Round 1)

1. the mean and median are usually not equal
2. \bar{x} is calculated using *all* the data whereas M ignores all but the middle values.
(SEE BOOK EXAMPLE pp. 52–53)
THINK: (100, 100, 100, 100, 100) vs. (0, 0, 100, 100, 100)

3. Definition: resistant

A numerical summary of the observations is *resistant* if extreme observations have little, if any, influence on its value.

the mean is affected by extreme values (outliers) but the median is not; that is, the median is *resistant* to outliers but the mean is not

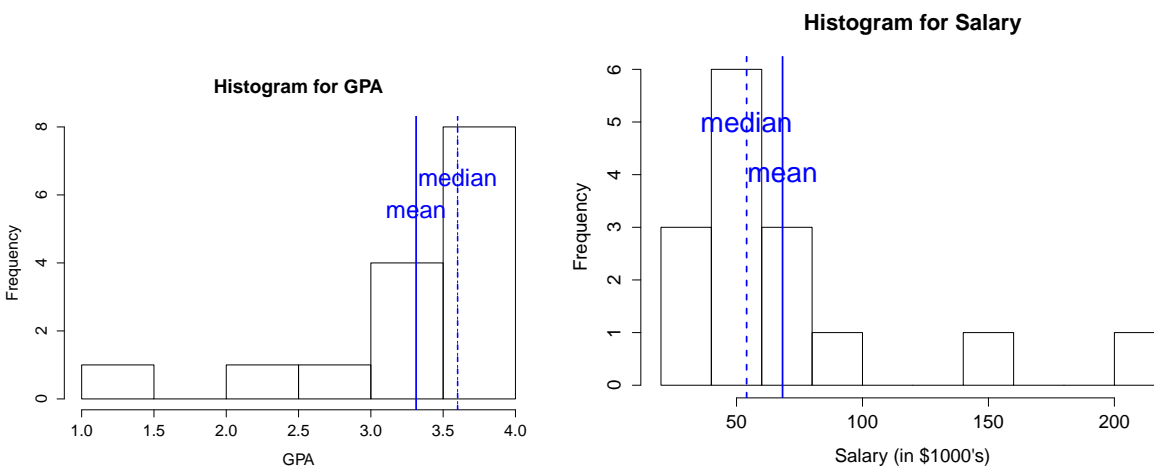
EXAMPLE 2.3

A survey of 15 recent graduates revealed the following information about their grade point averages (GPA) and their starting salaries upon graduation:

GPA:	1.1, 2.3, 2.9, 3.2, 3.3, 3.5, 3.5, 3.6, 3.6, 3.6, 3.6, 3.8, 3.8, 3.9, 4
Salary (in \$1000's):	31, 26, 44, 37, 55, 67, 52, 54, 143, 201, 90, 51, 43, 66, 64

Use R to look at the distributions of both variables (GPA and salary) and also calculate the mean and median for both.

```
> gpa <- c(1.1, 2.3, 2.9, 3.2, 3.3, 3.5, 3.5, 3.6, 3.6, 3.6, 3.6, 3.8, 3.8, 3.9, 4)
> salary <- c(31, 26, 44, 37, 55, 67, 52, 54, 143, 201, 90, 51, 43, 66, 64)
> hist(gpa, xlab="GPA", main="Histogram for GPA")
> hist(salary, xlab="Salary (in $1000's)", main="Histogram for Salary", breaks=7)
> mean(gpa)
[1] 3.313333
> median(gpa)
[1] 3.6
> mean(salary)
[1] 68.26667
> median(salary)
[1] 54
```



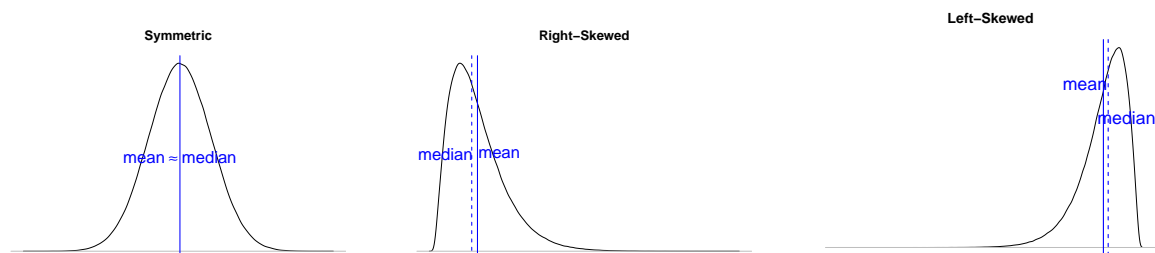
1. Describe the shape of the distribution of GPA's.

What is the relationship between the mean and median of the GPA's?

2. Describe the shape of the distribution of starting salaries.

What is the relationship between the mean and median of the salaries?

Mean vs Median (Round 2)



2.3.2 Measures of Spread

Motivating Example

Two separate statistics classes were given the same exam and received the following scores:

	Class 1	Class 2
Scores	40	60
	50	65
	70	70
	90	75
	100	80
\bar{x}	70	70
M	70	70

The mean and median exam scores of the two classes are the same. However, the exam scores for Class 1 are much more spread out than the scores for Class 2.

Looking at measures of center alone ignores other features of the distribution and can be misleading. We can get a better understanding of a distribution by looking at *both* measures of center *and* measures of spread!

Three Measures of Spread:

1. Range
2. Interquartile Range
3. Standard Deviation

I: RANGE

The *range* is the difference between the largest and smallest observations. That is,

$$\text{range} =$$

Example: Calculate the ranges of the exam scores for the two classes in the above example.

Class 1: range =

Class 2: range =

The Good:

Range is a simple measure of spread that is easy to calculate.

The Bad:

Range is only calculated using the most extreme values of a data set. Therefore, it can be misleading and is not resistant to outliers.

ex: 2, 100, 100, 100, 100, 100, 100 range = 98

II: INTERQUARTILE RANGE**Definition: percentile**

The p th *percentile* of a distribution is the value below which $p\%$ of the observations fall.

Example: Sam takes the GRE and scores in the 80th percentile. Therefore, Sam has performed better than 80% of the other students taking the test but worse than the top 20%.

NOTE: We can calculate any percentile (3rd, 56th, 81st, etc). However, it is most common to use the *quartiles* as a measure of the spread of a distribution.

1. **First Quartile (Q1)** = 25th percentile

The lowest 25% of the data lies below Q1.

Can also think of Q1 as the median of observations below M .

2. **Second Quartile (Q2)** = 50th percentile = median

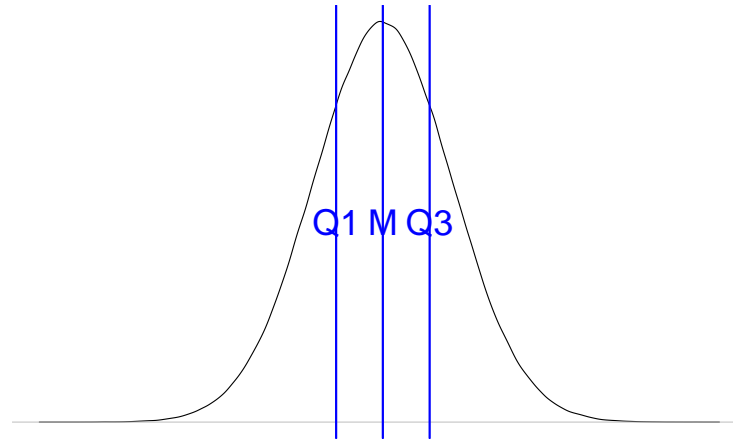
50% of the data are below and 50% are above M

3. **Third Quartile (Q3)** = 75th percentile

The highest 25% of the data lies above Q3.

Can also think of Q3 as the median of observations above M .

The quartiles split the distribution into 4 parts:



Definition: interquartile range (IQR)

The *interquartile range* is the difference between the first and third quartiles. That is,

$$\text{IQR} =$$

Notes about IQR:

1. The larger the IQR, the more spread out the data is.
2. IQR is resistant to outliers since it's calculated using only the middle 50% of the data set (outliers tend to be outside this range).

Definition: 5-number summary

The *5-number summary* is a brief numerical description of the center *and* spread of a distribution:

minimum Q1 M Q3 maximum

EXAMPLE 2.2 CONTINUED

Use R to find 5-number summaries of NBA team salaries for the Eastern and Western conferences.

```
> summary(Salary[Conference=="E"])
  Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
72290000 95450000 100200000 98580000 105100000 117700000
> summary(Salary[Conference=="W"])
  Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
77470000 83340000 95900000 96640000 110300000 116200000
```

1. IQR for Eastern Conference:

IQR =

2. IQR for Western Conference:

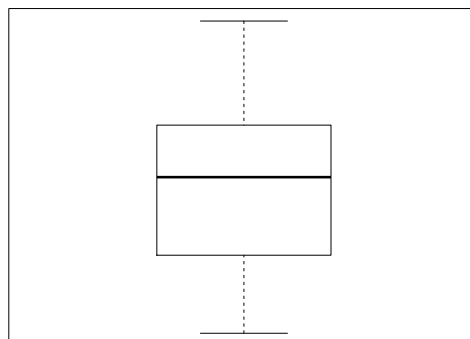
IQR =

3. Conclusions:

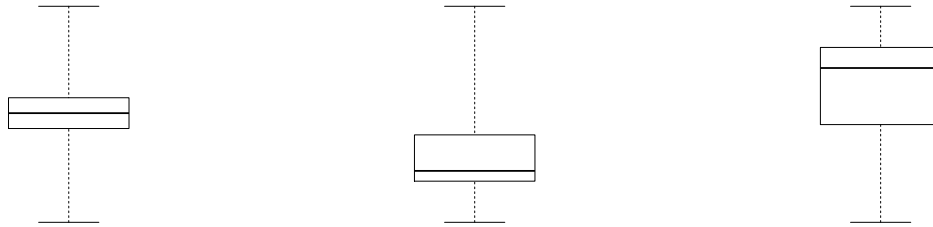
Salaries are more spread out in the _____, but the mean salary is higher in the _____.

Definition: box plot

The *box plot* is a plot of the five number summary.



NOTE: Not only do boxplots provide a picture of the center and spread of a distribution, they also give us an idea as to the shape or skewness of the distribution:



Definition: side-by-side box plot

A *side-by-side boxplot* graphs boxplots for more than one distribution (side by side). They allow us to compare the centers and spreads of different distributions.

EXAMPLE 2.4

To better understand his 98 students, a professor handed out a survey. Among other things, he asked about the amount of money they typically spend on a haircut:

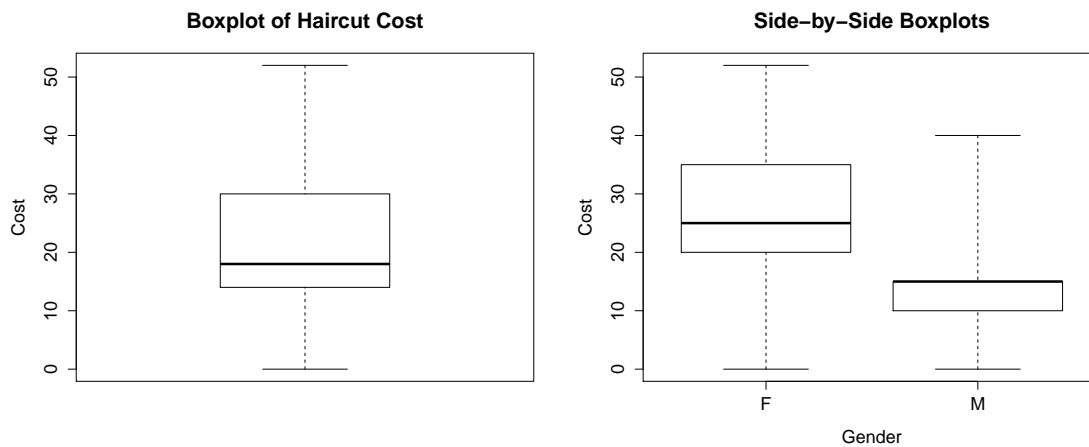
Year	Gender	Height	TV	Siblings	DistHome	Haircut
5	M	72	0	0	4	0
5	F	65	30	1	1	45
.
.

A full data set can be found at <http://www.stat.umn.edu/~wuxxx725/data/class.txt>. Use R to draw a boxplot and side-by-side boxplots of the costs of haircuts for both male (M) and female (F) students.

```
> dat <- read.table("http://www.stat.umn.edu/~wuxxx725/data/class.txt",
+                   header = TRUE)
> attach(dat)
# Regular boxplot:
> boxplot(Haircut, range=10, ylab="Cost", main="Boxplot of Haircut Cost")
# Side-by-side boxplots for males and females:
> boxplot(Haircut ~ Gender, range=10, xlab="Gender", ylab="Cost",
+         main="Side-by-Side Boxplots")
# For side-by-side boxplots always use this order: numerical variable ~ group variable
```

NOTES:

- By default, R marks any point that falls more than $1.5 \times \text{IQR}$ above $Q3$ or more than $1.5 \times \text{IQR}$ below $Q1$ as an outlier. These points are represented by open circles and the whiskers are only extended to values that aren't considered to be outliers.
- To obtain a boxplot whose 'whiskers' extend to the maximum and minimum values of a data set even if there are outliers, set 'range' to some large value such as 'range=10'. (By default, R sets 'range=1.5'.)



Observations:

- Overall cost is right skewed
- Women spend more than men
- Women are more spread out
- At least one outlier among men
- Women look fairly symmetric

III: STANDARD DEVIATION

Definition: sample variance (s^2)

The *sample variance* of a set of observations is the “average” of the squared deviations from the mean... WHAT?

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

NOTES:

1. $x_i - \bar{x}$ = the *deviation* of x_i from the mean.
2. s^2 describes how far a typical observation *deviates* from the mean.
3. s^2 is measured in units².
4. Why divide by $n - 1$ instead of n ?

It results in nicer mathematical properties.

Definition: sample standard deviation (s)

The *sample standard deviation* is the square root of the sample variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Example: Calculate the sample standard deviation for the following data: 4, 4, 6, 7, 9

Properties of s :

1. Interpretation: s = distance that a “typical” observation falls from the mean

2. s is measured in the same units as the original observations.

This makes it easier to interpret than s^2 .

3. Use s in conjunction with mean, \bar{x} .

Since s measures spread about the mean, only use s to describe the spread of a distribution when \bar{x} is used as the measure of center.

- 4.

5. The larger s is, the greater the spread of the data.

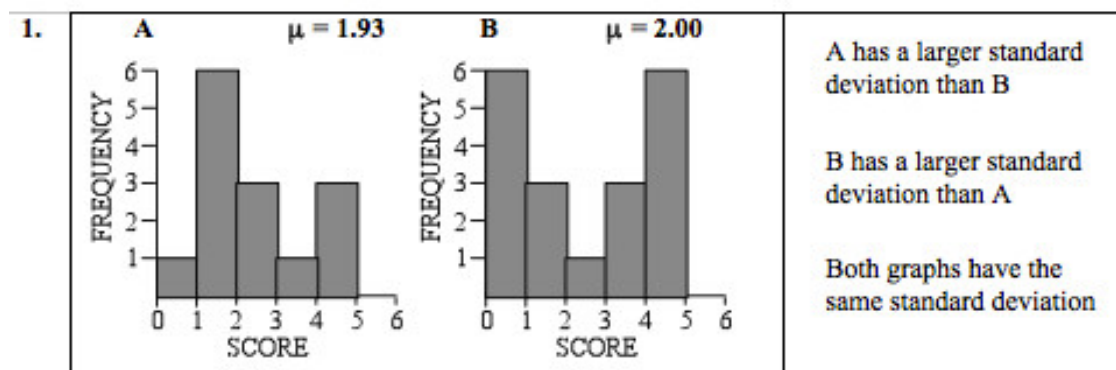
6. $s = 0 \Rightarrow$

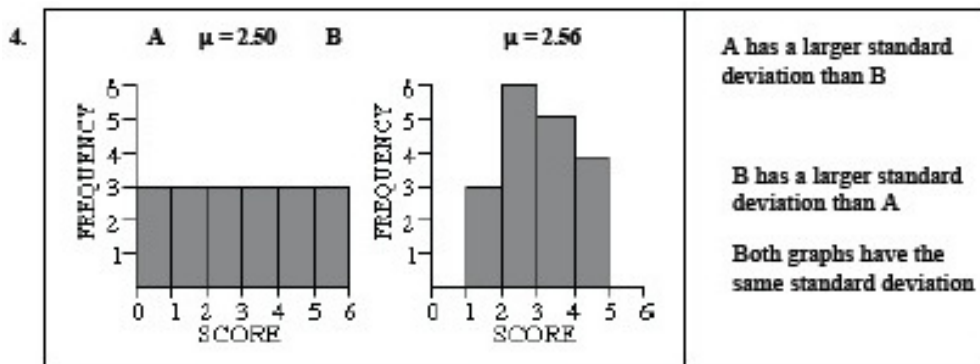
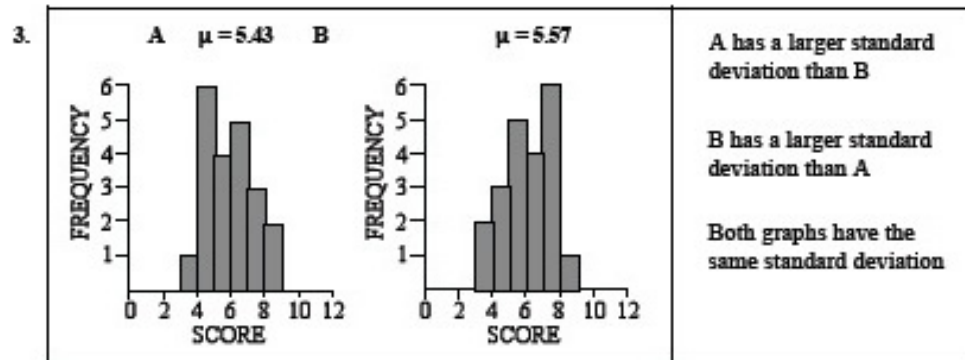
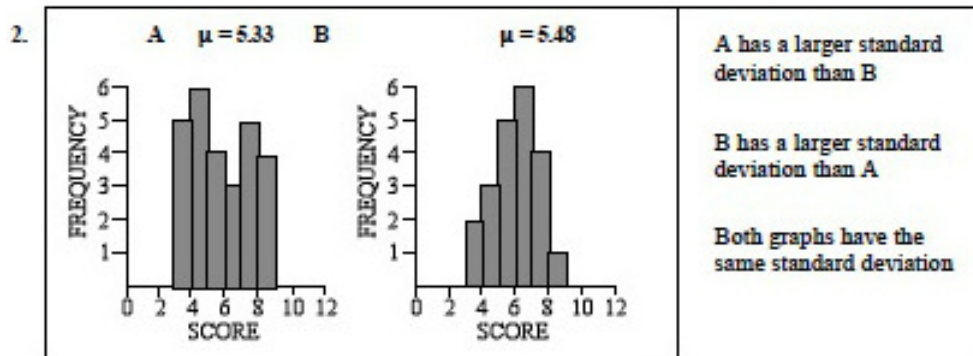
7. s depends on \bar{x} . Therefore,

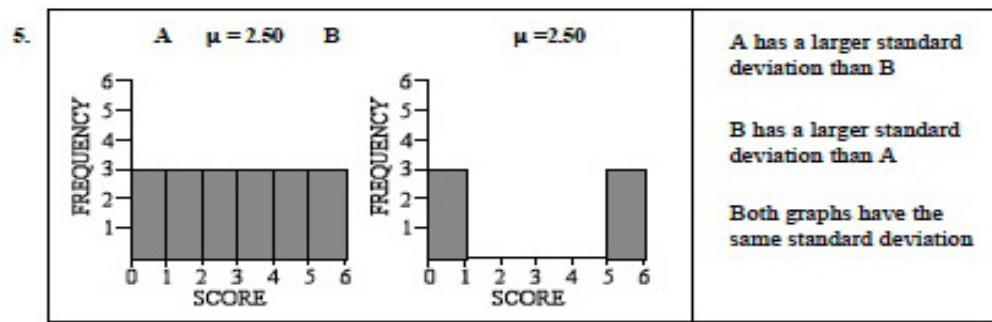
Comparing Standard deviations

Below you will find five pair of graphs. The mean for each graph μ is given just above each histogram. For each pair of graph presented,

- Indicate which one of the graphs has a larger standard deviation or if the two graphs have the same standard deviation
- Explain why. (Hint: Try to identify the characteristics of the graph that make the standard deviation larger or smaller)







Explain.

Interpreting the Magnitude of s

How big is big? That is, how can the value of s help us to evaluate whether or not a particular data set has a large spread or a small spread? The following rule helps us to answer this question *in the context of* the data set of interest.

General Rule:

Unless the data set is extremely skewed or has extreme outliers, a rough rule of thumb is that nearly all of the observations will fall within _____ s of _____.

EXAMPLE 2.4 CONTINUED

Use R to calculate the mean and standard deviation of the haircut costs for the women in the sample.

```
> mean(Haircut[Gender == "F"])
[1] 25.77358
> sd(Haircut[Gender == "F"])    #sd calculates the standard deviation
[1] 11.80065
```

From the boxplot for haircut costs for women, it appeared that the data was roughly symmetric. What does the ‘General Rule’ tell us about the spread of this data?

Almost all of the data falls between...

So?

Using the General Rule, we can see that s reflects a fairly large spread in the context of haircut costs.

EXAMPLE 2.5

In a study of repair costs for mid-sized luxury vehicles, 10 types of cars were crashed into a wall at 5 miles per hour. The following are the repair costs for these 10 vehicles.

Type	Repair Cost
Audi A6	0
BMW 328i	0
Cadillac Catera	900
Jaguar X	1254
Lexus ES300	234
Lexus IS300	979
Mercedes C320	707
Saab 95	670
Volvo S60	769
Volvo S80	4194

Use R to calculate the mean and standard deviation of the car repair costs. Repeat this analysis without the Volvo S80 (outlier).

```
> cost <- c(0,0,900,1254,234,979,707,670,769,4194)
> mean(cost)
[1] 970.7
> sd(cost)
[1] 1206.596
> mean(cost[-10]) #cost[-10] is the cost variable without the
                  #10th observation (the Volvo S80)
[1] 612.5556
> sd(cost[-10])
[1] 441.4188
```

(a) How many standard deviations away is the Volvo S80 repair cost from the average cost?

(b) Observation:

\bar{x} and s change a lot when we take out the Volvo S80. (They are not resistant to outliers.)

Choosing Numerical Summaries

We have seen that the sample median M is resistant to outliers, while the sample mean \bar{x} is not. The measure of spread we use most often in this course, the sample standard deviation s , is not resistant to outliers either.

So why do we use \bar{x} and s instead of more robust statistics?

- The mean and standard deviation involves all the values in the dataset, while the median and IQR are only determined by a few values in the dataset.
- When the dataset is large, the mean and standard deviation are easier to compute than the median and IQR, as the latter requires sorting the data.
- The mean and standard deviation have nicer theoretical properties than the median and IQR, when the distribution is not too skewed and is free from outliers.

Concluding Remarks:

The numerical summaries we have studied in this chapter (\bar{x} , s , median, etc.) are all examples of *sample statistics*. They are numbers that describe a sample taken from some population. In later chapters we will learn how we can use these sample statistics to learn about and estimate their unknown population counterparts, *population parameters*.

CHAPTER 4: GATHERING DATA

Variables: Explanatory variable (Independent variable) - Response variable (Dependent Variable)

4.1 Types of Studies

4.1.1 Experiment

A researcher conducts an experiment by assigning subjects to certain experimental conditions and then observing outcomes on the response variable.

The experimental condition are called treatments

4.1.2 Observational Study

The researcher observes values of the response variable and explanatory variables for the sampled subjects, without anything being done to the subjects (such as imposing a treatment).

Examples:

A sample survey selects a sample of people from a population and interviews them to collect data.

A census is a survey that attempts to count the number of people in the population and to measure certain characteristics about them.

In short:

an observational study merely observes rather than experiments with the study subjects. An experimental study assigns to each subject a treatment and then observes the outcome on the response variable.

4.1.3 Experiments vs. Observational Studies

An *experiment* reduces the potential for *lurking variables* to affect the result. Thus, an experiment gives the researcher more control over outside influences. Only an *experiment* can establish cause and effect. *Observational* studies can not.

Experiments are not always possible due to ethical reasons, time considerations and other factors.

4.2 Good and Poor Ways to Sample

Sampling Frame and Sampling Design

The *sampling frame* is the list of subjects in the population from which the sample is taken, ideally it lists the entire population of interest. The *sampling design* determines how the sample is selected. A *variable* is any characteristic of a subject in a population.

Simple Random Sampling (SRS)

Random Sampling is the best way of obtaining a sample that is representative of the population.

A *simple random sample* of n subjects from a population is one in which each possible sample of that size has the same chance of being selected.

A *simple random sample* is often just called a random sample.

Summary: Types of Bias in Sample Surveys

Bias: When certain outcomes will occur more often in the sample than they do in the population.

- Sampling bias occurs from using nonrandom samples or having undercoverage.
- Nonresponse bias occurs when some sampled subjects cannot be reached or refuse to participate or fail to answer some questions.
- Response bias occurs when the subject gives an incorrect response (perhaps lying) or the way the interviewer asks the questions (or wording of a question in print) is confusing or misleading.

A Large Sample Does Not Guarantee An Unbiased Sample!

Poor ways to sample

Convenience Sample: a type of survey sample that is easy to obtain.

- Unlikely to be representative of the population.
- Often severe biases result from such a sample.
- Results apply ONLY to the observed subjects.

Volunteer Sample: most common form of convenience sample.

- Subjects volunteer for the sample.
- Volunteers do not tend to be representative of the entire population.

Summary: Key Parts of a Sample Survey

- Identify the population of all subjects of interest.
- Construct a sampling frame which attempts to list all subjects in the population.
- Use a random sampling design to select n subjects from the sampling frame.
- Be cautious of sampling bias due to nonrandom samples (such as volunteer samples) and sample undercoverage, response bias from subjects not giving their true response or from poorly worded questions, and nonresponse bias from refusal of subjects to participate.

We can make inferences about the population of interest when sample surveys that use random sampling are employed.

CHAPTER 5: PROBABILITY

5.1 Randomness

Definition: random phenomenon

A phenomenon is *random* if individual outcomes are uncertain but there is a long-term regularity in the outcomes.

NOTE:

We use probability to describe and quantify randomness:

Definition: Probability

The *probability* of an outcome of a random phenomenon is the proportion of times the outcome occurs in a *very* long series of repetitions (i.e. the “long-run proportion”).

EXAMPLE 5.1: Probability as a Long-run Proportion

Flip a fair coin. We expect that half of the time we will see heads (H) and the other half of the time we will see tails (T). That is, the *probability* of flipping H is $1/2$ and the *probability* of flipping T is $1/2$.

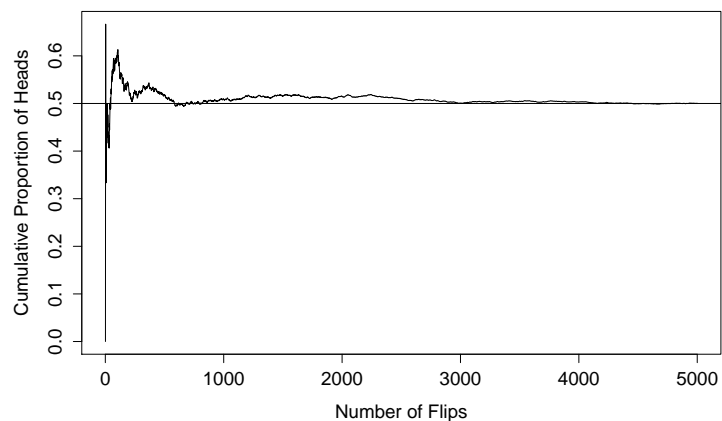
Use coin flipping to illustrate the idea of probability as a long-run proportion. Suppose we keep flipping the fair coin and keep track of the proportion of rolls that have turned up heads.

Flip	Result	Cumulative Proportion of Heads
1	T	0
2	H	0.5
3	H	
4	T	
5	T	
6	T	
\vdots	\vdots	\vdots

Keep doing this for 5000 rolls:

Flip	Result	Cumulative Proportion of Heads
1	T	0.000000
2	H	0.500000
3	H	0.666667
⋮	⋮	⋮
4998	H	0.5004002
4999	T	0.5005001
5000	H	0.5004000

Now, plot the cumulative proportion of heads vs. the number of flips:



Notice:

As the number of flips increases, the proportion of Heads

_____.

5.2 Probability Models

A probability model for a random phenomenon has two parts:

- 1.
- 2.

Definition: sample space (S)

A *sample space* is the collection of all possible outcomes of an experiment or random phenomenon.

EXAMPLE 5.2: Sample Spaces

Suppose we randomly select a student from class and ask a question. In the below situations, describe the sample space for the experiment.

1. How much time did the student spend studying during the last 24 hours?

$$S = [0, 24]$$

2. In what state was the student born, given that he/she was born in the US?

$$S = \{AL, AK, AZ, \dots WY\}$$

3. How many friends does the student have?

$$S = \{0, 1, 2, \dots\}$$

Definition: event

An *event* is a subset of the sample space.

Notation: We typically use capital letters A, B, C, etc to denote events.

EXAMPLE 5.3: Sample Spaces and Events

Suppose we toss a fair coin 3 times.

1. What is the *sample space*?

$$S =$$

2. Let A be the *event* that we get exactly 2 heads on the 3 tosses. Write down A .

$$A =$$

3. Let B be the *event* that at most one tail is flipped. Write down B .

$$B =$$

Special Events:

- complementary event: $A^c =$ “not A ”

$A^c =$ **collection of all outcomes in S that are** _____.

NOTE: $A^c \cup A =$ _____

- intersection: $A \cap B =$ “ A and B ”

$A \cap B =$ **all outcomes that are** _____.

- union: $A \cup B =$ “ A or B or both”

$A \cup B =$ **all outcomes that are** _____.

EXAMPLE 5.4: Special Events

Roll a fair die (one time).

1. Write down the sample space.

$S =$ _____

2. Let A be the event that you roll an even number and B be the event you roll a number bigger than 2. Write down the following events.

- (a) A, B

$A =$ _____ **and** $B =$ _____

- (b) A^c

$A^c =$ _____

- (c) $A \cap B$

$A \cap B =$ _____

- (d) $A \cup B$

$A \cup B =$ _____

Definition: disjoint

Two events A and B are disjoint if they have no outcomes in common.

EXAMPLE 5.4 CONTINUED

Let C be the event that you roll a value less than 3. Write down C and find $A \cap B$, $A \cap C$, and $B \cap C$. Which pairs of these events, if any, are disjoint?

$$C =$$

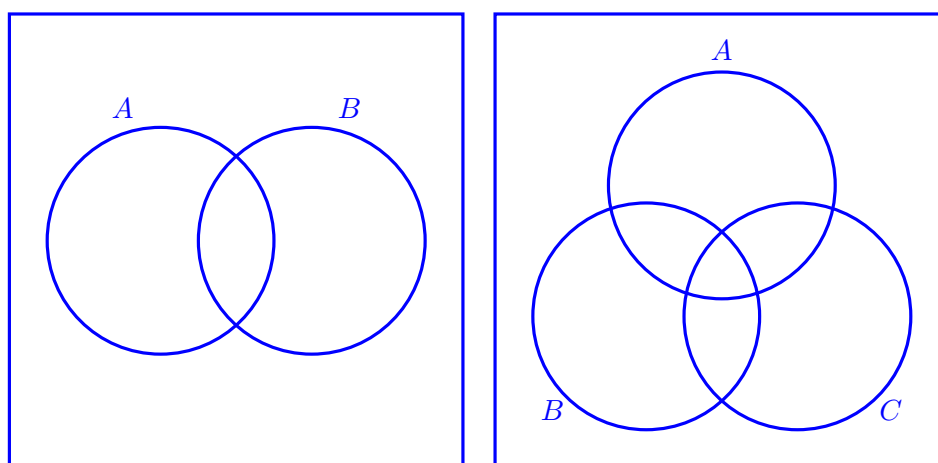
$$A \cap B =$$

$$A \cap C =$$

$$B \cap C =$$

(B, C) are disjoint but (A, B) and (A, C) are not.

Venn Diagrams: Allow us to visualize events.

**Probability Rules:**

Let A and B be events and let $P(A)$ and $P(B)$ denote the probabilities of these events occurring. Then Rules 1-5 are true for **any** A and B .

$$1. \quad 0 \leq P(A) \leq 1$$

$$P(A) = 0 \quad \Rightarrow \quad A \text{ will never occur}$$

$$P(A) = 0.0001 \quad \Rightarrow \quad A \text{ is very unlikely but } \textit{will} \text{ occur in a long series of trials}$$

$$P(A) = 0.6 \quad \Rightarrow \quad A \text{ will be observed more often than not in repeated trials}$$

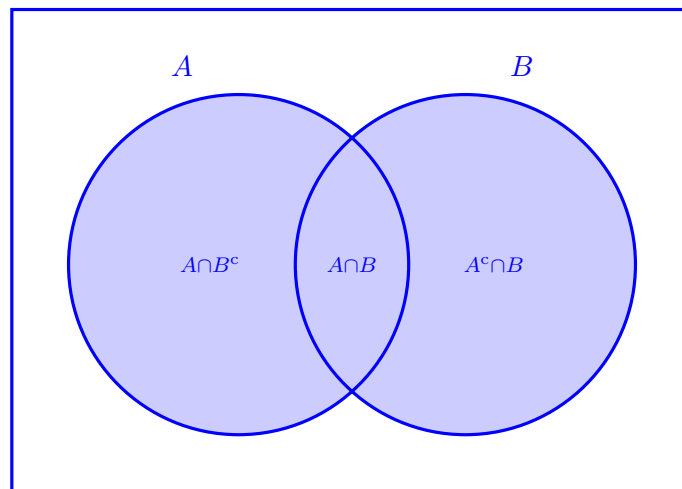
$$P(A) = 1 \quad \Rightarrow \quad A \text{ is certain to occur}$$

2. Law of Total Probability: $P(S) = \underline{\hspace{2cm}}$, where S is the sample space

3. Complement Rule: $P(A^c) = \underline{\hspace{2cm}}$

4. General Addition Rule: $P(A \cup B) = \underline{\hspace{2cm}}$

Using a Venn Diagram to Visualize the General Addition Rule:



5. Partitioning of Probability:

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Whereas Rules 1-5 hold for *any* events A and B , Rule 6 holds only if A and B are disjoint, and Rule 7 holds only if A and B are independent.

6. Addition Rule for Disjoint Events: Assume A and B are disjoint.

(a) $P(A \cap B) = \underline{\hspace{2cm}}$

(b) $P(A \cup B) = \underline{\hspace{2cm}}$

7. Multiplication Rule for Independent Events:

Definition: independent

Two events are *independent* if knowing that one occurs does not change the probability that the other occurs. For example,

- The event that it rains today is *not* independent of the event that it was cloudy this morning.
- The event that it rains today *is* independent of the event that 7 was one of the lottery numbers chosen last night.

When A and B are independent,

$$P(A \cap B) = P(A) \times P(B)$$

EXAMPLES:

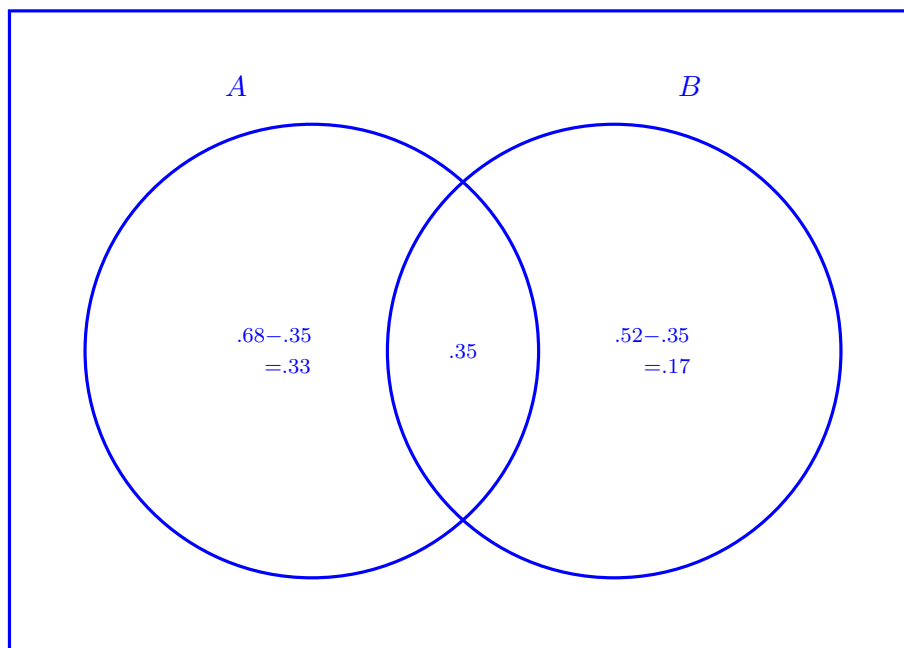
1. A survey of students found that in the last month:

68% had gone to see a movie (A)

52% had attended a sporting event (B)

35% had done both ($A \cap B$)

- (a) Draw a Venn diagram.



- (b) What is the probability that a randomly selected student has been to either a movie or a sporting event (or both) in the last month?

$$P(A \cup B) =$$

- (c) What is the probability that a randomly selected student has been to a movie but *not* a sporting event in the last month?

$$P(A \cap B^c) =$$

- (d) What is the probability that a randomly selected student has been to neither a movie nor a sporting event in the last month?

$$P((A \cup B)^c) =$$

2. SurveyUSA polled 451 Americans regarding their opinion on federal gun control laws:

		Opinion				Total
		Too Restrictive	Not Restrictive Enough	About Right	Not Sure	
Age	18–34	31	67	49	6	153
	35–54	36	82	59	3	180
	55+	21	60	33	4	118
Total		88	209	141	13	451

Select one person at random from the sample and define events A and B :

A = thinks that federal gun control laws are too restrictive

B = is under the age of 55

- (a) Find $P(A)$ and $P(B)$.

$$P(A) =$$

$$P(B) =$$

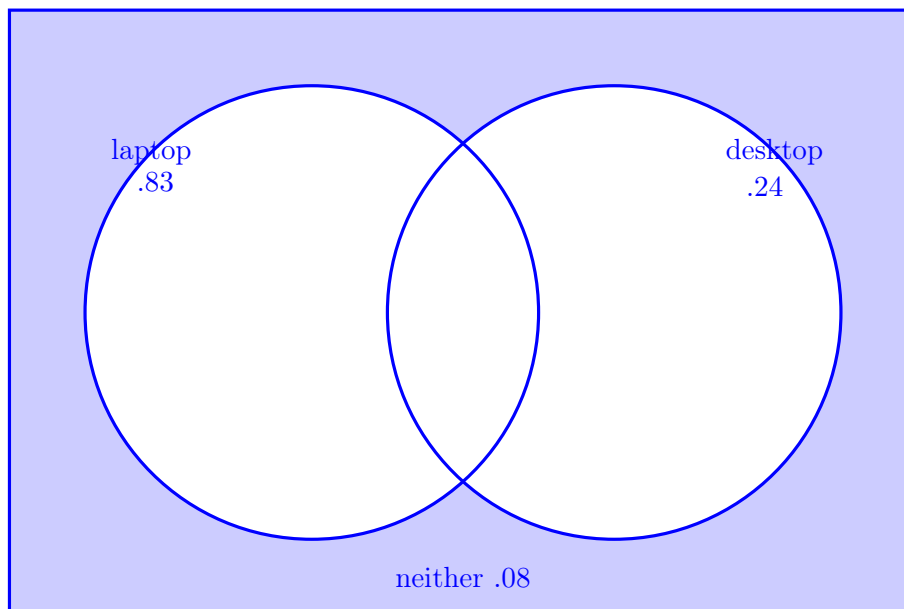
- (b) What is the probability that the person does *not* think federal gun control laws are too restrictive?

$$P(A^c) =$$

- (c) What is the probability that the person is under the age of 55 *and* thinks federal gun control laws are too restrictive?

$$P(A \cap B) =$$

3. According to an organization called Student Monitor, 83% of American college students own a laptop, 24% own a desktop, and 8% own neither a laptop nor a desktop.



- (a) What is the probability that a randomly selected student owns either a laptop or desktop or both?

$$P(\text{laptop or desktop}) = 1 - P(\text{neither}) = .92$$

- (b) What is the probability that a randomly selected student owns both a desktop and a laptop?

$$\begin{aligned} P(\text{laptop or desktop}) &= P(\text{laptop}) + P(\text{desktop}) - P(\text{both}), \\ \therefore P(\text{both}) &= P(\text{laptop}) + P(\text{desktop}) - P(\text{laptop or desktop}) \\ &= .83 + .24 - .92 = .15 \end{aligned}$$

- (c) Are owning a desktop and owning a laptop independent for the population of American college students?

$$P(\text{both}) = .15,$$

$$\text{but } P(\text{laptop}) \times P(\text{desktop}) = .83 \times .24 = .1992 \neq .15,$$

\therefore not independent.

5.3 Conditional Probability

Motivating Example:

Suppose I take a handful of M&M's and get 5 red, 4 blue, and 7 brown. I randomly pick one of the M&M's and eat it:

$$P(\text{red}) = 5/16$$

$$P(\text{blue}) = 4/16$$

$$P(\text{brown}) = 7/16$$

Now, suppose I tell you that the M&M was not brown. What is the probability the M&M was blue *given* (or knowing that) it was not brown?

That leaves 9 possible M&M's, 4 blue and 5 red.

Therefore, $P(\text{blue given not brown}) = 4/9$

Definition: conditional probability

The *conditional probability* of event A given event B is the probability that A occurs given the knowledge that B has occurred. When $P(B) > 0$, the conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Another Multiplication Rule:

Conditional probability gives us

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Therefore,

$$P(A \cap B) =$$

EXAMPLE 5.5

A study reported in the *The New Yorker* found that in the 2004 presidential election, 20% of voters identified moral values as the most important factor in their voting decision. Define events A and B where

A = identify moral issues as the most important factor in voting decision

B = vote for Bush

We know that Bush won 51% of the vote. We also know that 16% of voters both voted for Bush and identified moral issues as the most important factor. Use this information to calculate the following conditional probabilities.

- (a) What is the probability that a person who voted for Bush voted on moral issues?

$$P(A|B) =$$

Interpretation: _____ of Bush voters i.d.'d moral issues as the most important factor.

- (b) Find the conditional probability that a person voted for Bush given they voted on moral issues.

$$P(B|A) =$$

Interpretation: The majority (_____) of those who voted on moral issues voted for Bush.

- (c) Find the conditional probability that a person did not vote for Bush given they voted on moral issues.

$$P(B^c|A) =$$

Conditional Probability and Independence

Recall: If A and B are independent, knowing that B happened does *not* change the probability that A also occurs. Therefore, when A and B are independent,

$$P(A|B) = \underline{\hspace{2cm}}$$

Proof:

$$A \text{ and } B \text{ independent} \Rightarrow P(A \cap B) = P(A)P(B)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Three Ways to Check For Independence:

Two events A and B are independent if any of the following holds (you only have to check one):

1. $P(A \cap B) = P(A)P(B)$
2. $P(A|B) = P(A)$
3. $P(B|A) = P(B)$

If any one of these conditions do not hold, then none of them hold and the events are NOT independent.

More Conditional Probability Examples

- (a) In the gun control example on p. 42, what is $P(A|B)$? Interpret your answer in a sentence.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.149}{.738} = .202$$

Interpretation: 20.2% of Americans under the age of 55 think that federal gun control laws are too restrictive.

- (b) In the laptop/desktop example (p. 43), what proportion of students who own a laptop also own a desktop?

$$P(\text{desktop}|\text{laptop}) = \frac{P(\text{desktop} \cap \text{laptop})}{P(\text{laptop})} = \frac{.15}{.83} = .181$$

CHAPTER 6: PROBABILITY DISTRIBUTIONS

EXAMPLE 6.1

Toss a coin 3 times. By now we know that the sample space for this event is

$$S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}.$$

Let X = the number of heads (H). Therefore, X can be 0, 1, 2, or 3. The event that X takes on any one of these values is *random* and can be assigned a probability. So, X is a variable and X is random... X is a *random variable*!

Definition: random variable

A *random variable* is a variable whose value is a numerical outcome of a random phenomenon.

Notation:

We often use X , Y , and Z to denote random variables and use x , y , and z to denote their realized/observed values, respectively.

Two Types of Random Variable

1. **discrete:**
2. **continuous:**

6.1 Discrete Random Variables

A **discrete** random variable takes on values that can be listed.

examples:

X = # of M&M's in a bag

X = # of broken mirrors in a shipment

X = # of accidents/day in a factory

We can describe discrete random variables using its probability distribution...

6.1.1 Probability Distribution of A Discrete Random Variable

Probability distributions for discrete random variables have two properties:

- 1.
- 2.

EXAMPLE 6.2

Let X = the number of bases for a randomly selected at bat. In 2015, the probability distribution of X (excluding walks) for a Minnesota Twins player was as follows (all probabilities rounded to three decimals):

x	0	1	2	3	4
probability	.753	.159	.051	.008	.029

- (a) Is this a legitimate probability distribution?
- (b) For a randomly selected at bat, what is the probability the player got at least one base?

6.1.2 Center and Spread of A Probability Distribution

We can describe distributions of random variables just as we described distributions of data in Chapter 2. Specifically, we can describe the center and spread of a probability distribution using mean and standard deviation.

- Mean:

$$\mu = E(X) = \text{“Expected Value of } X\text{”}$$

Measures the center tendency of the distribution of X

- Standard Deviation:

$$\sigma = \text{“Standard Deviation of } X\text{”}$$

Measures the spread of the distribution of X .

Interpretation:

μ is the “long-run” average outcome. That is, μ is what we expect the average to be for a *very* long series of repetitions. Similarly, σ is a “long-run” standard deviation.

Calculating μ and σ for Discrete Distributions

Let x represent the possible outcomes of discrete random variable X . Then the formulas for the mean and standard deviation of the probability distribution for X are as follows:

$$\mu =$$

$$\sigma^2 =$$

$$\sigma =$$

EXAMPLE 6.2 CONTINUED

Recall: We let X be a random variable defined by the number of bases earned by a randomly selected Minnesota Twins at-bat from the 2015 season. The probability distribution of X is:

x	0	1	2	3	4
probability	.753	.159	.051	.008	.029

Find the mean and the standard deviation of this probability distribution.

If we define a similar random variable Y for the Rangers, who hit many more home runs than the Twins, would we expect the standard deviation of Y to be smaller or larger than the standard deviation of X ?

Larger. Since Y has a much higher probability to take the value of 4 than X , the distribution of Y is more spread out than that of X , and therefore Y has a larger standard deviation.

EXAMPLE 6.3

Let X denote the response of a randomly selected person to the question, “What is the ideal number of children for a family to have?”

According to a recent General Social Survey, the probability distribution of X for men in the U.S. is as follows:

x	0	1	2	3	4
probability	0.04	0.03	0.57	0.23	0.13

Find and interpret the mean of this probability distribution.

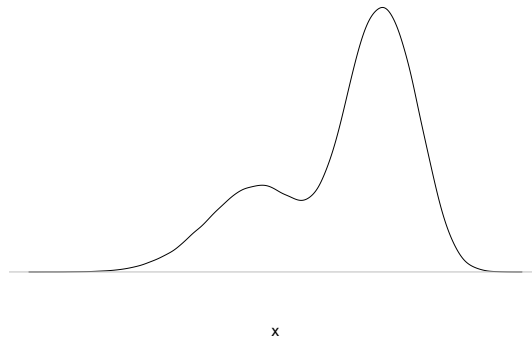
If we sample a large number of American men and compute their average ideal numbers of children, then in the long run the average will be close to _____.

6.2 Continuous Random Variables

6.2.1 Density Curves

A continuous random variable takes on values that form an interval. It is impossible to list all its values. Therefore, instead of writing out the probability distribution as we did for discrete random variables, we describe the distribution using a density curve.

A *density curve* specifies the probability distribution of a continuous random variable.



Properties of a Density Curve:

1. **always non-negative (≥ 0)**
2. **$P(a < X < b) = \text{area under the curve above } (a, b)$ (this sometimes requires calculus)
Therefore $P(X = a) = 0$ for all a**
3. **total area under density curve = 1**

EXAMPLE 6.4

A certain bus is equally likely to be anywhere from zero to twenty minutes late. Let $X =$ the number of minutes that the bus is late. Therefore X is uniformly distributed on the interval $[0,20]$.

1. Draw the density curve for X .



Total area = 1.

$$f(x) = \begin{cases} 1/20 & 0 \leq x \leq 20 \\ 0 & \text{otherwise} \end{cases}$$

2. What is the probability the bus is...

(a) At least 13 minutes late?



(b) Exactly 10 minutes late?



(c) Between 10 and 13 minutes late?

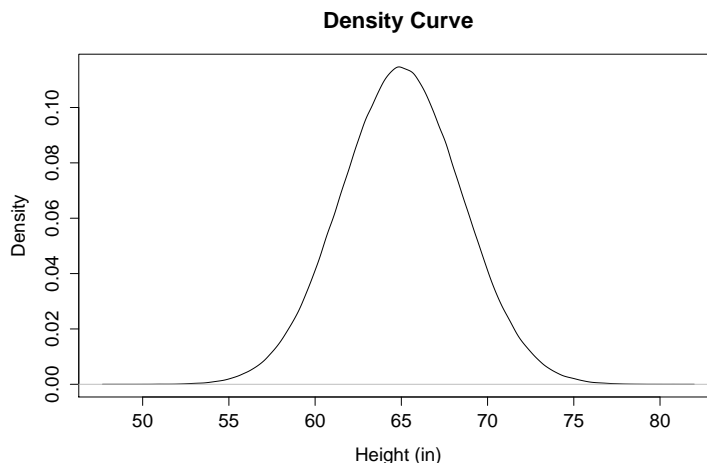


3. Exactly 75% of the waiting times are below what value?



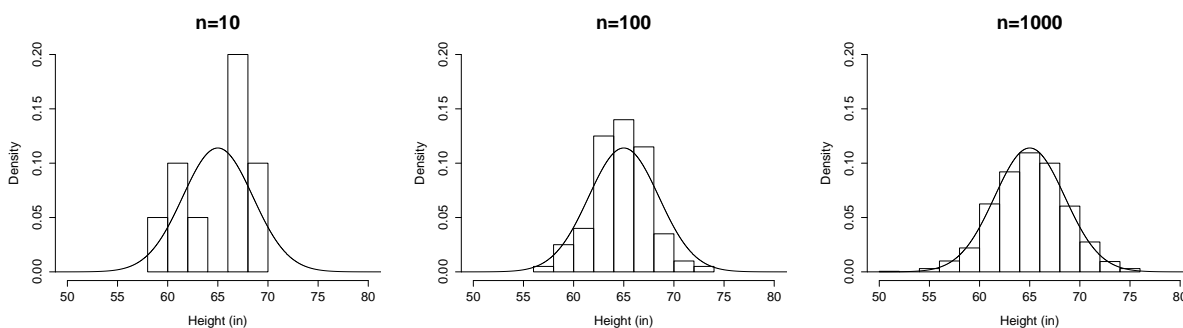
EXAMPLE 6.5: Histograms as Discrete Approximations of Density Curves

Let $X =$ height (in inches) of a North American female. The following density curve specifies the probability distribution of X :



Now, suppose we don't have this information. All is not lost! To get an idea of what the distribution of heights is, we can randomly sample n North American women and measure their heights.

The following are histograms for these samples where $n = 10$, $n = 100$, $n = 1000$ with the true density curve superimposed.



Observations:

As sample size increases, the shape of the histogram for the sample approaches the true density curve.

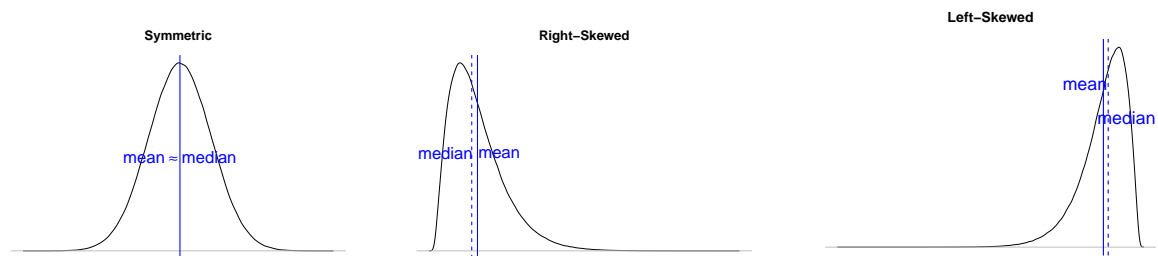
In general: The more data and finer the scale on the x -axis of a histogram, the better it approximates the true density curve.

CENTER & SPREAD OF A DENSITY CURVE

Measures of Center:

1. median =
2. mean =

Comparing the Mean and Median of a Density Curve



Measures of Spread:

We can measure the spread of a distribution using standard deviation. This value is tough to eyeball but can be calculated mathematically. We will return to this later...

6.2.2 The Normal Distribution

Notation Recall:

- μ = mean of a probability distribution
- σ = standard deviation of a probability distribution
- \bar{x} = sample mean
- s = sample standard deviation

The Normal Distribution

We now focus on a common continuous distribution, the *normal distribution*. The distributions of many quantitative random variables are well-approximated by the normal distribution. It will therefore become an indispensable tool in approaching statistical inference.

Normal distributions are specified by μ and σ for

$$-\infty < \mu < \infty \quad \text{and} \quad 0 < \sigma < \infty.$$

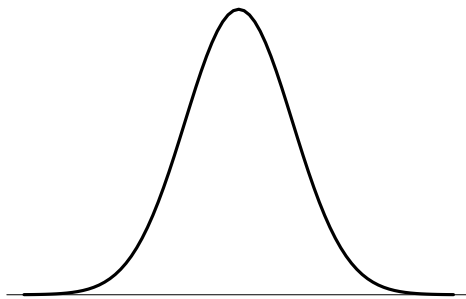
That is, if we know that a random variable has a normal distribution and we also know μ and σ , then we know exactly what the probability distribution looks like.

Notation:

$X \sim N(\mu, \sigma)$: X is “normally distributed” with mean μ and standard deviation σ

$Z \sim N(0, 1)$: Z is “standard normal” (normal with $\mu = 0$ and $\sigma = 1$)

The Normal Distribution Density Curve



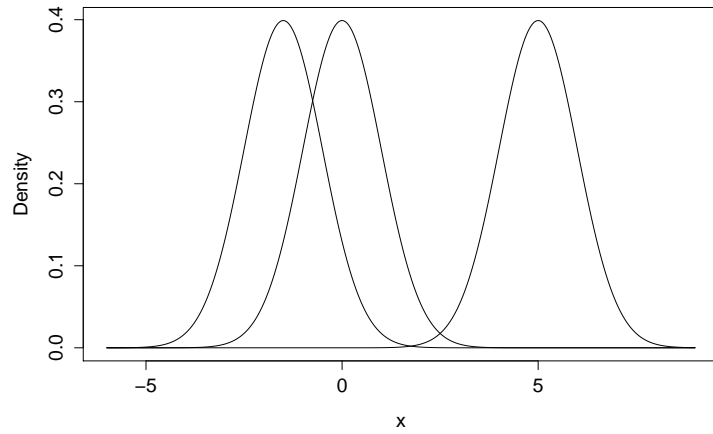
Features:

1. bell-shaped
2. symmetric
3. centered at μ (mark this on the plot)
4. mean = median = μ

NOTE: All normal curves have the same overall shape. So how does changing μ or σ affect the normal curve?

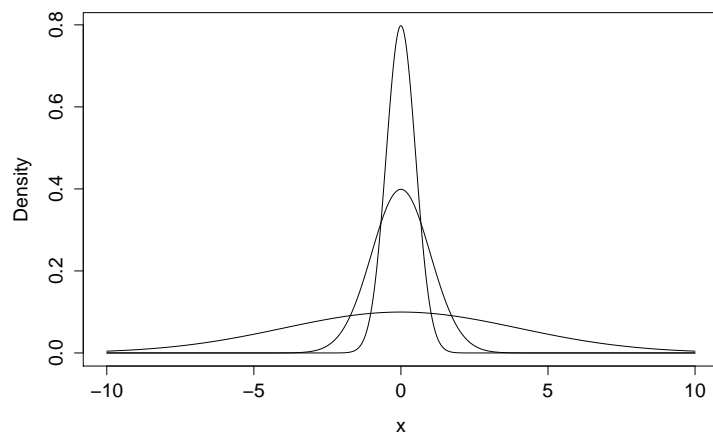
Changing μ :

Observations:



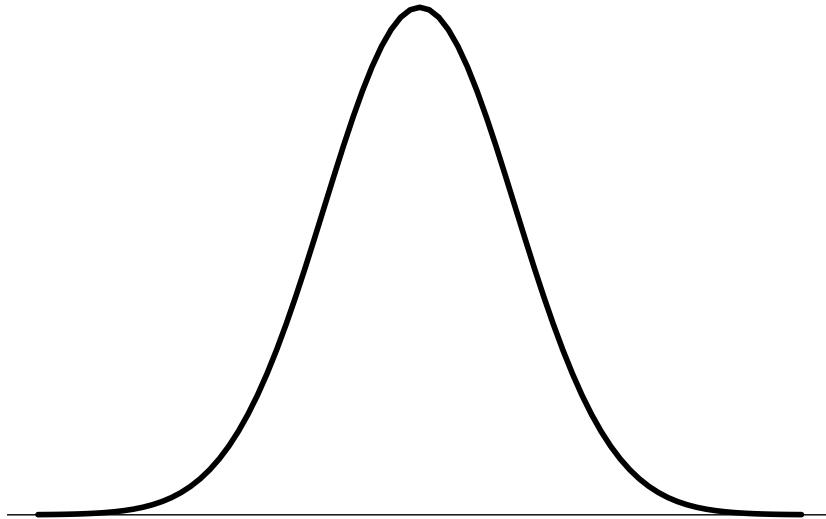
Changing σ :

Observations:



The 68–95–99.7 Rule

All normal distributions share common properties. One of these is the 68–95–99.7 Rule.



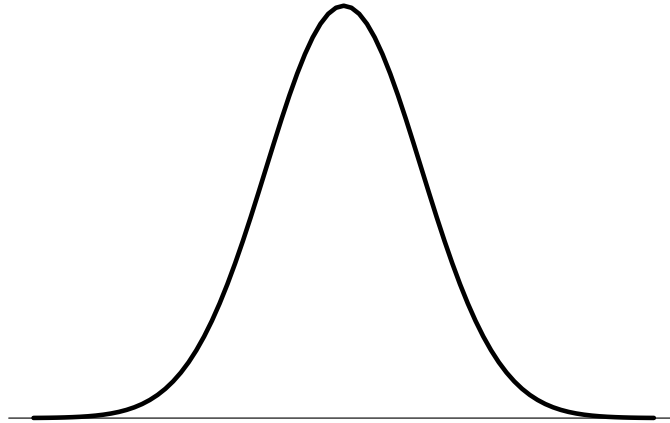
For **any** normal distribution

- Approximately _____ of the distribution falls within _____ σ of μ .
- Approximately _____ of the distribution falls within _____ σ of μ .
- Approximately _____ of the distribution falls within _____ σ of μ .

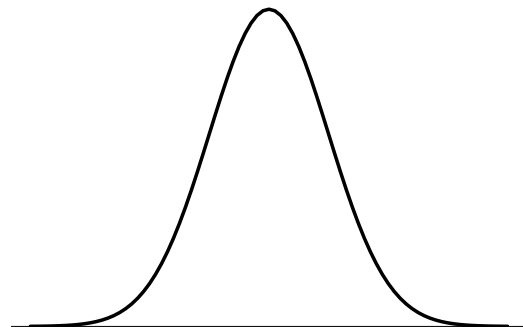
EXAMPLE 6.6

The time a customer has to wait for their food to arrive at a local restaurant has a normal distribution with a mean of 16 minutes and standard deviation of 4 minutes.

68–95–99.7 Rule:



- (a) What proportion of customers wait longer than 16 minutes?
- (b) What proportion of customers wait between 12 and 20 minutes?
- (c) What is the probability that a customer waits between 12 and 24 minutes?



(d) The shortest 2.5% of waiting times are smaller than what value?

(e) What is the probability that a customer waits less than 21 minutes?

Normal Distribution Calculations

Goal: Calculate quantities such as $P(X \leq a)$, $P(a < X \leq b)$, etc using the normal curve. For problems not covered by the 68–95–99.7 Rule, we will need to use Table A (in Appendix A).

Notice: There is only one table but infinitely many different normal distributions (think of all the combinations of μ and σ !). In order to use the same table for each of these distributions, we can *standardize* the distributions so that they are on the same scale.

Standardizing Normal Distributions

Suppose $X \sim N(\mu, \sigma)$ for *any* μ, σ . Then we can transform X to the standard normal scale:

Standardizing Observations

Definition: z -score

Let x be an observation from a normal distribution with mean μ and standard deviation σ . Then

$$z = \frac{x - \mu}{\sigma}$$

is the z -score of x .

Interpreting a z -score:

1. **By definition, $z =$ number of standard deviations (σ) that x is away from the mean (μ).**
2. **Almost all observations will be within 3σ of μ . Therefore, any observation with a z -score close to or more than 3 is an unusual observation (it falls far from μ).**

The Implication:

Suppose $X \sim N(\mu, \sigma)$ and we want to calculate $P(X < a)$ for some value a . Then, the standardization scheme guarantees that

where z^* is the z -score for a . Therefore, after standardizing any normal distribution, we can use the standard normal distribution for probability calculations.

Table A

Table A can be used to calculate areas under the curve for the standard normal distribution ($N(0, 1)$). Specifically, Table A allows us to approximate $P(Z < z^*)$ for any value z^* where $Z \sim N(0, 1)$:

Table A (page A-1):

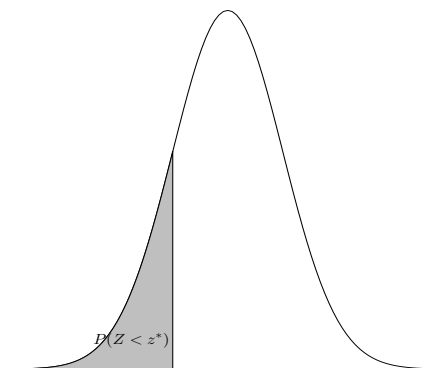
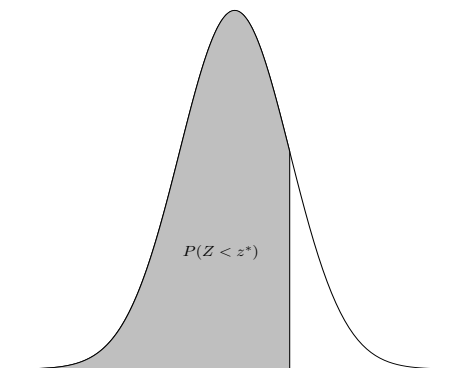


Table A (page A-2):

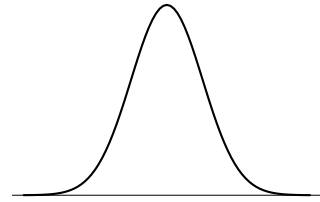


After standardizing, this table can be used to approximate *any* area beneath the normal curve corresponding to *any* normal distribution.

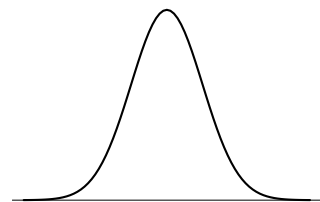
EXAMPLE 6.7

Let Z be a standard normal random variable. That is, $Z \sim N(0, 1)$.

- (a) What is the probability Z falls below -2.63 ?



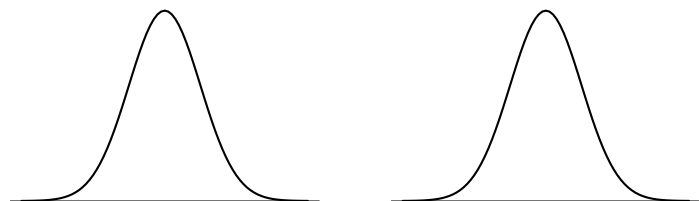
- (b) What is the probability Z is at least 2.63 ?



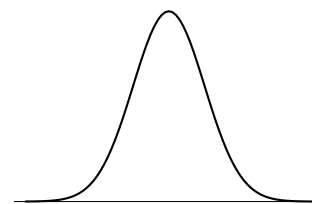
- (c) What is the probability Z is greater than -1.31 ?



- (d) Find the proportion of the distribution that falls between -0.97 and 1.31 .



- (e) What value marks the 57th percentile?



Steps for Calculating Normal Probabilities:

When given a value x and asked to find a probability or proportion, there are 3 steps to follow:

1. Draw a picture.
2. Convert x to a z -score.
3. Use Table A to calculate the appropriate probability. (Start from the outside and work in!)

Steps for Calculating x Values of a Normal Distribution:

When given a probability or proportion and asked for the corresponding x value, there are 3 steps to follow:

1. Draw a picture.
2. Use Table A to find the z -score for the specified probability.
(Start from the inside and work out!)
3. Unstandardize. Recall that $z = (x - \mu)/\sigma$. Therefore you must calculate $x = \mu + z\sigma$.

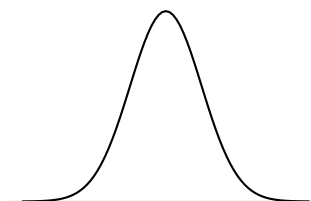
EXAMPLE 6.8

Let $X =$ SAT score and suppose that SAT scores are known to follow a normal distribution with mean 1026 and standard deviation 209.

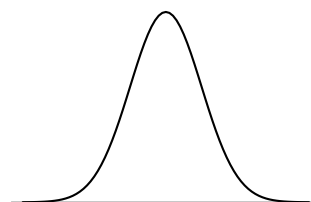
- (a) Write down the distribution of X .

(b) Calculate and interpret the z -score for an SAT performance of 1100.

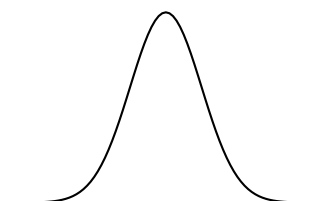
(c) What proportion of students score lower than 1100?



(d) What is the probability that a randomly selected student received a score of at least 820?



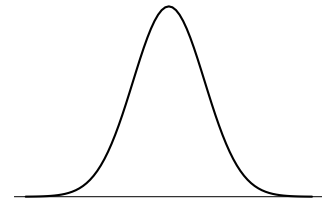
(e) What score must you earn to be in the top 10%?



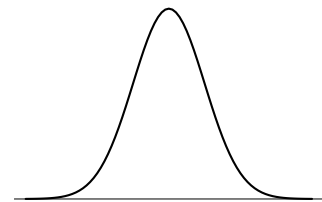
EXAMPLE 6.9

Let X = duration of a human pregnancy (in days). The distribution of X is well-approximated by a normal distribution with a mean = 266 days and standard deviation = 16 days.

- (a) What is the probability that the duration of the pregnancy is between 250 and 300 days?



- (b) How long do the shortest 2% of pregnancies last?

**Assessing Normality**

Not every continuous random variable follows a normal distribution. There is also no way to *guarantee* that a random variable is normally distributed. However, given a sample of observations from a certain distribution, there are some graphical tools that can help us to assess whether or not it is reasonable to *assume* normality.

1. **Histogram / boxplot: Check for skewness and outliers.**
2. **Q-Q plot: Check of skewness, outliers, and heavy-tailedness.**

The Histogram or Boxplot as a Graphical Tool for Assessing Normality

- Normality Assessment:

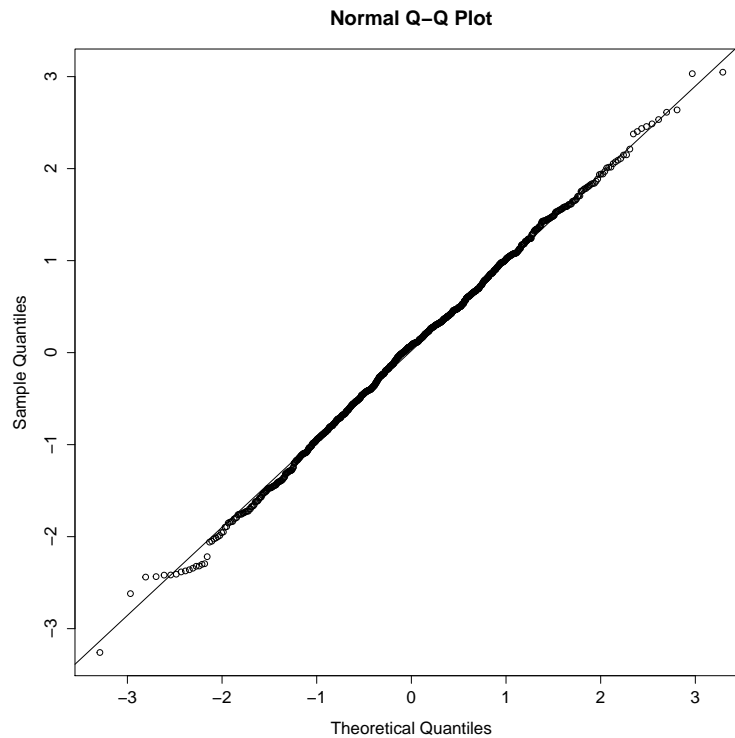
The normal distribution has a symmetric, bell-shaped shape. If the plot does not look symmetric and bell-shaped, then the data is skewed and/or contains outliers.

The QQ Normal Plot as a Graphical Tool for Assessing Normality

- Normality Assessment:

Plots the quantiles of the data against the quantiles of a normal distribution. Then the data are approximately normally distributed if ...

the points lie along the plotted line.

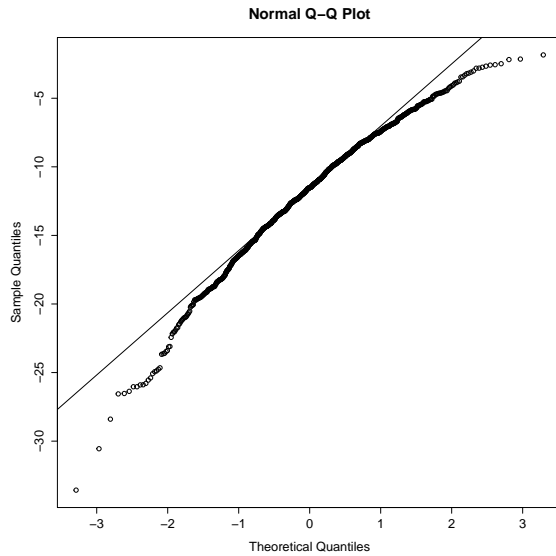


To create this plot, the following R commands were used:

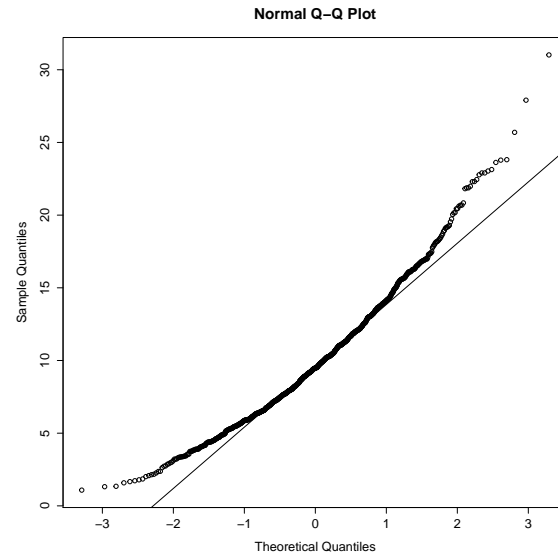
```
> samp <- rnorm(1000)
> qqnorm(samp)
> qqline(samp)
```

The first command created a random sample of 1000 observations ($n = 1000$) from $N(0, 1)$. The second command plotted the quantiles of the sample against the quantiles of the normal distribution. Then the third command created the line that we compare our points to.

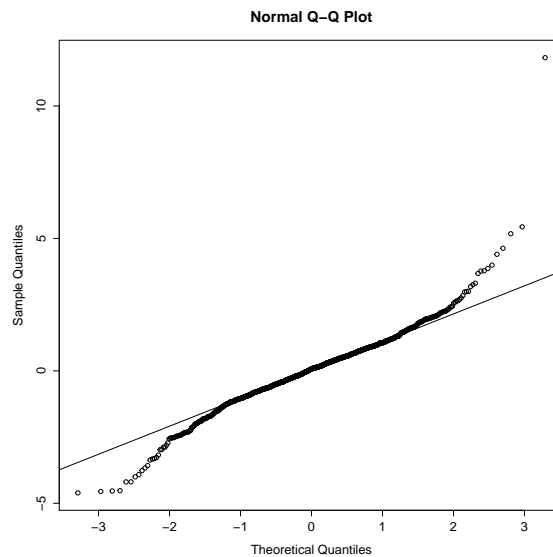
Obvious deviations of the points from the line may indicate non-normality in the data. For example, the following pictures show the Q-Q plots of data from three typical non-normal distributions.



Left skewed



Right skewed



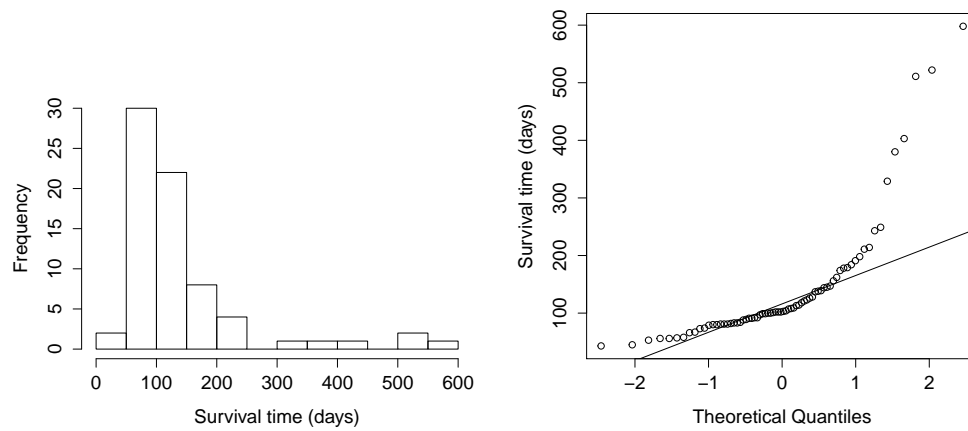
Heavy tailed

EXAMPLE 6.10

In the following two examples, we have two samples from unspecified distributions. Use histograms and Q-Q plots to assess the normality of these distributions.

1. The survival times (in days) are recorded for 72 guinea pigs in a medical experiment. The data can be found at <http://www.stat.umn.edu/~wuxxx725/data/guineapigs.txt>.

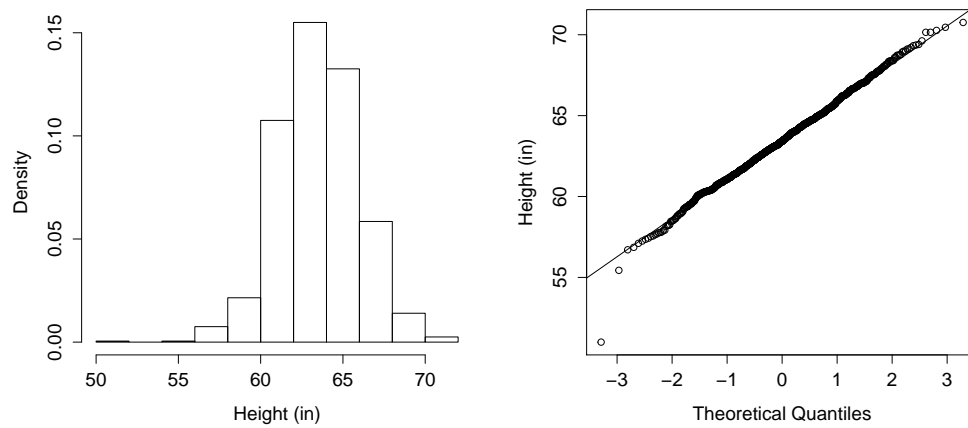
```
> hist(Survival, xlab="Survival time (days)", main="")
```



Assessment:

2. The heights of a random sample of 1000 women were recorded.

```
> hist(Height, xlab="Height (in)", main="", freq=F)
```



Assessment:

6.2.3 The Normal Distribution in R

Normal Distribution Calculations

1. Assume $Z \sim N(0,1)$.

(a) Calculate $P(Z \leq 1.645)$.

```
> pnorm(1.645)
[1] 0.950015
```

(b) Calculate $P(Z \geq -0.971)$.

```
> 1 - pnorm(-0.971)
[1] 0.8342259
> pnorm(-0.971, lower.tail=F)
[1] 0.8342259
```

NOTE: R calculates lower tail probabilities by default!

(c) Find the 97.5th percentile.

```
> qnorm(0.975)
[1] 1.959964
```

2. Assume $X \sim N(21,2)$.

(a) Calculate $P(X < 19)$.

```
> pnorm(19, mean=21, sd=2)
[1] 0.1586553
```

(b) Calculate $P(17 < X < 25)$ (within 2 sd of the mean).

```
> pnorm(25, mean=21, sd=2) - pnorm(17, mean=21, sd=2)
[1] 0.9544997
```

(c) Find the 60th percentile.

```
> qnorm(.60, mean=21, sd=2)
[1] 21.50669
```

Generating Normal Data

```
# random sample of 100 observations (n=100) from N(0,1)
> x <- rnorm(100)
# random sample of 10 observations (n=10) from N(90,5)
> y1 <- rnorm(10,mean=90,sd=5)
# random sample of 100 observations (n=100) from N(90,5)
> y2 <- rnorm(100,mean=90,sd=5)
# random sample of 1000 observations (n=1000) from N(90,5)
> y3 <- rnorm(1000,mean=90,sd=5)
> hist(y1)
> hist(y2)
> hist(y3)
```

CHAPTER 7: SAMPLING DISTRIBUTIONS

Recall from Chapter 1:

parameter: a number that describes a population

statistic: a number that describes a sample

inference: drawing conclusions about a population based on information from a sample

The Problem:

In the last chapter we studied the normal distributions, $N(\mu, \sigma)$. In doing so, we assumed that the *parameters* of these distribution, μ and σ , were known. However, in practice we will rarely know these values.

The Solution:

EXAMPLE 7.1

Let p = the proportion of Americans that approve of the job President Obama is doing. Notice that p is an unknown parameter, it would be impossible to determine Obama's *exact* approval rating.

Suppose ABC polled 1000 adults and 512 of them said they approved of the job Obama is doing. How would you use this information to estimate the *true* approval rating, p ?

Now, suppose CBS conducted their own poll of 1000 adults and found that 485 approved of the job Obama is doing. How would you estimate p based on this sample?

Which estimate of p is correct?

THE MAIN POINT: Different samples produce different results!

The value of a statistic (ex: \bar{x} , \hat{p}) will vary from sample to sample. Therefore statistics have their own distributions.

Reese's Pieces Part 1: Making Conjectures about Samples



Reese's Pieces candies have three colors: orange, brown, and yellow. Which color do you think has more candies (occurs more often) in a package: orange, brown, or yellow?

Guess the proportion of each color in a bag?

Color	Orange	Brown	Yellow
Predicted			
Proportion			

1. If groups (of 2-4 students) in the class take a sample of 25 Reese's Pieces candies, would you expect every group to have the same number of orange candies in their sample? Explain.
2. Make a conjecture: Pretend that 10 groups each took samples of 25 Reese's Pieces candies. Write down the number of orange candies you might expect for these 10 samples:

--	--	--	--	--	--	--	--	--	--

These numbers represent the **variability** you would expect to see in the number of orange candies in 10 **samples** of 25 candies.

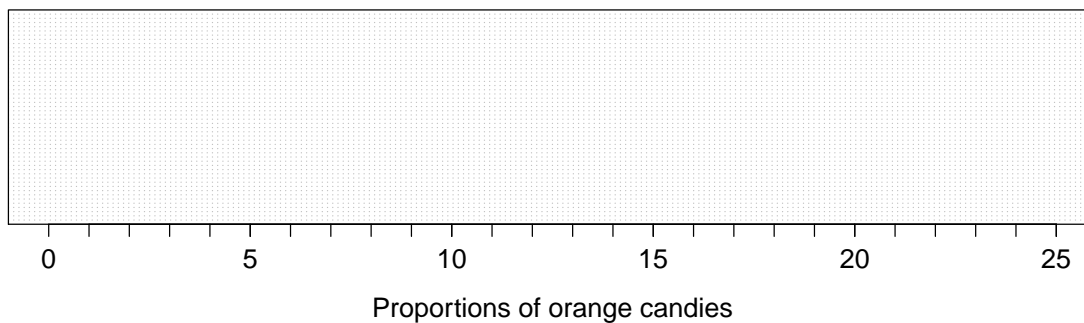
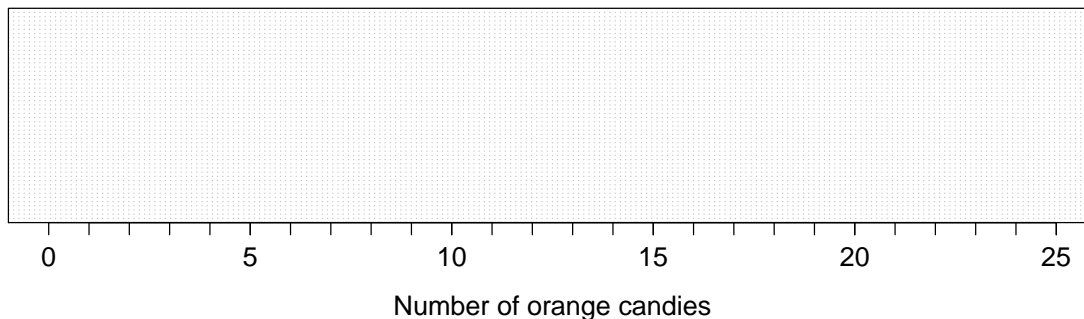
Your group will be given a cup that is a **random sample** of Reese's Pieces candies. Count out 25 candies from this cup without paying attention to color. In fact, try to **IGNORE** the colors as you do this.

3. Now, count the colors for your sample and fill in the chart below:

	Orange	Yellow	Brown	Total
Number of candies				
Proportion of candies				

Record both the number and proportion of orange candies on the board.

- Now that you have taken a sample of candies and see the proportion of orange candies, make a second conjecture: *If you took a sample of 25 Reese's Pieces candies and found that you had only 5 orange candies, would you be surprised?* Do you think that 5 is an unusual value?
- Record the number AND the proportion of orange candies in your sample on two dotplots on the board. Recreate both dotplots in the two figures below.



Part 2: Compare Sample Statistics to the Population Parameter

Discuss the following **Things to Consider** questions with your group. Be prepared to report back to the class.

Things to Consider

The proportions you have calculated are the sample statistics. For example, the proportion of orange candies in your sample is the statistic that summarizes your sample.

- Did everyone in the class have the same number of orange candies?
- How do the actual sample values compare to the ones you estimated earlier?
- Did all groups have the same proportion of orange candies?
- Describe the variability of the distribution of sample proportions on the board in terms of shape, center, and spread.
- Do you know the proportion of orange candies in the population? In the sample?
- Which one can we always calculate? Which one do we have to estimate?
- Does the value of the parameter change, each time you take a sample?
- Does the value of the statistic change each time you take a sample?
- How would this sample proportion compare to the population parameter (the proportion of all orange Reese's Pieces candies produced by Hershey's Company that are orange)?
- Based on the distribution we obtained (on the board), what would you ESTIMATE to be the population parameter, the proportion of orange Reese's Pieces candies produced by Hershey's Company?
- What if the groups in the class only took 10 candies in their sample instead of 25? Do you think the graphs on the board would look the same? If not, how would they change?
- What if the groups in the class each took 100 candies? Would the distributions on the board change at all? If so, how?

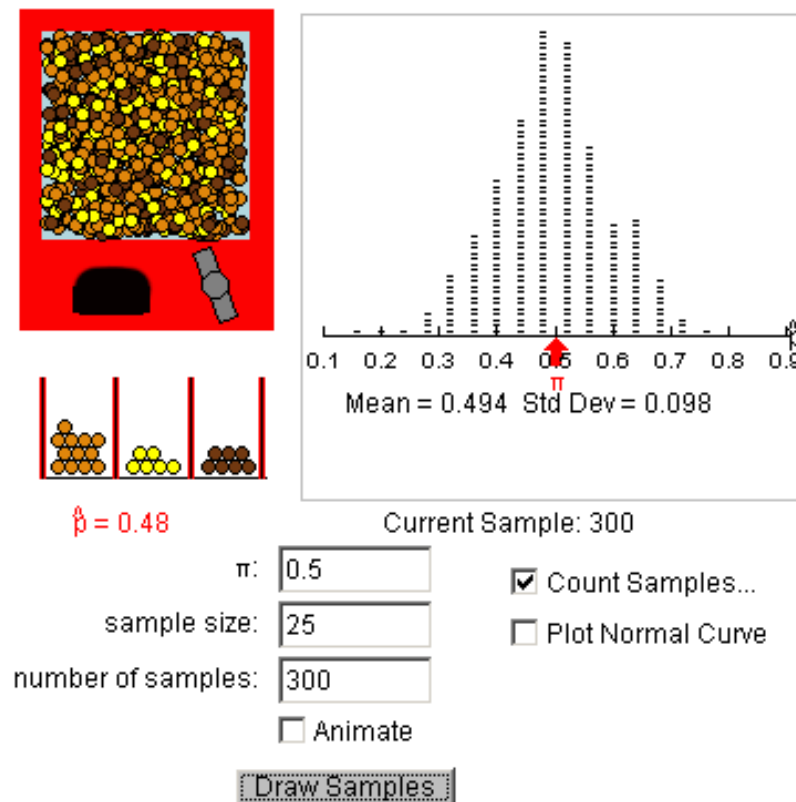
Part 3: Simulate the Sampling Process

We will now simulate additional data and tie this activity to the Simulation Process Model (SPM).

- Use the Web Applet: Reese's Pieces link below
- <http://www.rossmanchance.com/applets/OneProp/OneProp.htm?candy=1>

You will see a big container of colored candies that represents the POPULATION of Reese's Pieces candies.

Sampling Reese's Pieces



6. What is the proportion of orange candies in the population? (Note: In class we didn't know the parameter value but one catch in running a computer simulation is that we have to assume a value.)

You will see that the proportion of orange is already set at 0.5, so that is the **population parameter**. (People who have counted lots of Reese's Pieces candies came up with this number.)

7. How does 0.5 compare to the proportion of orange candies in your sample? Explain.

8. How does it compare to the center of the class distribution? Does it seem like a plausible value for the population proportion of orange candies? Explain.

Simulation

- Click on the “Draw Samples” button in the Reese’s Pieces applet. One sample of 25 candies will be taken and the proportion of orange candies for this sample is plotted on the graph.
 - Repeat this again. (Draw a second sample.)
9. Do you get the same or different values for each sample proportion?

 10. How do these numbers compare to the ones our class obtained?

 11. How close is each **sample statistic** (proportion) to the **population parameter**?

Further Simulation

- Uncheck the “Animate” box.
 - Change the number of samples (num samples) to 500.
 - Click on the “Draw Samples” button, and see the distribution of sample statistics (in this case proportions) build.
12. Describe the shape, center and spread of the distribution of sample statistics.
13. How does this distribution compare to the one our class constructed on the board in terms of shape? Center? Spread?
14. Where does the value of 0.2 (i.e., 5 orange candies) fall in the distribution of sample proportions? Is it in the tail or near the middle? Does this seem like a rare or unusual result?

Part 4: Examine the Role of Sample Size

Next we consider what will happen to the distribution of sample statistics if we change the number of candies in each sample (change the sample size).

Make a Conjecture

- 15 What do you think will happen to the distribution of sample proportions if we change the sample size to 10? Explain.

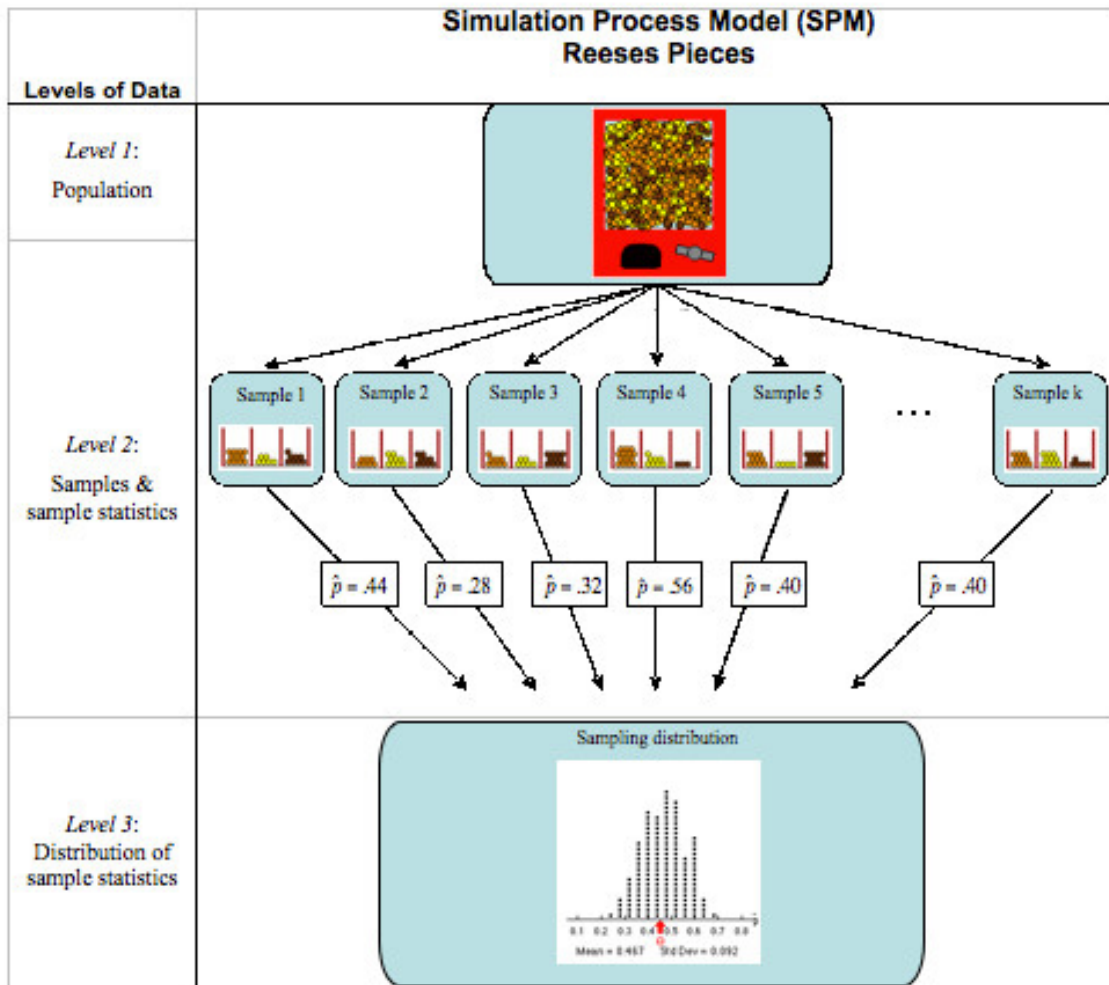
16. What do you think will happen if we change the sample size to 100? Explain.

Test your conjecture

- Change the “sample size” in the Reese’s Pieces applet to 10.
 - Be sure the number of samples (num samples) is 500.
 - Click on the “Draw Samples” button.
17. How close are the **sample statistics** (proportions), in general, to the **population parameter**?

- Change the “sample size” in the Reese’s Pieces applet to 100, and draw 500 samples.
 - Be sure the number of samples (num samples) is 500.
 - Click on the “Draw Samples” button.
18. How close are the sample statistics (proportions), in general, to the population parameter?
19. As the sample size increases, what happens to the distance the sample statistics are to the population parameter?
20. Now, describe the effect of sample size on the distribution of sample statistics in terms of shape, center and spread.

Note: When we generate sample statistics and graph them, we are generating an estimated **sampling distribution**, or a distribution of the sample statistics. It looks like other distributions we have seen of raw data.



Definition: Sampling Distribution

The *sampling distribution* of a statistic is the probability distribution that describes the possible values the statistic can take and assigns probabilities for those values. It is used to describe how the values of a statistic vary in *all* possible samples of the *same size* from the same population.

EXAMPLE 7.2

Jim just got a new haircut and wants to know what his family thinks of it. That is, Jim is interested in p , the true proportion of his family members that like his haircut. Their true opinions are:

Relative	Mom (M)	Dad (D)	Sister (S)	Brother (B)
Opinion	Y	Y	N	N

(Y=Yes, they like it; and N=No, they do not like it)

We can see that $p = 0.5$. However, Jim only has time to ask the opinion of 3 of his family members.

- (a) List all possible samples and for each sample calculate \hat{p} , the proportion of the sample that likes Jim's haircut.

- (b) Use part a to find the sampling distribution for the sample proportion, \hat{p} .

Sampling Distribution:

\hat{p}	1/3	2/3
probability	1/2	1/2

Notice $P(\hat{p} = p) = 0!$

- (c) Find the mean and standard deviation of this sampling distribution.

Although sampling distributions can be constructed for *any* statistic (including medians, minimums, maximums, etc.), we will focus on the sampling distributions for the sample mean and sample proportion.

Recall:

- A proportion is a value between 0 and 1. When multiplied by 100, it can be interpreted as a percentage.

Examples:

The proportion of students that are late, that get A's, etc.

- A sample mean is the average of a set of data.

Examples:

The mean number of hours per week spent on studying, the mean amount of money per week spent on food, etc.

7.1 The Sampling Distribution of A Sample Mean

Notation:

μ = population mean

σ = population standard deviation

\bar{x} = sample mean

Statistical Inference:

EXAMPLE 7.3

Suppose we want to determine the average income of the students in this class but do not have time to ask every single person. On average, would we rather make our estimate based on a sample of 2 students or a sample of 10 students?

Unless we take a very unrepresentative sample, the average income of 10 students will better approximate the overall average income than will the average income of just 2 students.

The Law of Large Numbers (LLN)

Let x_1, x_2, \dots, x_n be independent observations from *any* population with mean μ . Then, as the sample size n increases

Interpretation:

NOTE: The LLN only tells us that \bar{x} gets closer to μ as the sample size increases. However, in order to understand the variability in \bar{x} from sample to sample and to determine how large n needs to be to ensure we obtain a reasonable estimate of μ , we need to study its sampling distribution.

Mean and Standard Deviation of the Sampling Distribution of \bar{x}

Suppose x_1, x_2, \dots, x_n is a random sample from *any* population with mean μ and standard deviation σ . Then the sampling distribution of \bar{x} has

mean of the sampling distribution of $\bar{x} =$

Interpretation:

If we take a LOT of samples and calculate \bar{x} for each sample, the average of the \bar{x} 's will be close to μ .

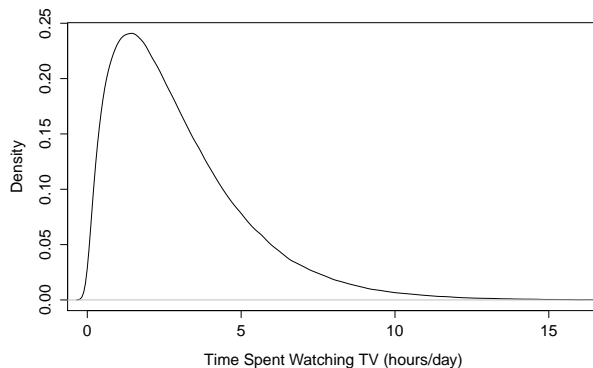
standard deviation of the sampling distribution of \bar{x} , $\sigma_{\bar{x}} =$

Notes:

1. $\mu_{\bar{x}} = \mu \Rightarrow$ **sample averages have the same mean as individual observations**
2. $\sigma_{\bar{x}} = \sigma/\sqrt{n} \Rightarrow$ **sample averages are _____ variable than individual observations.**
3. $\sigma_{\bar{x}} = \sigma/\sqrt{n} \Rightarrow$ **As sample size n increases, the variability of \bar{x} from sample to sample _____.**

EXAMPLE 7.4

Let X = number of hours an American watches TV on an average day and suppose it is known that X has a mean of 3 with a standard deviation of 2.25. The following is a density curve for random variable X :



- (a) Take a sample of 10 Americans. What are the mean and standard deviation of the sampling distribution of \bar{x} , the average number of hours per day spent watching TV for these 10 people?

$$\mu_{\bar{x}} =$$

$$\sigma_{\bar{x}} =$$

- (b) What if we take a sample of 100 people?

$$\mu_{\bar{x}} =$$

$$\sigma_{\bar{x}} =$$

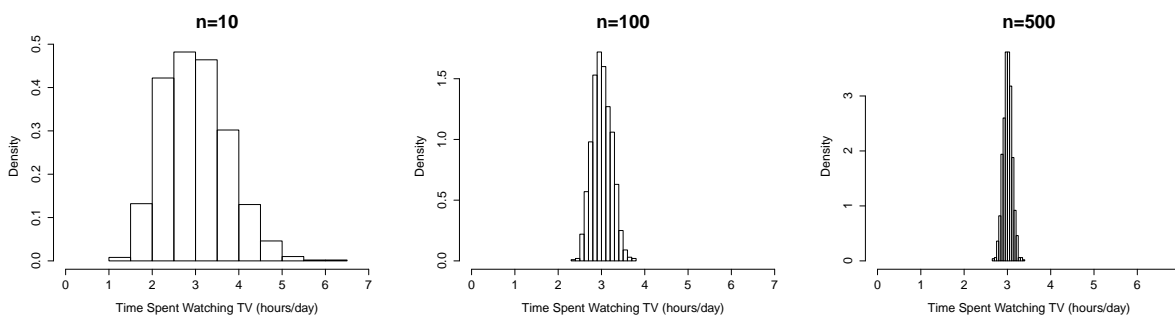
\bar{x} is _____ variable than when we only have $n = 10$.

- (c) How large of a sample would we need in order to obtain an estimate, \bar{x} , with a standard deviation less than or equal to .05?

We have found the mean and standard deviation of the sampling distribution of \bar{x} . However, as when we explored probability distributions, we would also like to know the *shape* of the sampling distribution. The following procedure will allow us to get an idea of this shape:

1. Take 1000 separate samples, each of size n .
2. For each sample, calculate the average time per day spent watching TV.
This gives us 1000 independent sample means.
3. Plot a histogram of the 1000 sample means.
This is an approximation of the density curve for the sampling distribution of \bar{x} based on a sample of size n .

Repeat this procedure for each of $n = 10$, $n = 100$, and $n = 500$:



Observations:

1. **sampling distribution has a different shape than the population distribution**
2. **bell-shaped, symmetric (appear normal)**
3. **centered around the true mean (3)**
4. **become much less variable as sample size increases**

Sampling Distribution of \bar{x}

1. If x_1, x_2, \dots, x_n is a random sample from $N(\mu, \sigma)$, then

2. Central Limit Theorem (CLT)

Suppose x_1, x_2, \dots, x_n is a random sample from *any* population with mean μ and standard deviation σ . Then, if n is large enough (rule of thumb: $n \geq 30$)

EXAMPLE 7.5

A machine fills empty glass bottles with soda. Let X = amount of soda poured into a bottle and assume X is normally distributed with a mean of 298 mL and standard deviation of 3 mL.

1. The label on the bottle says that it contains 295 mL of soda. Find the probability that a randomly selected bottle contains less soda than advertised.

2. Now, find the probability that the average amount of soda per bottle in a randomly chosen 6-pack is less than 295 mL.

Interpretation:

Only about ____ of any sampled 6-packs will have an average amount less than 295 mL/bottle, whereas almost _____ of individual bottles are underfilled.

EXAMPLE 7.6

Suppose the mean number of children per household in the U.S. is 1.90 with a standard deviation of 1.68. (These numbers are based on a 2006 General Social Survey.) Let X = number of children in a household.

1. Does X have a normal distribution?
2. Let \bar{x} = average number of children in a random sample of 50 households. What is the sampling distribution of \bar{x} ?
3. Sample 50 households. Find the probability that the average number of children per household in this sample is more than 5.

7.2 The Sampling Distribution of A Sample Proportion

Notation:

p = population proportion of “success”

\hat{p} = sample proportion of “success”

Statistical Inference: **use \hat{p} to estimate p**

EXAMPLE 7.7: Estimating p

Suppose we take a random sample of 60 U.S. households and 7 say they have been burglarized/robbed. Use this information to estimate p , the true proportion of U.S. households that have been burglarized.

GOAL:

\hat{p} varies from sample to sample.

Therefore, just as we did for \bar{x} , we want to describe the sampling distribution of \hat{p} .

Law of Large Numbers for \hat{p}

The Law of Large Numbers guarantees that as sample size n increases

That is, the larger the sample size, the better \hat{p} estimates p .

Mean and Standard Deviation of the Sampling Distribution of \hat{p}

Suppose \hat{p} is the sample proportion for a sample of size n from a population with proportion p . Then the mean and standard deviation of the sampling distribution of \hat{p} are

mean =

standard error =

Interpretation:

Central Limit Theorem for \hat{p}

Let \hat{p} be the sample proportion based on a sample of size n from a population with probability of success equal to p . Assume the expected number of successes (np) and the expected number of failures ($n(1 - p)$) are both at least 15. Then, by the Central Limit Theorem the sampling distribution of \hat{p} is

EXAMPLE 7.8

Suppose that 29% of University of Minnesota students feel that our campus is becoming more dangerous. Take a sample of 100 university students and let \hat{p} = the proportion of the sample that feel campus is becoming less safe.

- (a) Find the mean and standard deviation of the sampling distribution of \hat{p} .

- (b) What is the approximate sampling distribution of \hat{p} ?
- (c) Select one person at random from the sample of students. What is the probability that this student feels campus is becoming more dangerous?
- (d) What is the probability that fewer than 20% of those surveyed feel campus is becoming more dangerous?

About _____ of possible samples of University students will have sample proportions $< .20$.

RECAP:

The sampling distribution of a sample statistic describes how the value of the statistic varies depending on the sample we get. Specifically, it describes the possible values of the statistic and how likely each of those values is to occur.

CHAPTER 8: CONFIDENCE INTERVALS

How can we use sample data to draw conclusions about the population?

1.
 - (a)
 - (b)
- 2.

8.1 Point Estimation

Examples of Point Estimation: Using \bar{x} to estimate μ and \hat{p} to estimate p .

For any given unknown population parameter there are many (in fact, infinitely many) possible point estimates. For instance, we might estimate an unknown population mean μ using the sample mean \bar{x} or the sample median M (or anything else for that matter). So how do we decide which point estimate to use?

Measuring How Well a Point Estimate Approximates the “Truth”

- 1.

Definition: bias (of a statistic)

The *bias* of a statistic is the difference between the mean of its sampling distribution and the true parameter value. A statistic is *unbiased* if this difference is zero.

Interpretation:

An unbiased estimator doesn't systematically over- or underestimate the parameter. Therefore, we want an unbiased point estimator.

Examples of Unbiased Estimators:

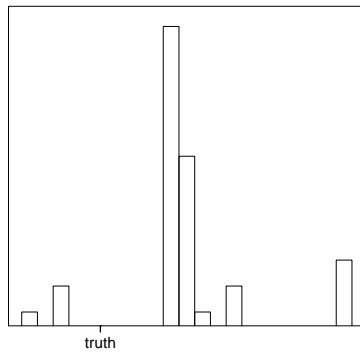
We saw in Chapter 7 that \hat{p} is unbiased for p and \bar{x} is unbiased for μ .

2.

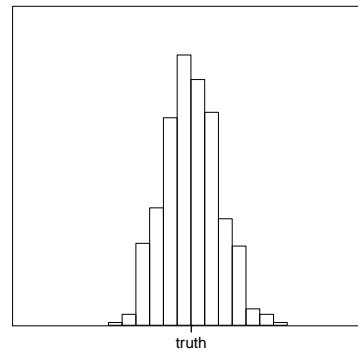
A good estimator will have small standard error compared to other estimators.

EXAMPLE 8.1

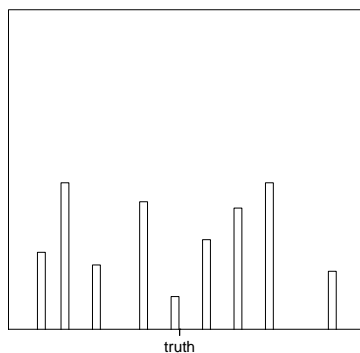
Consider a situation in which we want to choose from four different statistics, each estimating the same parameter. The value of each of these statistics will vary from sample to sample. Below are approximations of the sampling distributions for each statistic. The true value of the parameter is marked on the x -axes. Which statistic do you prefer?



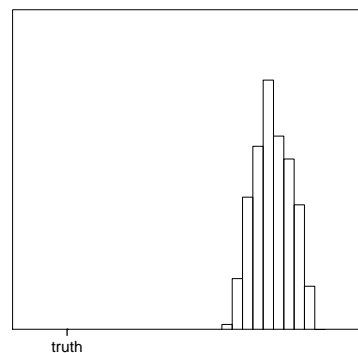
biased, large standard error



unbiased, small standard error



unbiased, large standard error



biased, small standard error

8.2 Interval Estimation

A **point estimate** serves as our best guess for the parameter. It might be too high, too low, or just about right. Therefore, we also need a measure of its reliability. That is, we want to be able to say how *confident* we are in the estimate.

Definition: interval estimate

An *interval estimate* is an interval of numbers within which the parameter value is believed to fall.

EXAMPLE 8.2:

According to the 2012 Resident Satisfaction Survey, 32% of Minneapolis residents felt that public safety is among the biggest challenges of the city in the next five years. Considering the reported margin of error of plus or minus 3 percentage points, find an interval estimate for p , the true proportion of Minneapolis residents who felt that public safety is among the biggest challenges of the city in the next five years.

$$\hat{p} \pm \text{moe} = .32 \pm .03 = (.29, .35).$$

We are fairly confident that the true proportion of Minneapolis residents who felt that public safety is among the biggest challenges of the city in the next five years is between 29% and 35%.

Definition: margin of error

The *margin of error* measures how accurate a point estimate is likely to be in estimating a parameter.

Definition: confidence interval (CI)

A *confidence interval* is an interval containing the most believable values for a parameter. The probability that this method produces an interval that contains the parameter is called the *confidence level*.

General form of a confidence interval:

NOTE: **The larger the moe, the _____ the interval and the _____ accurate our estimate is.**

8.2.1 Confidence Intervals for A Population Proportion, p

EXAMPLE 8.3

We want to estimate the number of contracts in a city that are awarded to minority-owned firms. We take a random sample of 389 contracts from the city and find that 58 were awarded to minority-owned firms. Our goal is to use the sample information to construct a confidence interval with a confidence level of .95 for p , the true proportion of contracts that are assigned to minority-owned firms.

- (a) What is the sampling distribution of \hat{p} , the sample proportion of contracts that went to minority-owned firms?

Since n is large enough, we have

- (b) Between what 2 values does the middle 95% of this distribution fall?

Interpretation:

In 95% of all possible samples, \hat{p} will be within _____ of p . The other 5% will be “unlucky” and \hat{p} will be further than _____ from p .

The problem:

The solution:

Definition: Standard Error

A *standard error* is an estimated standard deviation of a sampling distribution.

NOTE:

Standard errors are commonly used in practice, because TRUE standard deviations are usually unknown.

For example:

For finding a confidence interval for a population proportion p , the standard error (se) is $\sqrt{\hat{p}(1 - \hat{p})/n}$

- (c) Using the estimated standard deviation (i.e., the standard error) of \hat{p} , construct a 95% confidence interval for p .

Part (b) \Rightarrow if we take _____ from any sample and go _____ to the left or right of it, there is a 95% chance that it will include p . That is, there is a 95% chance that p is in the interval:

BUT we do not know p . To calculate the interval, estimate p by \hat{p} .

$\hat{p} = 58/389 = .149$. Therefore, a 95% CI for p is

Interpret the CI:

On average, approximately 95% of the samples drawn from the population would produce CIs that contain the true population proportion p

Suppose an equal opportunity law requires that at least 20% of all city contracts go to minority-owned firms. What information does the CI give us in this context?

We are 95% confident that the true proportion of contracts that go to minority-owned firms is only between _____ and _____. It is unlikely that the true proportion of contracts that go to minority-owned firms is at least 20%.

NOTE: We can calculate confidence intervals for any confidence level, not just .95.

Notation:

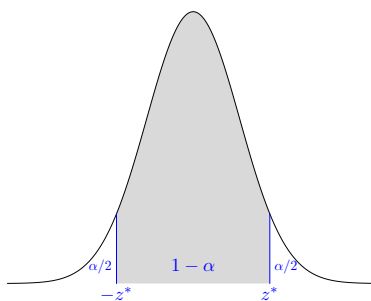
α = error probability of the confidence interval ($0 < \alpha < 1$)

$1 - \alpha$ = confidence level of the confidence interval

Large Sample Confidence Interval for p

Assume we take a random sample of size n and that the number of successes and the number of failures in that sample are both at least 15. That is, $n\hat{p}$ and $n(1 - \hat{p})$ are both at least 15. Then a large sample CI for p with confidence level $1 - \alpha$ is

where z^* depends on the confidence level. Specifically, $\pm z^*$ mark the *middle* $1 - \alpha$ proportion of the $N(0, 1)$ distribution:



Finding z^*

Though we can find z^* for any specified confidence level, 90%, 95%, and 99% confidence intervals are the most common.

Confidence Level	α	$\alpha/2$	z^*
90%			
95%			
99%			

Interpreting Confidence Intervals

1. For example: *In the long run, 95% of 95% CI's will contain the true parameter. The other 5% that do not, are based on the "unlucky" samples that produce unusually high or low values of the statistic.*
2. We are 95% confident that the true value of the parameter is between the upper and lower bounds of the 95% CI.
3. I'm 95% confident the interval will include the true value of the population parameter.

EXAMPLE 8.4 (Exercise 8.13)

When the 2010 General Social Survey asked subjects if they would be willing to accept cuts in their standard of living to protect the environment, 486 of 1374 subjects said yes.

- (a) Estimate the population proportion, p , who would answer yes.

$$\hat{p} = 486/1374 = .354$$

- (b) Can we use this sample to calculate a valid confidence interval?

Yes, we have more than 15 successes and 15 failures.

- (c) Calculate a 90% CI for p .

$$\begin{aligned} \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= .354 \pm 1.645 \sqrt{\frac{.354(1-.354)}{1374}} \\ &= .354 \pm .021 \\ &= (.333, .375) \end{aligned}$$

Interpretation: **We are 90% confident that the true proportion of people willing to lower their standard of living is between .333 and .375.**

- (d) Now calculate a 98% CI for p .

$$\begin{aligned} \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= .354 \pm 2.33 \sqrt{\frac{.354(1-.354)}{1374}} \\ &= .354 \pm .030 \\ &= (.324, .384) \end{aligned}$$

Interpretation: **We are 98% confident that the true value of p is between .324 and .384.**

- (e) Similarly, a 99% CI for p can be shown to equal (.321, .387). What pattern do you see?

As the confidence level increases, the CI becomes wider.

EXAMPLE 8.5

A woman who smokes during pregnancy increases health risks to her infant. Let p = proportion of women smokers that quit during pregnancy.

- (a) Suppose that in a random sample of 300 women who smoked prior to pregnancy, 51 quit smoking during pregnancy. Use this sample information to calculate a 98% CI for p in R. (Note: the assumptions of a random sample and at least 15 observed successes and 15 observed failures are met.)

```
#'x'=number of successes, 'n'=sample size
> prop.test(x=51, n=300, conf.level=0.98, alternative="two.sided")
      1-sample proportions test with continuity correction
data:  51 out of 300, null probability 0.5
X-squared = 129.3633, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
98 percent confidence interval:
 0.1240578 0.2280177
sample estimates:
      p
0.17
```

- (b) Now, suppose that we have a random sample of 1000 women smokers in which 170 quit smoking. Use this sample information to calculate a 98% CI for p in R. (Again, the assumptions for the CI are satisfied.)

```
> prop.test(x=170, n=1000, conf.level=0.98, alternative="two.sided")
      1-sample proportions test with continuity correction
data:  170 out of 1000, null probability 0.5
X-squared = 434.281, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
98 percent confidence interval:
 0.1436947 0.1999219
sample estimates:
      p
0.17
```

- (c) What pattern do you see?

Properties of a Confidence Interval

1. **moe _____ as confidence level increases**
2. **for fixed \hat{p} , moe _____ as sample size increases**

Small Sample Confidence Interval for p

When n is small the normal approximation to the sampling distribution of \hat{p} can be awful, thus causing the large sample confidence interval for p to be misleading. In fact, there exist alternative methods for constructing confidence intervals when the sample size is not large enough to satisfy the requirement of at least 15 successes and 15 failures. See p. 386 of the book for details.

Sample Size Calculations

Goal: Ideally, we would like **both**

1. _____ **level of confidence in our interval estimate; and**
2. _____ **margin of error**

If the sample size is large enough, we can achieve *both* of these goals simultaneously. If this is the case, then why don't we always just pick really large samples?

EXAMPLE 8.2 CONTINUED

Recall that in the 2012 Resident Satisfaction Survey, the City of Minneapolis found out that 32% of the residents felt that public safety is among the biggest challenges of the city in the next five years. However, suppose we want to obtain a more recent estimate of p , the true proportion of Minneapolis residents who felt that public safety is among the biggest challenges of the city in the next five years. How many people do we need to poll in order to estimate p within 3 percentage points with 95% confidence?

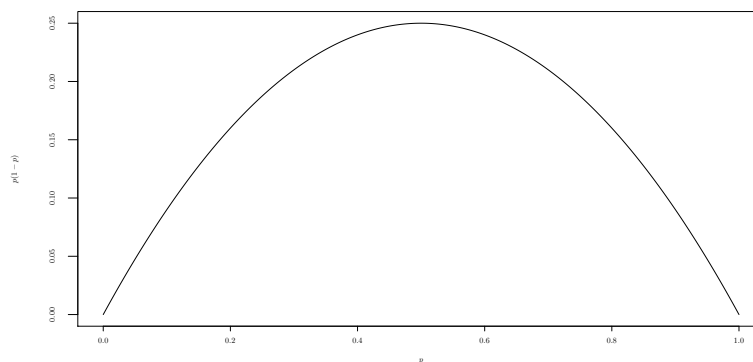
Problem:

Solution:

Back to the problem...

What if we had no prior guess for p ?

Use the worst case scenario. moe is maximized when $\sqrt{\hat{p}(1 - \hat{p})/n}$ is maximized (i.e. when $\hat{p}(1 - \hat{p})$ is max'd):



This happens at $\hat{p} = 1/2$.

Therefore, using $\hat{p} = 1/2$ will give a larger value for n than any other \hat{p} . However, this will sometimes lead us to collect a lot more data than is necessary if p is either close to 0 or close to 1.

Sample Size for a Desired Margin of Error

The sample size required to estimate p within a margin of error of (at most) m with a confidence level of $1 - \alpha$ is

where $z_{\alpha/2}$ depends on the desired confidence level $(1 - \alpha)$, and

1.

or

2.

NOTES:

1. Always round up for sample size calculations!
2. Notice that the sample size calculation is not influenced by the size of the population.

For instance, suppose we want to estimate the proportion of *all* Americans (not just Republicans) that prefer Romney. To estimate *this* proportion within 2.5 percentage points with 99% confidence, **we would only need to survey the same number of people** even though the target population of all Americans is much larger than the population of Republicans.

EXAMPLE 8.6

A campus group is interested in estimating the proportion of U students that feel that their binge drinking has negatively impacted their academic performance. How many students should they poll in order to estimate the proportion to within ± 0.05 with 90% confidence if...

- (a) we have no prior information?
- (b) based on UMADD statistics, we expect the proportion to be around 0.25?

8.2.2 Confidence Intervals for A Population Mean, μ

Constructing a confidence interval for a population mean is much like constructing a confidence interval for a population proportion. Mainly, the general form of the confidence interval will still be

Just as we used knowledge of the sampling distribution of \hat{p} to construct confidence intervals for p , we use the sampling distribution of \bar{x} to construct confidence intervals for μ . Recall that if x_1, x_2, \dots, x_n are a random sample from a population with mean μ and standard deviation σ

so that if we standardize \bar{x} , we have

The Problem:

The Solution:

NOTE: t does not have a normal distribution! In fact, in replacing σ by an estimate (s), t takes on an extra source of error that the z -score does not. Therefore, its distribution must account for this increased variability.

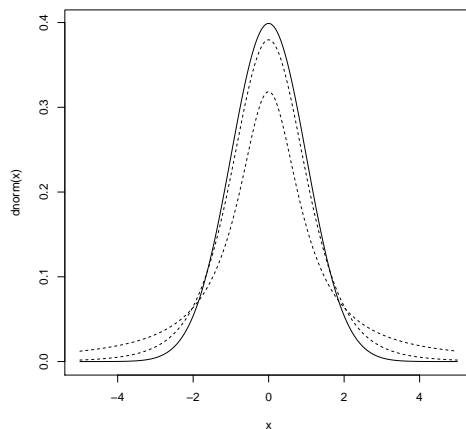
The t -distribution

If x_1, x_2, \dots, x_n are a random sample from a normal population with unknown mean μ and unknown standard deviation σ , then

where t_{n-1} denotes the t -distribution with $n - 1$ degrees of freedom (df).

NOTE: This is also *approximately* true when x_1, x_2, \dots, x_n is a random sample from *any* distribution with mean μ and standard deviation σ if n is large!

Properties of the t -distribution



- **Properties:**
 - bell-shaped
 - symmetric about 0
 - has fatter tails than $N(0, 1)$ (that is, it is more spread out)
- The exact shape and spread of the t distribution are determined by its degrees of freedom.
- Relationship to $N(0, 1)$:
As df increase, the t -distribution gets closer to $N(0, 1)$.

The idea: the estimate of σ improves as n (and therefore df) increases.

t-distribution Calculations

Table B in Appendix A gives right tail probabilities for *t*-distributions with varying df.

EXAMPLE 8.7:

1. If $df = 11$, find the value of t for which $P(t_{11} \geq t) = .005$.

```
> qt(.005, df=11, lower=F) #‘qt’ = quantile of the t distribution with
                           #‘df’ = degrees of freedom
[1] 3.105807                #R calculates lower tail probabilities by default
                           #Tell it not to with ‘lower=F’
```

2. If $df = 21$, find the value of t for which $P(t_{21} \leq t) = .99$.

```
> qt(.99, df=21)
[1] 2.517648
```

3. If $df = 14$, what is $P(t_{14} \geq 2)$?

```
> pt(2, df=14, lower=F)
[1] 0.03264398
```

$$P(t_{14} \geq 2.145) < P(t_{14} \geq 2) < P(t_{14} \geq 1.761) \Rightarrow .025 < P(t_{14} \geq 2) < .05$$

Have to use R to find the exact calculation.

4. Rule:

If $100 \leq df \leq 1000$, use $df = 100$ in Table B.

If $df \geq 1000$, use $df = \infty$ in Table B.

(NOTE: t_∞ is “approximately” the same as $N(0, 1)$!)

Example: $P(t_{5000} \geq 1.96) \approx$

Confidence Interval for μ

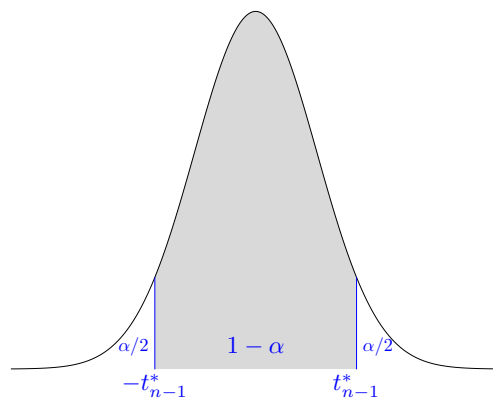
Assumptions:

1. randomness: x_1, x_2, \dots, x_n are a random sample from some population with unknown mean μ and unknown standard deviation σ

2. normality: the population distribution is approximately normal

If the assumptions are satisfied, the confidence interval with confidence level $1 - \alpha$ for μ is

where t_{n-1}^* depends on the confidence level. Specifically, $\pm t_{n-1}^*$ mark the *middle* $1 - \alpha$ proportion of the t_{n-1} distribution:



Properties of the confidence interval for μ :

1. **moe** _____ as the confidence level increases. (same as for p)
2. **moe** _____ as sample size increases. (same as for p)
3. **the smaller the standard deviation, the** _____ **the moe**
(There is more precision in the interval estimate if there's less variation among the subjects.)

EXAMPLE 8.8

A survey of 51 current adult smokers in the U.S. asked, “On average, how many cigarettes do you smoke per day?”

- (a) Based on the following stem-and-leaf plot of the raw data, does it appear that the underlying assumptions of the confidence interval are satisfied?

```
> cigs <- c(1,3,3,8,9,9,9,10,11,11,12,13,14,14,15,15,15,16,16,16,16,17,17,17,17,
  18,19,19,20,20,20,20,20,22,22,23,23,24,25,25,25,28,30,30,30,30,32,32,35,38,40)
> stem(cigs)
0 | 133
0 | 8999
1 | 0112344
1 | 55566667777899
2 | 0000022334
2 | 5558
3 | 000022
3 | 58
4 | 0
```

- (b) Calculate and interpret a 99% confidence interval for μ , the true mean number of cigarettes smoked per day by U.S. smokers.

```
> mean(cigs)
[1] 19.09804
> sd(cigs)
[1] 8.775545
> length(cigs)
[1] 51
```

Interpretation:

We are 99% confident that the average number of cigarettes that a smoker has each day is between _____ and _____.

(c) Repeat this analysis using the `t.test` function in R.

```
> t.test(x=cigs, conf.level=.99, alternative="two.sided")
t = 15.5417, df = 50, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 15.80751 22.38857
sample estimates:
mean of x
 19.09804
```

Robustness of t -Procedures

When we use the t distribution to construct a confidence interval for μ , we assume that the sample is *randomly* drawn from some population that is *approximately normal*. However, in practice these assumptions are rarely perfectly satisfied. How does this affect the reliability of interval estimates?

Definition: robustness

A statistical method is *robust* with respect to a particular assumption if it performs adequately even when that assumption is violated.

Notes: The t procedure for calculating a confidence interval...

1. ***is robust to non-normality (when no outliers are present)***
2. ***does not perform well when there are extreme outliers***
3. ***is not robust to violations of the assumption of a random sample***

Sample Size Calculations

The Goal: Determine how large of a sample size is needed to obtain *both* a small margin of error *and* high confidence in our interval estimate (so our method almost always captures μ).

For instance, how large of a sample do we need if we want a confidence interval for μ with a $1 - \alpha$ confidence level and a margin of error that is no more than m ?

t^* depends on the confidence level and the sample size.

The problem:

The solution:

-
-

Sample Size for a Desired Margin of Error

The sample size required to estimate μ within a margin of error of m with a confidence level of $1 - \alpha$ is approximately

Reminder: Always round up for sample size calculations!

EXAMPLE 8.10

Suppose the Minnesota wildlife service wishes to estimate the mean number of days spent hunting, per hunter, for all licensed hunters in Minnesota. How many hunters must they survey in order to be 95% confident that their estimate is within 1 day of the true mean? (Use the fact that for a previous study conducted in 2000 the sample standard deviation was $s = 10$.)

CHAPTER 9: HYPOTHESIS TESTS

Motivating Example

A diet pill company advertises that at least 75% of its customers lose 10 pounds or more within 2 weeks. You suspect the company of falsely advertising the benefits of taking their pills. Suppose you take a sample of 100 product users and find that only 5% have lost at least 10 pounds. Is this enough to prove your claim? What about if 72% had lost at least 10 pounds?

Goal:

9.1 Elements of A Hypothesis Test

1. Assumptions

The reliability of any hypothesis test relies on a certain set of assumptions being satisfied.

2. Hypotheses

Each hypothesis test has two hypotheses about the *population*:

Null Hypothesis (H_0):

a statement about the population we want to *disprove*. It often represents *no effect*

Alternative Hypothesis (H_a):

what we hope to find evidence for. It is an *alternative* to the null hypothesis and should be stated *before* looking at the data (to avoid bias)!

Diet Pill Example:

Let p = true proportion of diet pill customers that lose at least 10 pounds. State the null and alternative hypotheses for the diet pill example.

H_0 :

H_a :

3. Test Statistic

Definition: Test Statistic

A test statistic is a *measure* of how compatible the data is with the null hypothesis. The larger the test statistic, the less compatible the data is with the null hypothesis.

Most test statistics we will see have the following form:

What does an extreme value of T reflect?

T is extreme

⇒ sample estimate is many standard deviations away from the hypothesized value

⇒ the data don't agree with the hypothesized value.

NOTE:

T is a function of the sample data

⇒ T is a statistic

⇒ T has a sampling distribution under H_0 (assuming H_0 is true)

Studying the sampling distribution of T helps us measure how likely our sample data is when H_0 is true

4. p -value

The p -value helps us to interpret the test statistic.

Definition: p -value

Assume H_0 is true. Then the p -value is the probability that the test statistic T takes a value (in support of H_a) as or more extreme than the one we observed.

Diet Pill Example:

Suppose that in your sample of 100 customers, 65% had lost at least 10 pounds in 2 weeks. Recall our hypotheses:

$$H_0 : p = 0.75$$

$$H_a : p < 0.75$$

- (a) Assuming H_0 is true, what is the sampling distribution of \hat{p} ?
- (b) Use part (a) to determine the sampling distribution of test statistic T .
- (c) Calculate a test statistic for this hypothesis test.
- (d) Calculate and interpret the p -value for this test statistic.

p -value =



Interpretation: If the proportion of customers that lose at least 10 pounds is truly .75, it is very unlikely that we would choose a sample with so few people achieving this weight loss (prob=). Thus, our data provides evidence that H_0 is false.

In General:

5. Conclusion and Interpretation

We assume that H_0 is true and put the *burden of proof* on H_a . Therefore, we can use the p -value to decide whether or not the data provide sufficient evidence to reject H_0 in favor of H_a .

The Idea:

Recall that a small p -value indicates that our observation based on the sample is unlikely to occur under H_0 .
How small is small enough to reject H_0 ?

Rejection Rule:

Determine “statistical significance” by comparing a p -value to a *significance level*, α .

- $p\text{-value} < \alpha \Rightarrow$ we have strong evidence against H_0 . We conclude that the results are “statistically significant” at level α and reject H_0 .
- $p\text{-value} \geq \alpha \Rightarrow$ we fail to reject H_0 in favor of H_a . We do not have enough evidence to conclude H_0 is false.

Interpreting the Significance Level

If, for example, we choose $\alpha = 0.05$, we require strong enough evidence *against* H_0 that when H_0 is true there is only a 5% chance that we mistakenly reject it.

Common Choices for α :

Our choice of α depends on how *confident* we want to be in our decision about H_0 . It should always be chosen *prior* to collecting the data!

$\alpha = 0.01 \Rightarrow$ need to be _____ confident that H_0 is false in order to **reject**

$\alpha = 0.05 \Rightarrow$ need to be _____ confident that H_0 is false in order to **reject**

$\alpha = 0.10 \Rightarrow$ need to be _____ confident that H_0 is false in order to **reject**

9.2 Normal Hypothesis Test for Population Proportion p

Assumptions

- 1.
- 2.

Rule of thumb: at least 15 expected successes ($np_0 \geq 15$) and 15 expected failures ($n(1 - p_0) \geq 15$)

Hypotheses

1. $H_0: p = p_0$
 $H_a:$

2. $H_0: p = p_0$
 $H_a:$

3. $H_0: p = p_0$
 $H_a:$

- The value of p_0 is the same in both H_0 and H_a .

- Hypotheses 1 and 2 are called “**one-sided tests**” and hypothesis 3 is called a “**two-sided test**”.

Test Statistic

p-value

$p\text{-value} = P(\text{observe a value as or more extreme than } z^* \text{ in favor of } H_a \mid H_0 \text{ is true})$

Hypothesis 1:

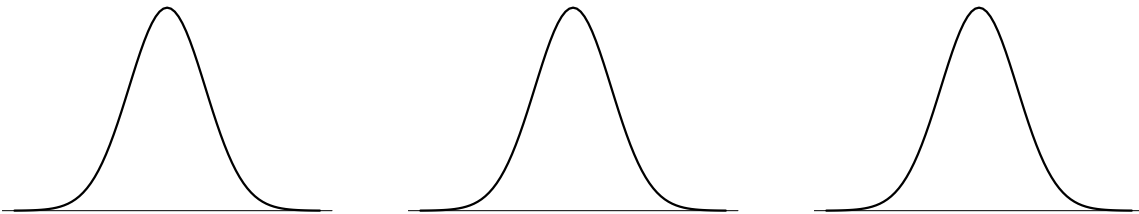
$p\text{-val} =$

Hypothesis 2:

$p\text{-val} =$

Hypothesis 3:

$p\text{-val} =$

**Conclusion**

$p\text{-value} < \alpha \Rightarrow$

$p\text{-value} \geq \alpha \Rightarrow$

EXAMPLE 9.1

The manufacturer of a new breast cancer screening method claims that it detects cancer in more than 83% of women who have it. To investigate, they apply their screening method to a sample of 203 randomly selected women known to have breast cancer. The test detects cancer in 184 of these women. Use this information to test their claim at the 0.01 level.

Assumptions:

Hypothesis: Let $p =$ true proportion of detection

$H_0:$

$H_a:$

Test statistic:

p -value:

p -val =

There's only a _____% chance that our test performs so well in a sample of this size if it doesn't actually detect as well as claimed

Conclusion:

Interpretation: We have _____ to conclude that the true proportion of breast cancers detected by the new screening method is significantly greater than 83%.

EXAMPLE 9.2 (Exercise 9.19)

Among the employees eligible for management training at a large supermarket chain in Florida, 40% are women. However, since management training began, only 12 of the 40 employees (30%) chosen for the training were women. At the 0.05 level, test the claim of a women's group that women are being passed over for management training in favor of their male colleagues. That is, test the claim that a disproportionate number of the people selected for the training are men.

Let p = proportion of employees selected for training that are women

Hypothesis:

H_0 :

H_a :

Note: The assumptions hold since we have a random sample and there are $40(.40) = 16$ expected 'successes' and $40(1 - .40) = 24$ expected failures under H_0 .

```
> prop.test(x=12, n=40, conf.level=0.95, p=0.40, alternative="less")
1-sample proportions test with continuity correction
data: 12 out of 40, null probability 0.4
X-squared = 1.276, df = 1, p-value = 0.1293
alternative hypothesis: true p is less than 0.4
95 percent confidence interval:
 0.0000000 0.4416485
sample estimates:
 p
0.3
```

NOTES:

1. If $H_a: p > 0.40$, then we would type alternative = "greater".
If $H_a: p \neq 0.40$, then we would type alternative = "two.sided".
2. We would get slightly different values if we did this by hand. R uses a 'continuity correction' that we do not...

Interpret p -value:

Under H_0 , there is a nearly _____% probability to observe a sample proportion of .3 or lower. This is not _____, so it is _____ that this sample is drawn from a population with proportion .4.

Conclusion:

We _____ H_0 at the significance level .05.

Interpretation: We _____ evidence to conclude that the true proportion of employees selected for training that are women is significantly less than 40%.

9.3 The t -Test: Hypothesis Testing for Population Mean μ

The basic structure of hypothesis tests regarding population means is the same as for hypothesis tests regarding population proportions, only the details change.

Assumptions

1. Random sample
2. Normality and/or Large n : Population distribution is approximately normal and/or sample size is large

Hypotheses

- | | | |
|---------------------------------|---------------------------------|---------------------------------|
| 1. $H_0: \mu = \mu_0$
$H_a:$ | 2. $H_0: \mu = \mu_0$
$H_a:$ | 3. $H_0: \mu = \mu_0$
$H_a:$ |
|---------------------------------|---------------------------------|---------------------------------|

Test Statistic

p -value

p -value = $P(\text{observe a value as or more extreme than } t \mid H_0 \text{ is true})$

Hypothesis 1:

p -val =

Hypothesis 2:

p -val =

Hypothesis 3:

p -val =

Conclusion

p -value $< \alpha \Rightarrow$ reject H_0

p -value $\geq \alpha \Rightarrow$ fail to reject H_0

EXAMPLE 9.3 (Exercise 9.31)

In the 2004 General Social Survey, men were asked how many hours they worked in the previous week. For the random sample of 895 male workers, the mean was 45.3 hours with a standard deviation of 14.8 hours. Does this data support the claim that the true average number of hours that men work each week (μ) exceeds the standard 40 hour work week? (Test this claim at the 0.05 level.)

Assumption:

Hypothesis:

H_0 :

H_a :

Test Statistic:

p -value:

Interpretation of the p -value:

Conclusion:**Interpretation:**Hypothesis Tests and Confidence Intervals**EXAMPLE 9.4**

It is important for nutritional information on food packaging to be accurate. A random sample of $n = 10$ frozen dinners of a certain brand was selected from a production line. The mean calorie content for these dinners was 246.6 with a standard deviation of 10.803 calories. The data is approximately normal with no substantial outliers.

- (a) The packaging for this dinner lists the calorie content as 240 calories. Let $\mu =$ true mean calorie content of the frozen dinner and test whether or not the information on the package is accurate at the 0.05 level.

Assumption:

Random sample and though n is small, the population from which the data are sampled is approximately normal according to the problem description.

Hypothesis:

$$H_0: \mu = 240$$

$$H_a: \mu \neq 240$$

Test Statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{246.6 - 240}{10.803/\sqrt{10}} = 1.93$$

p-value:

p-val = $2P(t_9 > 1.93)$, where $P(t_9 > 1.93) > P(t_9 > 2.262) = 0.025$.

Therefore, $2P(t_9 > 1.93) > 0.05$

Conclusion:

p-val > .05 \Rightarrow Fail to reject H_0 at the .05 level. There is not enough evidence to conclude that true mean content is significantly different from what the label says.

- (b) Calculate and interpret a 95% confidence interval for μ .

$$\begin{aligned}\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} &= 246.6 \pm t_{.025, 9} \frac{10.803}{\sqrt{10}} \\ &= 246.6 \pm 2.262 \frac{10.803}{\sqrt{10}} \\ &= 246.6 \pm 7.73 \\ &= (238.87, 254.33)\end{aligned}$$

We are 95% confident that the true mean calorie content is between 238.87 and 254.33 calories.

- (c) Repeat these analyses in R.

```
> cal <- c(240, 253, 243, 267, 258, 239, 235, 252, 246, 233)
> t.test(x=cal, conf.level=0.95, alternative="two.sided", mu=240)
      One Sample t-test
data:  cal
t = 1.9319, df = 9, p-value = 0.08541
alternative hypothesis: true mean is not equal to 240
95 percent confidence interval:
 238.8718 254.3282
sample estimates:
mean of x
 246.6
```

Notes about the t.test function:

- 'mu' = hypothesized value of μ
- If $H_a: \mu < 240$, then we would type alternative = "less".
- If $H_a: \mu > 240$, then we would type alternative = "greater".

- Always need alternative = “two.sided” to get the correct CI.

- (d) What conclusions can you make about the package’s claim based on the hypothesis test and the confidence interval?

Notice that 240 calories is in the 95% CI for μ . Therefore, we can use both the hypothesis test and CI to conclude that 240 calories is a plausible value of μ .

Equivalence Between a CI and 2-sided Hypothesis Test for μ

Inference about μ has an exact equivalence between the *two-sided* hypothesis test at level α and the confidence interval with confidence level $1 - \alpha$. Specifically, suppose we wish to test $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$ at level α . Then,

1. If μ_0 is in the $1 - \alpha$ confidence level CI for $\mu \Rightarrow$

2. If μ_0 is *not* in the $1 - \alpha$ confidence level CI for $\mu \Rightarrow$

For example, we can use a 95% CI for μ to make a conclusion for the two-sided test at the 0.05 level and can use a 99% CI for μ to make a conclusion for the two-sided test at the 0.01 level (and so on).

NOTES:

1. Confidence intervals and *one-sided* tests for μ are *not* compatible!

2. Inference about *proportions* does *not* have an exact equivalence between the confidence interval and 2-sided hypothesis test.

9.4 Possible Errors in Hypothesis Testing

Inference based on a hypothesis test may not always reflect the “truth”!

	H_0 true	H_a true
Do not reject H_0		
Reject H_0		

Note: “Error” doesn’t mean we did anything wrong - it just means that the conclusion based on our data does not reflect the true state of nature.

EXAMPLE 9.5

According to the Journal of Psychology and Aging, older workers have an average job satisfaction rating of 4.3 (on a scale from 0 to 5). We are interested in knowing if the average satisfaction rating is lower among younger workers. That is, we want to test

$$H_0: \mu = 4.3$$

$$H_a: \mu < 4.3$$

where μ = the mean job satisfaction rate for younger workers. What are the Type I and Type II errors in the context of this problem?

Type I:

Type II:

How likely are we to commit a Type I or Type II error?

Probability of a Type I Error

Recall the interpretation of the significance level, α , of a hypothesis test:

If we set significance level = α , we require strong enough evidence *against* H_0 that the probability of mistakenly rejecting H_0 when it is actually true is only α . Therefore...

P(Type I Error) =

Controlling the Probability of a Type I Error

Our choice of α controls the chances of making a Type I Error. How should we choose α ?

Probability of a Type II Error

Calculating the probability of a Type II Error can be complex. It is also not as easy to control as the probability of a Type I error. In general,

P(Type II error) _____ as P(Type I error) increases.

Why?

Insight: As P(Type I error) increases, we require _____ evidence for rejecting H_0 . This means that we are _____ likely to make a Type II error.

9.5 Limitations and Common Misinterpretations of Hypothesis Testing

1. Statistical significance does *not* mean practical significance. Statistical significance relates to the *existence* of an effect (or difference), whereas practical significance relates to the *size* of a possible effect (or difference).

Example: Let μ = true average IQ of children in a certain region of the U.S.. Based on a sample of 5000 children with an average IQ of 100.8 and standard deviation of 16.21, the p -value for the test of $H_0 : \mu = 100$ versus $H_a : \mu > 100$ is approximately .0002. Therefore, though there is not much of a *practical* difference between the sample mean (100.8) and the hypothesized mean (100), this difference *is* highly statistically significant. That is, there is

statistical significance but not practical significance. Why did this happen?

When n is large, standard error will be small. Therefore, the test statistic will be large and the p -value small. However, the p -value measures extent of evidence against H_0 , not how far the true parameter is from H_0 .

2.

Example: p -values of .047 and .052 have almost no practical difference, but they will result in different conclusions if $\alpha = .05$. Therefore, always report the exact p -value when it is close to α .

3.

When we ‘fail to reject H_0 ’, we are saying that we do not have enough evidence to conclude that H_0 is false. We are *not* saying that H_0 is true.

4. $p\text{-value} = P(\text{test statistic is more extreme than the one we observed} \mid H_0 \text{ is true})$

$p\text{-value} \neq P(H_0 \text{ is true} \mid \text{observed test statistic})$

5. It is misleading to report results only if they are statistically significant.

Suppose we run 20 similar tests, of which only one is statistically significant. If we report the one significant test, it is quite likely that we are reporting a Type I error.

6. Don’t always believe what you read! Since researchers *do* often only report results that are significant and beneficial to their cause, be sure to keep in mind that these results may actually be the result of a Type I error!

CHAPTER 10: COMPARING TWO GROUPS

Instead of focusing on one population mean or proportion, we might be interested in comparing parameters from 2 distinct populations.

Examples:

- compare average GRE scores before and after taking a prep course
- compare the proportion of Minnesotans who own guns to the proportion of Wisconsinites who own guns

10.1 Comparing Two Proportions

Goal: Compare two (unknown) proportions corresponding to two independent samples.

Procedure:

Take two *independent* random samples of size n_1 and n_2 from the two populations of interest and count up the number of “successes” in each. We use the following notation to describe the situation:

Population	Population Proportion	Sample Size	# of Successes	Sample Proportion
1	p_1	n_1	X_1	$\hat{p}_1 = X_1/n_1$
2	p_2	n_2	X_2	$\hat{p}_2 = X_2/n_2$

How can we compare p_1 and p_2 ?

Just as with inference regarding a single mean, μ , or proportion, p , we can make inferences about $p_1 - p_2$ using

- 1.
- 2.

10.1.1 Point Estimation for $p_1 - p_2$

The obvious point estimate of $p_1 - p_2$ is _____.

Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

The statistic $\hat{p}_1 - \hat{p}_2$ varies from sample to sample. Just as the sampling distributions for \bar{x} and \hat{p} were used to make inferences about μ and p , respectively, the sampling distribution of this statistic will be used to construct confidence intervals and hypothesis tests regarding $p_1 - p_2$. This sampling distribution has

In addition, when n_1 and n_2 are both “large enough”, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal:

How large is “large enough”?

10.1.2 Confidence Intervals for $p_1 - p_2$

Recall the basic form of a confidence interval:

Large Sample Confidence Interval for $p_1 - p_2$

Assumptions:

- 1.
- 2.

Then the large sample $1 - \alpha$ confidence level CI for $p_1 - p_2$ is

where, as usual, $\pm z^*$ mark the middle $1 - \alpha$ proportion of the $N(0, 1)$ distribution.

Interpreting the Confidence Interval for $p_1 - p_2$

- 1.
- 2.

EXAMPLE 10.1

The NCAA requires colleges to report the graduation rates of their athletes. In a sample of former male and female student athletes, 43 of the 53 females surveyed had graduated whereas 58 of the 102 males had graduated. Calculate and interpret a 99% confidence interval for the difference in true proportions of female athletes who graduate (p_F) and male athletes who graduate (p_M).

Assumptions:

99% CI for $p_F - p_M$:

We are 99% confident that the true proportion of female student athletes that graduate is between _____ and _____ higher than the true proportion of males student athletes that graduate.

10.1.3 Hypothesis Tests for Comparing p_1 and p_2

Hypothesis tests for comparing two population proportions consist of the same five elements as were introduced in Chapter 9.

Assumptions

1. 2 independent random samples
2. at least 5 successes and 5 failures in both samples

Hypotheses

The null hypothesis is always $H_0: p_1 = p_2$, that is, that there is no difference between the 2 proportions.

- | | | |
|-------------------------------|-------------------------------|-------------------------------|
| 1. $H_0: p_1 = p_2$
$H_a:$ | 2. $H_0: p_1 = p_2$
$H_a:$ | 3. $H_0: p_1 = p_2$
$H_a:$ |
|-------------------------------|-------------------------------|-------------------------------|

Test Statistic

WHY?

Recall that the basic form of a test statistic is

Point Estimate and Hypothesized Value:

We can rewrite $H_0: p_1 = p_2$ as $H_0: p_1 - p_2 = 0$ where $\hat{p}_1 - \hat{p}_2$ is a point estimate for $p_1 - p_2$.

Standard Error of $\hat{p}_1 - \hat{p}_2$:

When H_0 is true, $p_1 = p_2 = p$ for some (unknown) value of p so that

Since p is unknown, we estimate it using _____.

Note: \hat{p} is called the *pooled* estimate of p since it is obtained by pooling the information from both samples to provide one estimate of p .

Finally, we can substitute \hat{p} for p and obtain the above test statistic.

p-value

Hypothesis 1: $p\text{-val} = P(Z < z)$

Hypothesis 2: $p\text{-val} = P(Z > z)$

Hypothesis 3: $p\text{-val} = 2P(Z > |z|)$

Conclusion

If $p\text{-value} < \alpha$, reject H_0

If $p\text{-value} \geq \alpha$, fail to reject H_0

EXAMPLE 10.2

SurveyUSA polled 500 Americans and asked if marijuana should be legalized for medicinal purposes. The results of this survey are summarized in the following contingency table:

	Legalize	Don't legalize	Total
< 50 years old	202	108	310
≥ 50 years old	118	72	190
Total	320	180	500

- (a) Use this information to test at the 0.05 level whether younger people are more likely than older people to think marijuana should be legalized for medicinal purposes.

Assumption:

Hypothesis:

$H_0 :$

$H_a :$

Test Statistic:

p -value:

Conclusion:

Repeat this analysis in R:

```
#Create a vector of the number of successes for both populations:
> NumLegalize <- c(202,118)
#Create a vector of the sample sizes from both populations:
> SampSize <- c(310,190)
#Run the hypothesis test:
> prop.test(x=NumLegalize, n=SampSize, conf.level=0.95, alternative="greater")
      2-sample test for equality of proportions with continuity correction
data:  NumLegalize out of SampSize
X-squared = 0.3541, df = 1, p-value = 0.2759
alternative hypothesis: greater
95 percent confidence interval:
 -0.04670842  1.00000000
sample estimates:
   prop 1    prop 2 
0.6516129 0.6210526
```

NOTES: This is the same ‘prop.test’ function we used to make inference about one proportion. Therefore,

1. If $H_a: p_1 < p_2$, then we would type `alternative = "less"`.
If $H_a: p_1 \neq p_2$, then we would type `alternative = "two.sided"`.
2. We would get slightly different values if we did this by hand. R uses a ‘continuity correction’ that we do not...

(b) Compute a 95% CI for $p_1 - p_2$ in R.

```
> prop.test(x=NumLegalize, n=SampSize, conf.level=0.95, alternative="two.sided")
      2-sample test for equality of proportions with continuity correction
data:  NumLegalize out of SampSize
X-squared = 0.3541, df = 1, p-value = 0.5518
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.06069792  0.12181846
sample estimates:
   prop 1    prop 2 
0.6516129 0.6210526
```

NOTE: To get a correct confidence interval in R, the alternative must be ‘two.sided’!

10.2 Comparing Two Means - Matched Pairs

Comparisons of two means are based on either two independent samples or two dependent samples:

Independent Samples

ex: compare mean heights of men and women

ex: compare average exam scores for two different stat classes

Dependent Samples

The dependent samples we will consider result from *matched pairs* experiments.

ex: compare the mean weight loss resulting from 2 different diet pills

Sample $2n$ people who are “matched” by weight.

Randomly assign one person in each pair to pill 1 and the other to pill 2.

This natural pairing causes the samples to be dependent.

ex: compare mean weights before and after taking a diet pill

Matched Pairs

Goal:

Definition: matched pairs

Data are *matched pairs* if each subject in one sample is matched with a subject in another sample in some meaningful way. A matched pair may be “before” and “after” observations on one subject or observations on two subjects that have been matched by characteristics that may influence the response variable (ex: gender, age, weight).

Why use matched pairs?

Method for Studying Matched Pairs:

Notation:

μ_D = true mean of paired differences

\bar{x}_D = sample mean of paired differences

s_D = sample standard deviation of paired differences

n_D = number of pairs in the sample

EXAMPLE 10.3

Six adults are chosen for a blood alcohol content (BAC) study. Each adult is given 3 beers at sea level and 3 beers at high altitude. The order of two drinking sessions is chosen at random for each subject.

Sea Level (SL)	High Altitude (HA)	Difference
0.007	0.013	-0.006
0.010	0.017	-0.007
0.009	0.015	-0.006
0.011	0.014	-0.003
0.008	0.010	-0.002
0.006	0.009	-0.003

Based on this data, we can show that

$$\bar{x}_D = -0.0045 \quad \text{and} \quad s_D = 0.0021.$$

Suppose the distribution of differences is normal. At the 0.01 level, test the hypothesis that the average BAC at sea level is less than the average BAC at high altitude.

Assumptions: random sample and distribution of differences is normal

Hypotheses:

$$H_0 : \mu_D = 0$$

$$H_a : \mu_D < 0$$

Test statistic:

$$t = \frac{\bar{x}_D - \mu_0}{s_D / \sqrt{n_D}} = \frac{-0.0045}{0.0021 / \sqrt{6}} = -5.25$$

p -value:

p -val = $P(t_5 < -5.25)$, where $0.001 = P(t_5 < -5.894) < P(t_5 < -5.25) < P(t_5 < -4.032) = 0.005$

Conclusion:

p -val $< 0.005 < 0.01 \Rightarrow$ **Reject H_0 . We have enough evidence (at the 0.01 level) to conclude that alcohol reaction is greater at HA than at SL.**

EXAMPLE 10.4

Suppose we are interested in the protein concentration (in grams/kg of wheat) of a winter wheat and a spring wheat. The growth and protein level of these wheats may be influenced by the location in which they're grown (temperature, altitude, soil characteristics, etc). Therefore, we select 20 different locations and plant both a winter wheat and a spring wheat in each location. That is, winter and spring wheat observations are matched by location. The following is a partial data set for our experiment:

Location	Spring Wheat Protein	Winter Wheat Protein	Difference (spring - winter)
1	122	87	35
2	171	145	26
3	144	116	28
\vdots	\vdots	\vdots	\vdots

The following are summary statistics for the sample data:

$$\bar{x}_D = 29.2 \quad \text{and} \quad s_D = 9.35117.$$

Calculate a 95% confidence interval for μ_D , the true mean difference (spring – winter) in protein concentration.

$$\begin{aligned} \bar{x}_D \pm t_{\alpha/2, n_D-1} \frac{s_D}{\sqrt{n_D}} &= 29.2 \pm 2.093 \cdot \frac{9.35117}{\sqrt{20}} \\ &= 29.2 \pm 4.38 \\ &= (24.82, 33.58) \end{aligned}$$

We are 95% confident that the average protein content for spring wheat is between 24.82 and 33.58 higher than for winter wheat.

10.3 Comparing Two Means - Independent Samples

Goal: Compare two (unknown) means corresponding to two independent samples.

The procedure for comparing two means of independent samples is similar to comparing two proportions of independent samples, only the details change. Mainly, we make inferences about the differences between two means ($\mu_1 - \mu_2$) through point estimation, interval estimation, and hypothesis testing.

To this end, we independently collect a random sample from each population:

Population	Pop. Mean	Pop. St. Dev.	Sample Size	Sample Mean	Sample St. Dev.
1	μ_1	σ_1	n_1	\bar{x}_1	s_1
2	μ_2	σ_2	n_2	\bar{x}_2	s_2

NOTE: We will assume for the rest of this chapter that σ_1 and σ_2 are unequal. Along with this case, the book also presents this material under the assumption that $\sigma_1 = \sigma_2$. The assumption that $\sigma_1 = \sigma_2$ results in a different standard error formula for $(\bar{x}_1 - \bar{x}_2)$, which we will not cover in class.

10.3.1 Point Estimation for $\mu_1 - \mu_2$

The obvious point estimate of $\mu_1 - \mu_2$ is _____.

The statistic $\bar{x}_1 - \bar{x}_2$ varies from sample to sample. That is, it has its own sampling distribution:

However, since σ_1 and σ_2 are unknown, we instead estimate the standard deviation using the standard error of $\bar{x}_1 - \bar{x}_2$:

In this case, we construct confidence intervals and hypothesis tests using the fact that

Note: This is a *conservative* approximation of the degrees of freedom (df). That is, the df are likely to be larger than this rule gives us. This means that if we used a more accurate approximation for the df, we would likely get a narrower confidence interval. There *is* a more accurate approximation for the df called ‘Welch’s df’. The formula is very complicated, but R will do it automatically.

10.3.2 Confidence Intervals for $\mu_1 - \mu_2$

Assumptions:

- 1.
- 2.

The $1 - \alpha$ confidence level confidence interval for $\mu_1 - \mu_2$ is

Note: The interpretation of the confidence interval for $\mu_1 - \mu_2$ is similar to that of confidence intervals for $p_1 - p_2$.

10.3.3 The Two-Sample t -Test for Comparing μ_1 and μ_2

Assumptions

1. 2 independent random samples
2. Both populations are approximately normal.
(Though 2-sample t procedures are robust to non-normality for larger sample sizes.)

Hypotheses

The null hypothesis is always $H_0: \mu_1 = \mu_2$, that is, that there is no difference between the 2 means.

- | | | |
|-------------------------|-------------------------|-------------------------|
| 1. $H_0: \mu_1 = \mu_2$ | 2. $H_0: \mu_1 = \mu_2$ | 3. $H_0: \mu_1 = \mu_2$ |
| $H_a:$ | $H_a:$ | $H_a:$ |

Test Statistic

p -value

Hypothesis 1: $p\text{-val} = P(t_{df} < t)$

Hypothesis 2: $p\text{-val} = P(t_{df} > t)$

Hypothesis 3: $p\text{-val} = 2P(t_{df} > |t|)$

where df is the smaller of $n_1 - 1$ and $n_2 - 1$.

Conclusion

If $p\text{-value} < \alpha$, reject H_0

If $p\text{-value} \geq \alpha$, fail to reject H_0

EXAMPLE 10.5

Verbal SAT scores were recorded for independent samples of students who intend to major in engineering and students who intend to major in literature. Suppose histograms of both samples show no strong skewness and no outliers. From the data we calculate

Intended Major	Sample Size	Sample Mean	Sample St. Dev.
(1) Engineering	44	446.9	42.0
(2) Literature	44	534.2	45.5

Let

μ_E = true mean verbal score for intended engineering majors

μ_L = true mean verbal score for intended literature majors

- (a) Calculate a 90% confidence interval for $\mu_E - \mu_L$, the true difference in mean verbal scores for students intending to major in engineering and students intending to major in literature.

- (b) Determine if there is strong evidence at the 0.01 level that the true mean verbal score of intended engineering majors is less than that of intended literature majors.

Assumptions:

Hypotheses:

$H_0 :$

$H_a :$

Test statistic:

p-value:

Conclusion:

EXAMPLE 10.6

The data set <http://www.stat.umn.edu/~wuxxx725/data/class.txt> contains data collected from a recent survey of U of M students.

Year	Gender	Height	TV	Siblings	DistHome	Haircut
1	M	75	60	5	10	15
5	F	65	30	1	1	45
3	M	70	0	1	1	18
.
.

Along with other questions, the students were asked about the amount of money spent on their latest haircut.

1. Construct and interpret a 95% confidence interval for the true difference in the mean amount of money men and women spend on a haircut.

```
#Sample of Male Haircuts:
> MaleHair <- Haircut[Gender == "M"]
#Sample of Female Haircuts:
> FemHair <- Haircut[Gender == "F"]
> t.test(x=MaleHair, y=FemHair, alternative="two.sided", conf.level=0.95)
Welch Two Sample t-test
t = -6.5631, df = 92.461, p-value = 3.011e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.333509  -9.280328
sample estimates:
mean of x mean of y
 12.46667  25.77358
```

2. Test at the 0.05 level whether male students spend less money on average than female students on their haircuts.


```
> t.test(x=MaleHair, y=FemHair, alternative="less", conf.level=0.95)
      Welch Two Sample t-test
data:  MaleHair and FemHair
t = -6.5631, df = 92.461, p-value = 1.506e-09
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -9.938168
sample estimates:
mean of x mean of y
 12.46667  25.77358
```

NOTE:

- If $H_a: \mu_x > \mu_y$, then we would type alternative = "greater".
- If $H_a: \mu_x \neq \mu_y$, then we would type alternative = "two.sided".

EXAMPLE 10.3 (DONE THE WRONG WAY)

Treat the data as independent samples, ignoring the fact that the BAC measurements at sea level and at high altitude were made on the same 6 people. At the 0.01 level, test the hypothesis that the average BAC at sea level is less than the average BAC at high altitude.

```
> BAC.sl<-c(0.007, 0.010, 0.009, 0.011, 0.008, 0.006)
> BAC.ha<-c(0.013, 0.017, 0.015, 0.014, 0.010, 0.009)
> mean(BAC.sl)
[1] 0.0085
> mean(BAC.ha)
[1] 0.013
> sd(BAC.sl)
[1] 0.001870829
> sd(BAC.ha)
[1] 0.00303315
> t.test(x = BAC.sl, y = BAC.ha, alternative = "less", conf.level = 0.95)
```

Welch Two Sample t-test

```
data:  BAC.sl and BAC.ha
t = -3.093, df = 8.323, p-value = 0.007062
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.001808149
sample estimates:
mean of x mean of y
 0.0085    0.0130
```

CHAPTER 14: ANALYSIS OF VARIANCE

EXAMPLE 14.1:

What is the most effective method for treating anorexia? 72 anorexic teenage girls were randomly assigned to one of three treatments. The first treatment group received cognitive behavioral therapy in which girls are taught to identify the thinking that triggers their eating disorder and to replace it with other thoughts meant to prevent this behavior. The second treatment group attended family therapy and the third group served as a control group and did not receive any therapy. Each girl was weighed before her treatment began and weighed again upon completion of the treatment. This data can be found at

<http://www.stat.umn.edu/~wuxxx725/data/anorexia.txt>.

Let

μ_1 = mean weight gain after completing cognitive behavioral therapy

μ_2 = mean weight gain after completing family therapy

μ_3 = mean weight gain with no treatment

Are any of the treatments better or worse than the others? That is, are there any significant differences among μ_1 , μ_2 , and μ_3 ?

GOAL:

Why not just do a two-sample t-test for each pair of means?

1. We would have to do $\binom{g}{2} = \frac{g(g-1)}{2}$ tests, where g = number of means we want to compare.

BUT: Type I error rate increases as the number of tests increases

2. This would only allow us to compare 2 groups at a time but we want to know if means differ among *all* groups – want to answer this with one test.

14.1 One-Way ANOVA

Notation:

g	=	number of groups we are comparing
n_i	=	sample size of the i th group
N	=	overall sample size = $n_1 + n_2 + \dots + n_g$
\bar{y}_i	=	sample mean of the i th group
\bar{y}	=	overall sample mean (sample mean of all the observations)
s_i	=	sample standard deviation of the i th group

One-Way ANOVA compares population means among g different groups.

Hypothesis:

H_0 :

H_a :

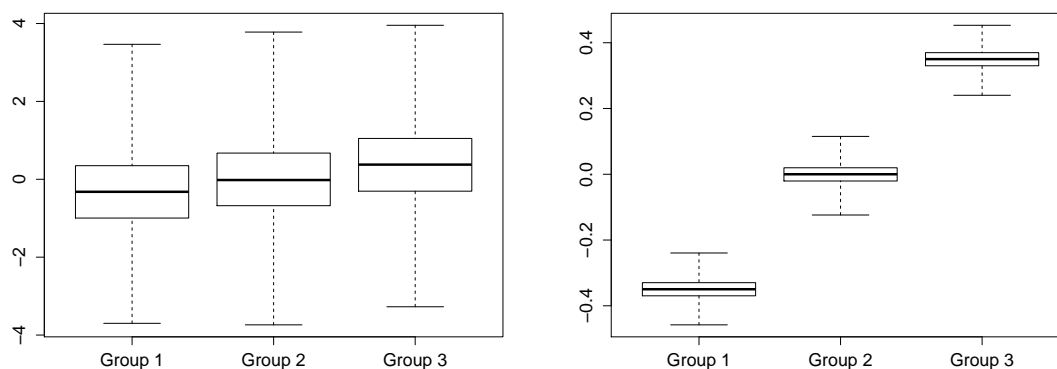
To test H_0 we do an *analysis of variance*. That is, we compare

- 1.
- 2.

Question:

If our goal is to compare the *means* of several populations, why are we doing an analysis of *variance*?

Answer:



Notice that the means of the three groups are the same in both pictures!

plot 1: variation between sample means is small in comparison to variability within groups

plot 2: variability between sample means is large in comparison to variability within groups

Conclusions:

Differences in means in plot 1 are probably due to chance whereas in plot 2 the differences are probably due to true difference among groups and not to chance

Two means are significantly different only if their difference is large *relative* to the variability of observations within each group.

Side-by-side boxplots provide a *graphical* comparison of the variability within and between groups. However, in order to test H_0 we also need a numerical summary of this information.

Measuring Variability *Between* Groups

SSG = Sum of Squared Deviations for Groups = $\sum_{i=1}^g n_i(\bar{y}_i - \bar{y})^2$

$MSG = \frac{SSG}{g-1}$ = a measure of how much means vary from group to group.

Measuring Variability *Within* Groups

SSE = Sum of Squared Deviations for Error = $\sum_{i=1}^g (n_i - 1)s_i^2$

$MSE = \frac{SSE}{N-g}$ = a measure of how much observations vary within each group.

MSE is also an estimate of σ^2 , the population variance for each group.

Measuring *Overall* Variability

$SST = SSG + SSE = \sum (y - \bar{y})^2$

Measures how much observations vary from the overall mean.

The ANOVA F Test Statistic

The F statistic measures how compatible the data is with $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$ by comparing the between group variability to the within group variability:

Interpretation:

The ANOVA Table

The analysis of variance is summarized using an ANOVA Table:

Source of Variation	df	SS	MS	F
Group				
Error				
Total				

The One-Way ANOVA F Test

Assumptions:

1. g independent random samples from g populations
2. Normality: **Each population has a normal distribution with unknown mean**
3. Equal Variance: **Each population has an equal (but unknown) standard deviation σ**

Notes:

- The F test is robust to departures from normality and equal variance.
- Graphical methods (such as histograms or boxplots) can be used to check the assumptions of normality and equal variance.

Hypothesis:

H_0 :

H_a :

Test Statistic:

p -value: Recall that the larger F is, the more evidence we have *against* H_0 .

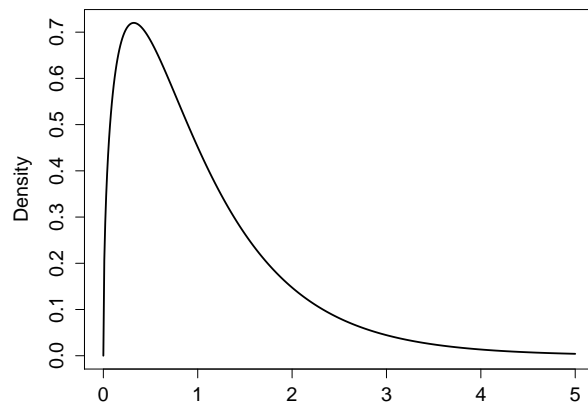
Conclusion:

If p -value $< \alpha$, reject H_0

If p -value $\geq \alpha$, fail to reject H_0

The F Distribution

The $F(3, 60)$ density curve:



Properties:

1. Non-negative
2. Right-skewed
3. Mean is approximately 1 (exact value: $\frac{df_2}{df_2-2}$ for $df_2 > 2$.)
4. Mode is also approximately 1 (exact value: $\frac{df_1-2}{df_1} \frac{df_2}{df_2+2}$ for $df_1 > 2$.)

 F calculations:

Table D in Appendix A only gives the 95th percentiles of the F distribution. Thus, Table D can tell us only whether an F statistic will result in a p-value greater or less than .05. The pf function in R can be used to compute exact p-values for a given F statistic:

```
> pf(3.493, df1=2, df2=20, lower=F) #p-value for F=3.493 with df1=2 and df2=20
[1] 0.04999364
> pf(7.42, df1=5, df2=15, lower=F) #p-value for F=7.42 with df1=5 and df2=15
[1] 0.001103249
```

EXAMPLE 14.1 CONTINUED

Recall that we want to compare anorexia treatment methods. Let

$$\begin{aligned} \text{change} &= \text{weight change (post-treatment - pre-treatment weight)} \\ \text{therapy} &= \text{treatment group} \end{aligned}$$

Suppose the assumptions needed for the one-way ANOVA F test are fulfilled. Fill in the following ANOVA table and use it to perform the test for equal mean weight gain among the three anorexia treatments at the 0.05 level:

Source of variation	df	SS	MS	F
groups		614.6		
error		3910.7		
Total				

Assumptions:

Hypothesis:

H_0 :

H_a :

Test statistic:

p -value:

Conclusion:

We can also use R to run the one-way ANOVA F test and obtain an exact p-value:

```
> aov1 <- aov(change ~ therapy) #The form of the equation is always:
                                # aov(numerical variable ~ group variable)
> summary(aov1)                 #This returns the ANOVA table and p-val of the F test
      Df Sum Sq Mean Sq F value    Pr(>F)
therapy    2  614.6   307.3  5.4223 0.006499 **
Residuals 69 3910.7    56.7
```

14.2 Follow-Up to the ANOVA F -test

When we reject the one-way ANOVA F -test for equal means we can only conclude that at least two of the group means are significantly different than each other. However, we cannot use this test to determine *which* means are different or by *how much* they differ. In this case, we should perform further analysis, but we need to be careful...

Example:

Suppose we reject $H_0: \mu_1 = \mu_2 = \mu_3$. To determine which means are significantly different, we could use the methods of Chapter 10 to construct $100(1 - \alpha)\%$ confidence intervals for each pair of means:

$$\mu_1 - \mu_2$$

$$\mu_1 - \mu_3$$

$$\mu_2 - \mu_3$$

What is the problem here?

To construct the confidence intervals so that the desired confidence level extends to the entire *set* of intervals as opposed to each *individual* interval, *multiple comparison* procedures can be used.

Definition: Multiple Comparisons

Multiple comparison methods perform several separate statistical analyses with a confidence level that applies simultaneously to the entire set of analyses rather than to each analysis separately.

Applied to confidence intervals:

Several multiple comparison methods have been developed. A common one is the **Tukey Honest Significant Difference** (Tukey H.S.D.) method. The math involved in the Tukey H.S.D. is complicated, so we'll just show how to do it in R and discuss the interpretation.

EXAMPLE 14.1 CONTINUED

Use R to construct Tukey H.S.D. multiple comparison confidence intervals for the mean weight gain for the three different anorexia treatments. Use an overall confidence level of $1 - \alpha = 0.95$.

```
> aov1 <- aov(change ~ therapy)
> TukeyHSD(aov1, "therapy", conf.level=0.95) #Always list the ANOVA first and the
                                             #name of the group variable second.

    Tukey multiple comparisons of means
      95% family-wise confidence level
Fit: aov(formula = change ~ therapy)
$therapy
      diff      lwr      upr    p adj
control-cog -3.456897 -8.327276  1.413483 0.2124428
family-cog   4.257809 -1.250554  9.766173 0.1607461
family-control 7.714706  2.090124 13.339288 0.0045127
```

Conclusion:

Mean weight gain for the family therapy group was significantly higher than for the control group (all values in the interval are > 0).

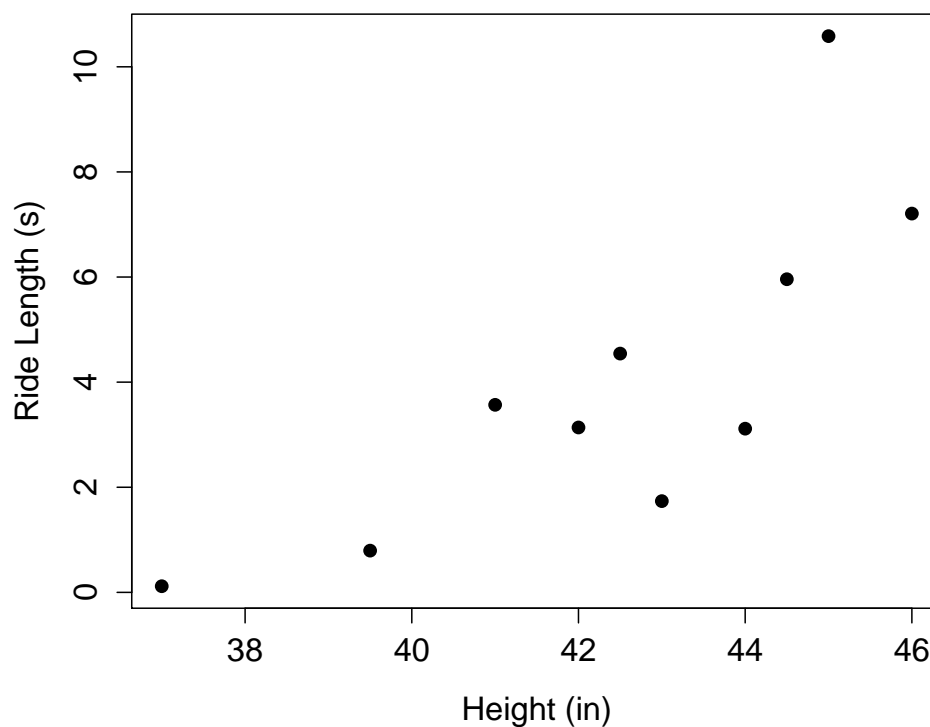
However, since 0 falls between the lower and upper values of the other 2 intervals, the mean weight gain is not significantly different among the control and cognitive behavior groups nor among the family and cognitive behavior groups.

CHAPTER 3: TWO-VARIABLE ASSOCIATIONS

Motivating Example

“Mutton busting” is a popular rodeo intermission event in which children are placed on top of a sheep (mutton) and ride around until they fall off (busting). The following table contains the ride times and heights of 10 little mutton busters:

Ride Length (s)	0.12	0.79	3.57	3.14	4.54	1.74	3.11	5.96	10.58	7.21
Height (in)	37.0	39.5	41.0	42.0	42.5	43.0	44.0	44.5	45.0	46.0



Using what we learned in previous chapters, we can *separately* explore the heights of the mutton busters as well as their ride times. However, we are also interested in the *relationship* between the height of a child and how long they can stay on the mutton. For instance, from the plot it appears that taller children tend to have longer ride times.

Goal:

Definition: association

Two variables measured on the same subjects are *associated* if some values of one variable tend to occur more often with some of the second variable.

When exploring the relationship between two variables, we typically distinguish between which is the *response* variable and which is the *explanatory* variable.

Sometimes this distinction is obvious, while at other times different variables can be considered a response, depending on the goal of the analysis.

Definition: response and explanatory variables

A *response variable* measures an outcome that is thought to occur in response to an *explanatory variable*.

Examples:

- In the relationship between blood alcohol content (BAC) and the # of beers one drinks,

response variable:

explanatory variable:

- In the relationship between one's gender and their political party affiliation,

response variable:

explanatory variable:

In the next couple of chapters we will explore the following types of associations:

Chapter 11: Associations between 2 **categorical** variables

Chapter 12: Associations between 2 **quantitative** variables

CHAPTER 11: ASSOCIATION BETWEEN TWO CATEGORICAL VARIABLES

EXAMPLE 11.1

“Entrance polls” from the 2012 Iowa Republican caucuses collected information from 1,787 caucus voters. This survey data is summarized in the following contingency table which breaks down the number of votes for each candidate by family income level:

		Candidate				Total
		Romney	Santorum	Paul	Other	
Family Income Level	Under \$50K	94	112	183	201	590
	\$50K-\$100K	146	202	147	203	698
	\$100K or more	179	120	70	130	499
Total		419	434	400	534	1787

Name the two *categorical* variables in this study:

Is there an association between candidate and income level among the entire *population* of Iowa Republican caucus voters?

11.1 Chi-Squared Test for Independence

Definition: independent variables

Two categorical variables are *independent* if the distribution of one of the variables is not influenced by the observed value of the other.

Goal:

How?

Expected Cell Counts:

In general, when two categorical variables are independent, we can calculate the expected value of each cell in the contingency table:

Chi-Squared Test for Independence

1. Assumptions

- (a) random sample
- (b) large enough sample size so that expected cell count ≥ 5 in all cells

2. Hypothesis

 $H_0:$ $H_a:$

3. Test Statistic

Recall: A test statistic is a measure of how compatible the data is with H_0 . In this case, a test statistic should measure the degree to which the observed contingency table agrees with the assumption of independence between the two variables.

When H_0 is true, do you expect X^2 to be a large or a small number?

What is the distribution of X^2 when H_0 is true?

That is, X^2 has a _____ distribution with $df =$ _____, where r =number of rows and c =number of columns in the contingency table. (See below.)

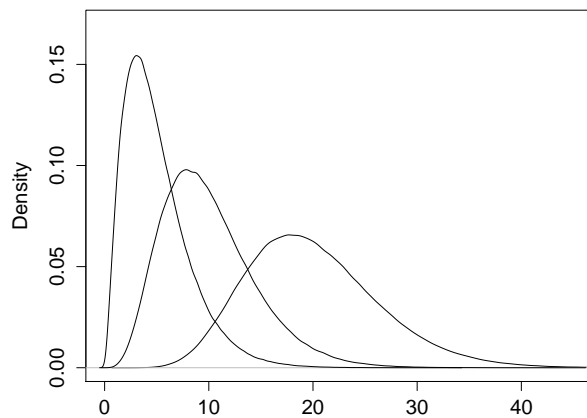
4. p -value

5. Conclusion

If p -value $< \alpha$, reject H_0

If p -value $\geq \alpha$, fail to reject H_0

The Chi-Squared Distribution



Properties:

1. continuous
2. right skewed
3. only takes on non-negative values (≥ 0)
4. shape is specified by the df
5. the larger df , the more spread out

χ^2 calculations

Use Table C in Appendix A to calculate a range for the p-value of the chi-squared test. The `pchisq` function in R can be used to compute an exact p-value for a given test statistic X^2 .

1. Let $df=20$ and find $P(\chi^2 \geq 34.17)$.

```
> pchisq(34.17, df=20, lower=F)
[1] 0.02499745
```

2. Let $df=13$ and estimate $P(\chi^2 \geq 24)$.

```
> pchisq(24, df=13, lower=F)
[1] 0.03113006
```

EXAMPLE 11.1 CONTINUED

- (a) Calculate the expected cell counts under the assumption that one's family income level and candidate preference are independent.

	Romney	Santorum	Paul	Other	Total
Under \$50K	94 (138.3)	112 (143.3)	183 (132.1)	201 (176.3)	590
\$50K-\$100K	146 (163.7)	202 (169.5)	147 (_____)	203 (_____)	698
\$100K or more	179 (117.0)	120 (121.2)	70 (_____)	130 (_____)	499
Total	419	434	400	534	1787

- (b) At the 0.05 level, test for an association between one's family income and their candidate preference.

Assumptions:

Hypotheses:

H_0 :

H_a :

Test statistic:

***p*-value:**

Conclusion:

EXAMPLE 11.2

Are smoking and divorce related? A random sample of 1669 adults were interviewed about their marriage and smoking statuses:

		Divorced?		Total
		Yes	No	
Smoke?	Yes	238	247	485
	No	374	810	1184
Total		612	1057	1669

Use R to test for an association between smoking and divorce at the 0.01 level.

- Step 1: Put the data in matrix form.

```
> dat <- matrix(c(238,247,374,810), nrow=2, byrow=T)
> dat
      [,1] [,2]
[1,]  238  247
[2,]  374  810
```

NOTE: 'nrow=2' tells R that there are 2 rows in the table and 'byrow=T' tells R that we are entering in the data for the first row followed by the data for the second row (as opposed to entering data for the first column followed by the second column).

- Step 2: Run the Chi-square test.

```
> mytest <- chisq.test(dat, correct=F)
> mytest
      Pearson's Chi-squared test
data:  dat
X-squared = 45.292, df = 1, p-value = 1.697e-11
```

Assumptions:

random sample, expected cell counts > 5

Hypotheses:

H_0 : smoking and divorce are independent

H_a : smoking and divorce are dependent

Test statistic:

$$X^2 = 45.29, \text{ df} = 1$$

p -value:

$$p\text{-val} = P(\chi_1^2 \geq 45.29) = 1.697 \times 10^{-11}$$

Conclusion: Reject H_0 . There is a significant association between smoking and divorce at the 0.01 level.

11.2 Measures of Association

The chi-squared test addresses:

The chi-squared test does *not* address:

Definition: risk

The “*risk*” of an outcome is the probability of its occurrence.

Definition: relative risk

The ratio of risks for two groups is called the *relative risk* and can be used to measure the strength of the association between two categorical variables.

EXAMPLE 11.2 CONTINUED

We previously showed that there is a significant association between the incidence of smoking and divorce. We now want to *describe* this association.

- (a) What is the estimated risk of divorce among smokers?

- (b) What is the estimated risk of divorce among non-smokers?

- (c) Calculate and interpret the estimated relative risk of divorce among smokers and non-smokers.

CHAPTER 12: REGRESSION ANALYSIS

EXAMPLE 12.1

How well does the production cost of a movie predict how well it will do at the box office? Data (courtesy of *Houghton Mifflin*) was collected from a random sample of 10 Hollywood movies:

Box	Prod	Promo	Book
85.1	8.5	5.1	4.7
106.3	12.9	5.8	8.8
.	.	.	.
.	.	.	.

where

Box = 1st year box office receipts (millions)

Prod = production costs (millions)

Promo = promotional costs (millions)

Book = book sales (millions)

A full data set can be found at <http://www.stat.umn.edu/~wuxxx725/data/movies.txt>.

GOAL:

Explore the association between 2 *quantitative* variables.

ex: production cost and box office sales

The Approach:

1. Get to know the sample data using both graphical and numerical summaries.
2. Use the sample data to make inferences about the true population relationship.

12.0 Exploring the Data (A Return to Chapter 3)

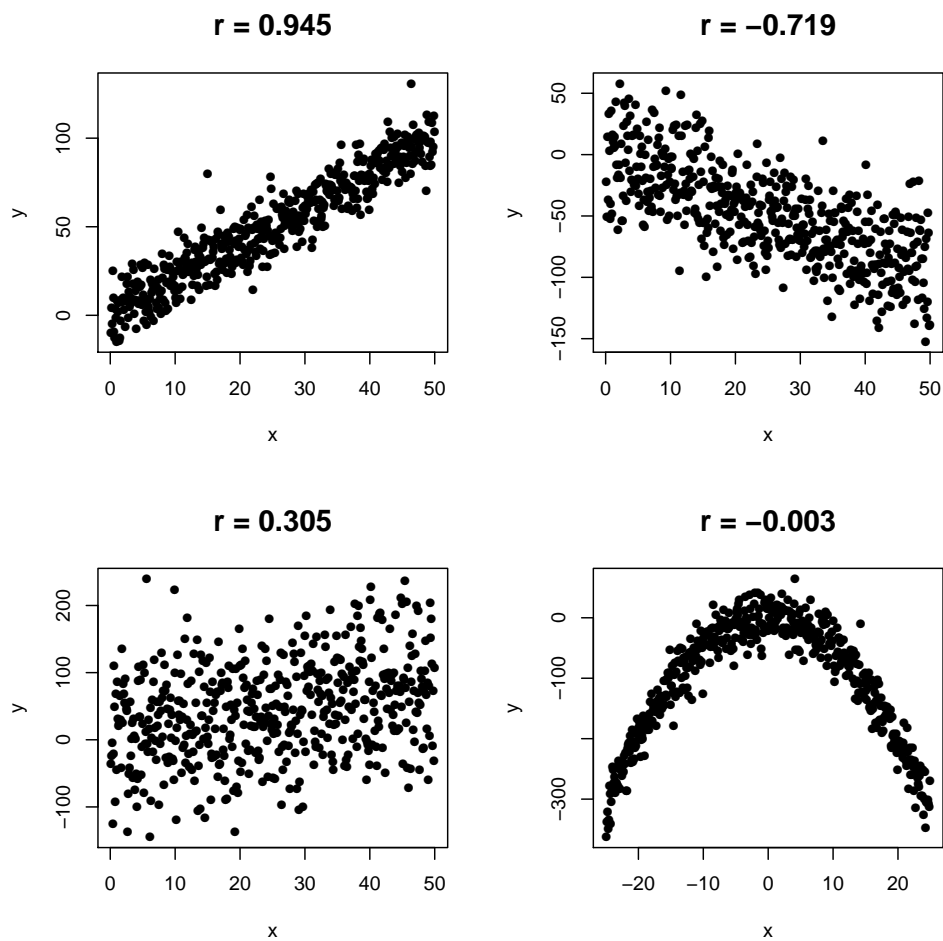
3.2.1 Graphical Summaries - The Scatterplot

The *scatterplot* is a graphical tool used to display the relationship between two quantitative variables.

Constructing a Scatterplot:

1. Label the x -axis (horizontal) with the explanatory variable.
Label the y -axis (vertical) with the response variable.
2. Represent each observation with a point in the graph at its (x, y) coordinate.

Examples:



What To Look For In A Scatterplot

1. overall pattern

ex: linear, curved, etc

2. strength

How closely do the points follow a pattern?

ex: weak, moderate, strong

3. direction

Do the variables have a positive or negative association?

Definition: positive association

Two quantitative variables are *positively associated* when high values of one of the variables tend to occur with high values of the other.

Definition: negative association

Two quantitative variables are *negatively associated* when high values of one of the variables tend to occur with low values of the other (and vice versa).

4. outliers

Are there any unusual points that fall outside the cloud of data points?

EXAMPLE 12.2

Is there an association between Olympic winning times for the 200 meter dash and the year in which the Olympics took place? The following is a partial data set of the Olympic winning times (in seconds) starting in 1908 and ending in 1996 where

Year = number of years after 1900

Time = Olympic winning times for the 200m dash (in seconds)

(The full data set can also be found at <http://www.stat.umn.edu/~wuxxx725/data/dash.txt>.)

	Year	Time
1	8	22.60
2	12	21.70
3	20	22.00
.	.	.
.	.	.

Which is the explanatory variable and which is the response?

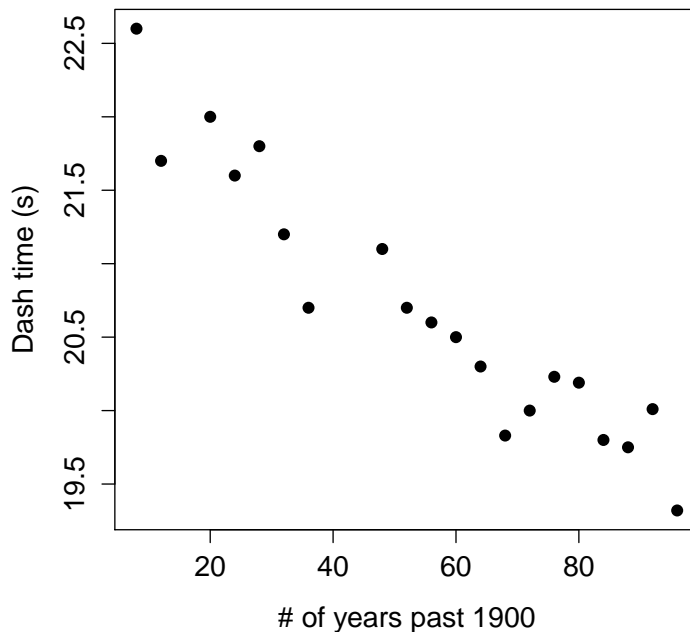
response = Time

explanatory = Year

Use R to draw a scatterplot of the data.

```
> plot(Year, Time, xlab="# of years past 1900", ylab="Dash time (s)", pch=16)
```

- When using the “plot” function, always list the x -axis (explanatory) variable first and the y -axis (response) variable second.
- Recall: “xlab” = x -axis label, “ylab” = y -axis label, and “main” = title
Also, “pch = 16” tells R to use solid dots (the default is open circles).



Describe the relationship between the Olympic winning time of the 200m dash and the year in which the Olympics took place:

strong, negative, linear relationship

200m dash times have been improving (decreasing) over time

3.2.2 Numerical Summaries - Correlation

A scatterplot only allows us to *eyeball* the strength of a linear relationship between two quantitative variables. We also want to *quantify* the strength of this relationship.

Notation:

Let $x = \{x_1, x_2, \dots, x_n\}$ and $y = \{y_1, y_2, \dots, y_n\}$ denote two paired samples of size n corresponding to two different quantitative variables. Also, let

\bar{x} = sample mean of the x data

s_x = sample standard deviation of the x data

\bar{y} = sample mean of the y data

s_y = sample standard deviation of the y data

Definition: correlation (r)

Correlation is a measure of the strength and direction of the **linear** relationship between two quantitative variables. The sample correlation between two variables x and y can be calculated using

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Notice the z-scores!

Properties of r :

1. The value of r does not change if we reverse the roles of x and y .

(It doesn't matter which is labeled "response" and which is labeled "explanatory".)

2. The value of r does not change if we change the units of the variable.

ex: converting pounds to kilograms won't affect r

3. r *only* measures the strength and direction of a *linear* (not curved) relationship.

4. r is not resistant to outliers.

This is obvious since the formula for r involves \bar{x} and \bar{y}

5. $-1 \leq r \leq 1$

Direction:

$r > 0 \Rightarrow$ positive association

$r < 0 \Rightarrow$ negative association

$r = 0 \Rightarrow$ uncorrelated (no linear relationship)

Strength:

$r \approx 0 \Rightarrow$ weak linear relationship.

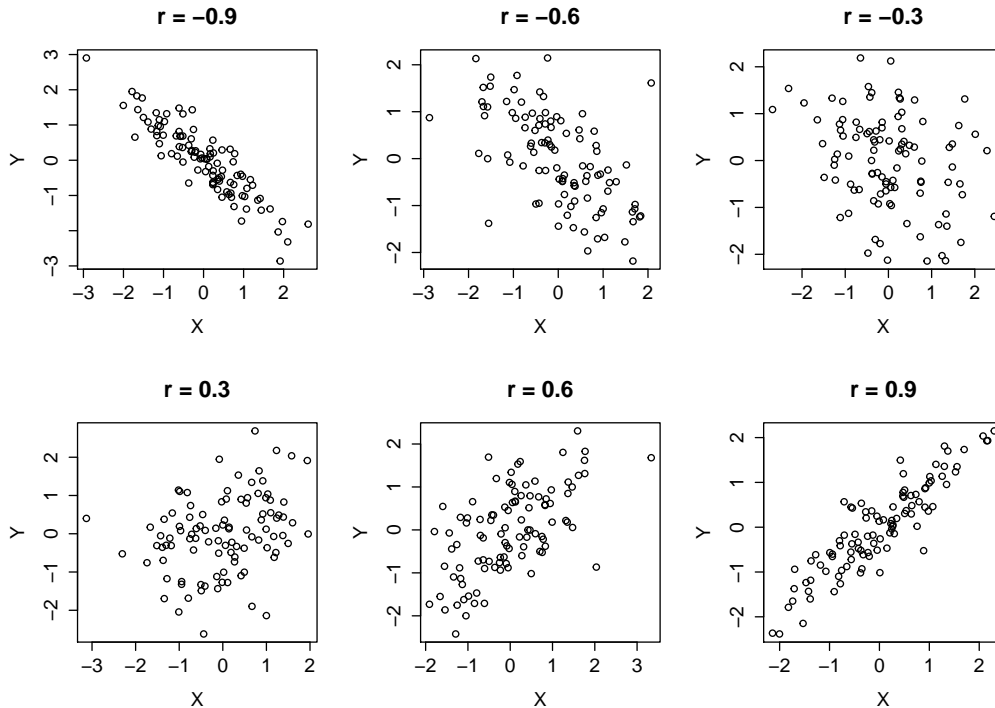
$r \approx \pm 1 \Rightarrow$ strong linear relationship.

$r = 1$ or $r = -1$ only if all points fall exactly on a straight line.

Draw a picture:

$[-1, -.8) =$ strong, negative; $[-.8, -.5) =$ moderate, negative;

$[-.5, .5] =$ weak; $(.5, .8] =$ moderate, positive; $(.8, 1] =$ strong, positive



Refer back to the scatterplots on page 161 for examples.

EXAMPLE 12.2 CONTINUED

Use R to calculate the correlation between the year in which the Olympics took place and the winning time for the 200 meter dash:

```
> cor(Year, Time)
[1] -0.9496326
> cor(Time, Year)
[1] -0.9496326
```

$r = -0.950$. Therefore, there is a strong, negative linear association between Time and Year. This agrees with our interpretation of the scatterplot.

3.2.3 Numerical Summaries - Least Squares Regression

Correlation provides a measure of the *strength* and *direction* of a linear relationship. We will now learn how to use **least squares regression** to provide a numerical description of the *pattern* of the linear relationship between a (quantitative) explanatory variable and a (quantitative) response variable.

Definition: regression line

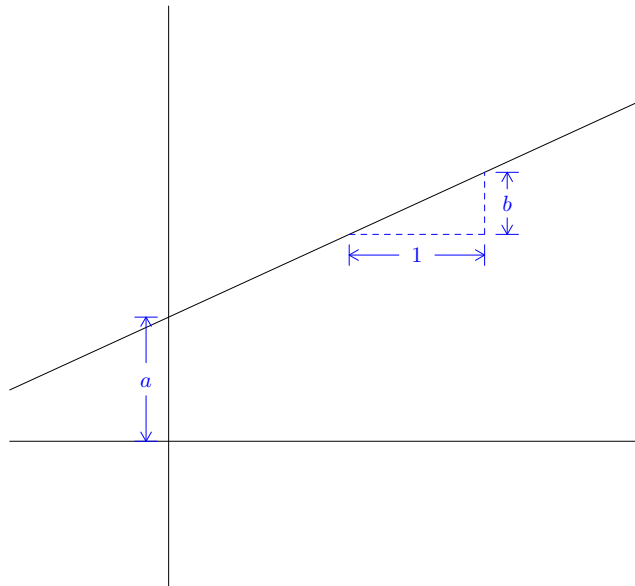
A *regression line* formulates the linear relationship between a response variable (y) and explanatory variable (x) and can be used to predict the value of y for a given value of x .

Sample Regression Line

Suppose we have explanatory variable x and a response variable y with observed data pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Then the equation for the sample regression line is

$$\hat{y} = a + b x$$

- \hat{y} = the predicted value of y at x
- a = intercept (value of y when $x = 0$)
this may not have any interpretive value if no observations have x values near 0
- b = slope of the line = expected change in y per unit change in x



Relationship Between r and b :

$b > 0 \Rightarrow$ **positive slope (positive association)**

Therefore if $b > 0$, $r > 0$ (and vice versa)

Similarly, if $b < 0$, $r < 0$ (and vice versa)

Using the Regression Line for Prediction:

We can predict the value of the response variable at some value of x by plugging x into the equation of the regression line.

Example: Suppose we have a regression line with $a = 2$ and $b = -1/4$.

- (a) Write down the formula and draw a picture of the regression line.

$$\hat{y} = 2 - \frac{1}{4}x$$

- (b) Predict y for $x = 100$.

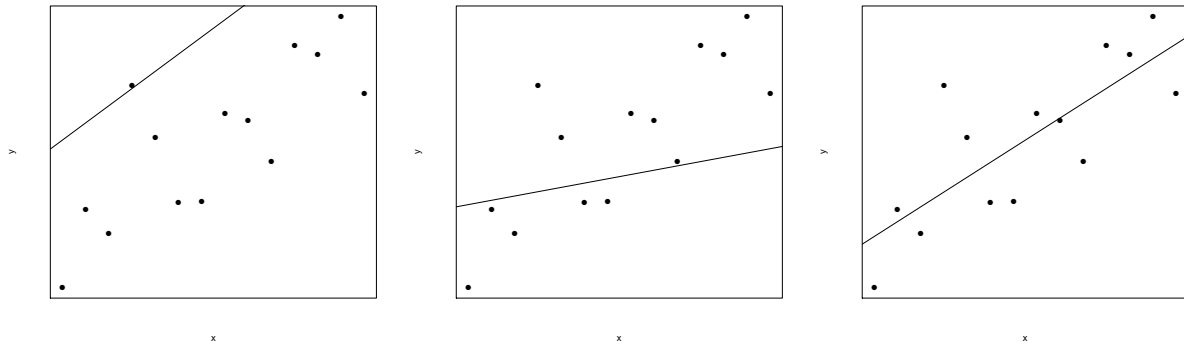
$$\hat{y} = 2 - \left(\frac{1}{4}\right)(100) = -23$$

- (c) Predict y for $x = 4$.

$$\hat{y} = 2 - \left(\frac{1}{4}\right)(4) = 1$$

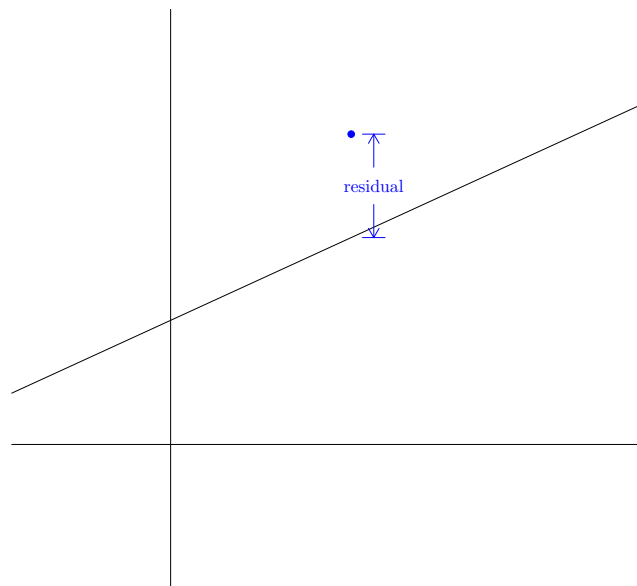
Finding a and b

As we've seen before, the observations will (most likely) not all fall on a straight line. So how do we pick a and b for the regression line? Consider a simple example:

**Definition: residual**

A *residual* is the prediction error for an observation. Specifically, it is the difference between an observed value of the response variable (y) and the value predicted by the regression line ($\hat{y} = a + b x$):

$$\text{residual} = y - \hat{y}$$



NOTES:

1. Cannot switch the order of y and \hat{y} in the formula for the residual.

2. What does it mean when the residual = 0?

The predicted value is spot on.

3. What does it mean when the residual < 0 ? > 0 ?

residual $< 0 \Rightarrow y < \hat{y} \Rightarrow$ predicted value is too high.

residual $> 0 \Rightarrow y > \hat{y} \Rightarrow$ predicted value is too low.

GOAL:

Choose a and b so that

- residuals are small

- not systematically over- or under- predicting the true value.

Least Squares RegressionDefinition: residual sum of squares

The *residual sum of squares* (RSS) is one measure of how well a regression line predicts values of the response variable. It is *literally* the sum of the squared residuals:

$$\text{RSS} = \sum (\text{residual})^2 = \sum (y - \hat{y})^2$$

Least Squares Criterion:

choose a and b to minimize RSS

Least Squares Line:

$$\hat{y} = a + b x$$

where

$$b = r \frac{s_y}{s_x}$$

$$a = \bar{y} - b \bar{x}$$

Properties of the Least Squares Regression Line (LSL)

1. Changing x and y around will give a different regression line.
2. Some residuals are positive and some residuals are negative, but it is *always* true that

$$\sum \text{residual} = \sum (y_i - \hat{y}_i) = 0$$

Interpretation:

low predictions are balanced by high predictions

3. RSS for the LSL is smaller than for any other line.
4. The LSL always passes through the point (\bar{x}, \bar{y}) .
i.e. $\bar{y} = a + b\bar{x}$
5. r^2 provides a measure of how well the LSL describes the relationship between x and y (where r is the correlation).
 - $r^2 =$ **proportion of variation in y that is explained by its linear relationship with x .**
 - $0 \leq r^2 \leq 1$
 - **The closer r^2 is to 1 the better x can be used to predict y**
 - $r^2 = 1 \Rightarrow$ **all variability in y is explained by its linear relationship with x**

EXAMPLE 12.2 CONTINUED

Recall: We are interested in quantifying the linear relationship between the Olympic winning time of the 200m dash and the year in which the Olympics took place. Let

$x = \text{Year} = \text{number of years after 1900}$

$y = \text{Time} = \text{Olympic winning times for the 200m dash (in seconds)}$

Step 1: Graphical Summary

We already looked at a scatterplot for this data. (see page 163)

Step 2: Numerical Summaries using R

```
> LSL <- lm(Time ~ Year) # "lm" fits the least squares line using Time as the response
                           #variable and Year as the explanatory variable
                           #We store this information in "LSL"
> summary(LSL)             #Gives values of a, b, r-squared, etc
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.355087   0.143763  155.50 < 2e-16 ***
Year         -0.030266   0.002354  -12.86 1.65e-10 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2838 on 18 degrees of freedom
Multiple R-squared:  0.9018, Adjusted R-squared:  0.8963
F-statistic: 165.3 on 1 and 18 DF,  p-value: 1.649e-10
```

Notes:

- Function “lm” stands for “linear model”.
(That is, “lm” is the *letter* l and the letter m. Don’t make a typo!)
- The form of the “lm” function is always “lm(response ~ explanatory)”.
- “LSL” (or whatever other name you assign your results) has more information than just coefficient estimates and r^2 . Let’s take a look:

```
> names(LSL)           #shows what LSL holds
[1] "coefficients" "residuals" "effects" "rank"
[5] "fitted.values" "assign" "qr" "df.residual"
[9] "xlevels" "call" "terms" "model"
> LSL$fitted.values   #predicted values of Time based on the regression line
> LSL$residuals       #residuals
```

- (a) Use the R output to find the equation of the least squares regression line.

$$\hat{y} = 22.36 - 0.0303x$$

- (b) Interpret the value of the slope.

On average, 200 meter dash times improve (decrease) by 0.0303 seconds every year.

- (c) By how much would we expect the 200m dash time to improve from one Olympics to the next (4 years apart)?

$$-0.0303(4) = -0.1212$$

- (d) Estimate what the winning 200m dash time might have been had the Olympics been held in 1950.

$$\hat{y} = 22.36 - 0.0303(50) = 20.84$$

- (e) State and interpret r^2 .

$$r^2 = 0.902$$

Therefore, about 90% of the variability in winning times is accounted for by its linear relationship with year.

The other 10% of the variation in 200m dash time is accounted for by other factors that are not in this study (ex: temperature, running surface, etc)

- (f) Calculate the interpret the sample correlation r .

$$r = \text{sgn}(b)\sqrt{r^2} = (-1)(.950) = -.950$$

- (g) The 2016 Olympics 200m dash was won by Usain Bolt of Jamaica in 19.78 seconds. Find the residual for this observation.

$$\hat{y} = 22.355 - 0.0303(116) = 18.84$$

$$\text{residual} = y - \hat{y} = 19.78 - 18.84 = 0.94$$

The prediction provided by the regression line for the 2016 Olympics was not very good, even though the relationship in the data looks quite linear and the r^2 value is high.

Why did the prediction fail? (There could be multiple reasons.)

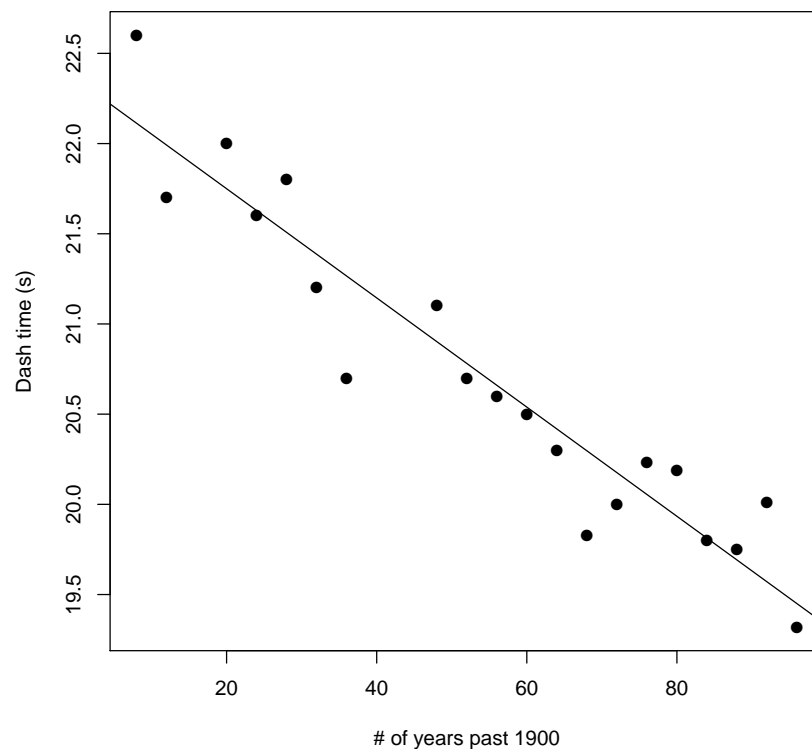
Extrapolation: The value of x in the new observation is too far outside of the range of the observed values of x .

Non-linearity: While the r^2 value of this model appears to be high, there might exist other non-linear models that work even better. For example, we may consider regressing the winning *speeds* of the Olympics 200m dash (instead of winning times) on the years that the Olympics took place.

Plotting the Least Squares Line in R

After using R to find the least squares line, we can draw a graph that includes the data points along with the fitted line.

```
#Fit the line:  
> dashline <- lm(Time ~ Year)  
#Draw the scatterplot:  
> plot(Year, Time, xlab="# of years past 1900", ylab="Dash time (s)", pch=16)  
#Add the least squares line to the scatterplot:  
> abline(dashline)
```



12.1 Regression Analysis

Thus far, we have learned to explore and describe the *observed* relationship between 2 variables for a *sample* of data. In other words, we have only learned how to describe some data that's sitting in front of us. In the remainder of this chapter we will learn how to use these sample statistics to make inferences about the *true* relationship between 2 variables among an entire population.

Population Regression Model

Let

y = value of the response variable

x = value of the explanatory variable

Notice: The trend of the linear relationship between x and y is described by

$$\mu_x = \alpha + \beta x$$

What is the interpretation of the trend?

μ_x is the expected value of y given a particular value of x .

Model Assumptions:

When using such a model to describe the linear relationship between two quantitative variables, x and y , we are making 3 assumptions.

1. A linear relationship is appropriate.
2. Normality: The distribution of the y values at a given value of x is

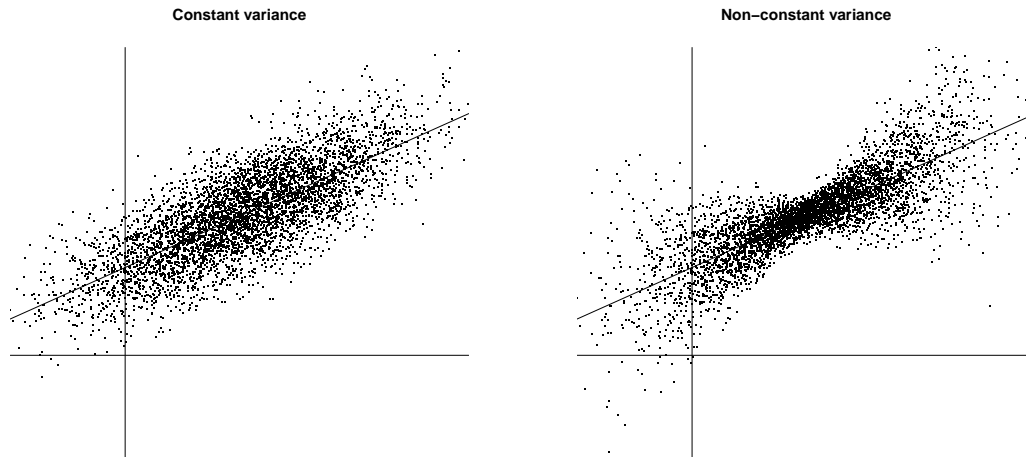
$$N(\mu_x, \sigma) = N(\alpha + \beta x, \sigma)$$

3. Constant variance:

The standard deviation of y , σ , is the same for all values of x .

In pictures:

draw 2 pictures, one with non-constant variance



12.2 Inference About the Population Regression Model

As usual, we can make inferences about the *unknown* true population regression model through

1. point and interval estimation
2. hypothesis testing

12.2.1 Estimating α and β

Recall that the population regression model can be written as

$$y = \alpha + \beta x + \varepsilon .$$

However, *parameters* α and β are unknown. That is, the *true* linear relationship between x and y is unknown.

Goal:

Choose estimates of α and β that will result in small prediction errors.

The Bottom Line:

The least squares line

$$\hat{y} = a + bx$$

serves as a *point estimate* of the true population regression model:

a is an unbiased estimate of α

b is an unbiased estimate of β

EXAMPLE 12.1 CONTINUED

Recall: We are interested in exploring the relationship between box office sales and the cost of production. Which is the response and which is the explanatory variable?

Response: box office sales

Explanatory variable: production cost

Use R to draw a scatterplot and estimate the true linear relationship between box office success and the cost of production:

```
> plot(Prod, Box, xlab="Production Cost", ylab="Box Office Sales")
> myfit <- lm(Box ~ Prod)
> summary(myfit)
```

Coefficients:

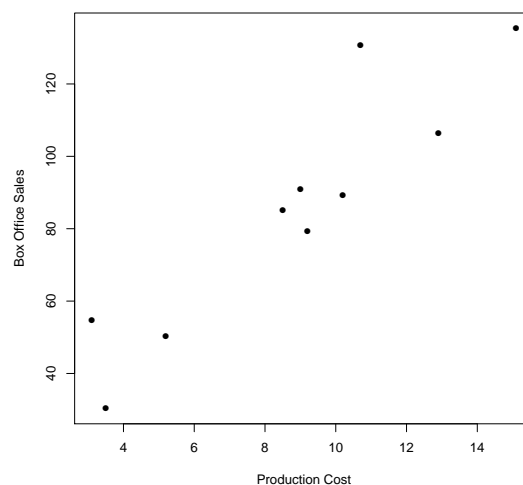
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	15.513	11.603	1.337	0.217989
Prod	7.978	1.223	6.522	0.000184 ***

--

Residual standard error: 14.26 on 8 degrees of freedom

Multiple R-squared: 0.8417, Adjusted R-squared: 0.8219

F-statistic: 42.54 on 1 and 8 DF, p-value: 0.0001838



- (a) What is the estimate of
- α
- ?

$$a = 15.513$$

- (b) State and interpret the estimate of
- β
- .

$$b = 7.978$$

we expect box office sales to increase by almost \$8 million for every one million spent on the movie's production

- (c) Write down the least squares estimate of the population regression line.

$$\widehat{\text{Box}} = a + b(\text{Prod}) = 15.513 + 7.978 (\text{Prod})$$

- (d) Predict the box office sales for a movie with production costs of \$7 million.

$$\widehat{\text{Box}} = a + b(\text{Prod}) = 15.513 + 7.978 (7) = \$71.4$$

12.2.2 Hypothesis Tests and Confidence Intervals for β

Most of the interesting information about the linear relationship between two quantitative variables is captured by the slope of the population regression line, β . Therefore, we will focus on making inferences about β . We can also construct hypothesis tests and confidence intervals for the intercept α in a similar fashion. However, recall from Chapter 3 that it is not always reasonable to interpret α in the context of the problem. In such cases, it is also meaningless to construct and interpret confidence intervals and hypothesis tests for α .

The construction of hypothesis tests and confidence intervals for β will rely on the following information:

b = point estimate of β

standard error of $b = se(b)$

(Calculate in R, the formula is too complicated)

sampling distribution of b :

$$\frac{b - \beta}{se(b)} \sim t_{n-2}$$

where n = sample size

Confidence Intervals for β

Assumptions:

1. Random sample of n pairs of data
2. Model Assumptions:
 - (a) Linear relationship is appropriate
 - (b) Normality
 - (c) Constant variance

$1 - \alpha$ Confidence Level CI for β :

$$\text{point est} \pm \text{margin of error} = b \pm t_{df}^* \cdot se(b)$$

where $df = n - 2$

Hypothesis Tests for β

Assumptions: Same assumptions as for the confidence interval.

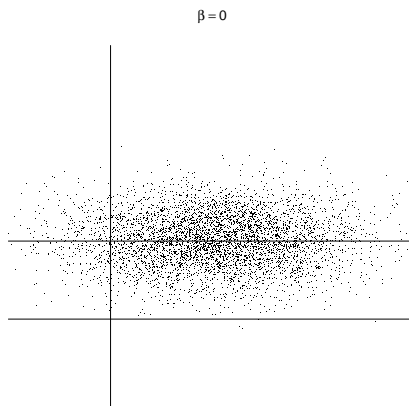
Hypothesis:

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

Interpretation:

Draw a picture. When $\beta = 0$, the outcome of y does not depend on the value of x .



The following are equivalent:

$$\begin{aligned} H_0 : \beta = 0 &\iff H_0 : x \text{ and } y \text{ are independent} \\ &\iff H_0 : \text{we wouldn't want to use } x \text{ to predict } y \end{aligned}$$

Test Statistic:

$$t = \frac{\text{point est} - \text{hyp value}}{\text{se}(\text{point est})} = \frac{b}{\text{se}(b)} \sim t_{n-2} \text{ when } H_0 \text{ true}$$

p-value:

$$p\text{-value} = 2P(t_{n-2} > |t|)$$

Conclusion:

If *p*-value < α , reject H_0

If *p*-value $\geq \alpha$, fail to reject H_0

12.2.3 Measuring the Strength of the Linear Relationship

Rejecting $H_0: \beta = 0$ only tells us that a significant linear association between our two quantitative variables *exists*. It does not give us an idea of how *strong* this association is!

1. Correlation $R =$ **a measure of the strength and direction of the linear association between 2 quantitative variables**

R is unknown. Estimate R using r , the sample correlation.

2. $R^2 =$ **proportion of variation in y that is explained by its linear relationship with x . The remaining variation is captured by or reflected in the scatter**

R^2 is unknown. Estimate R^2 using r^2 .

EXAMPLE 12.1 CONTINUED

Use the R output on page 176 to answer the following questions.

- (a) At the 0.05 level, use the R output to test $H_0: \beta = 0$ versus $H_a: \beta \neq 0$ where β is the slope of the regression line between box office sales and production costs.

Assumptions:

We will learn how to check these assumptions later...

Test statistic:

$$t = \frac{b}{se(b)} = \frac{7.978}{1.223} = 6.522$$

p-value:

$$p\text{-val} = 2P(t_{n-2} > t) = 2P(t_8 > 6.522) < 2(0.001) = 0.002$$

Actually, from R we have an exact p-value: p-val = 0.000184

Conclusion:

p-val < 0.05. Reject H_0 . We have strong evidence that a linear association exists between box office sales and production costs.

- (b) Calculate and interpret the 95% confidence interval for β :

$$\begin{aligned} b \pm t_{\alpha/2, n-2} se(b) &= 7.978 \pm 2.306(1.223) \\ &= 7.978 \pm 2.820 = (5.158, 10.798) \end{aligned}$$

We are 95% confident that Box office sales increase somewhere between \$5.158 and \$10.798 million for \$1 mil increase in production cost.

- (c) What percentage of the variation in box office sales is accounted for by its linear relationship with production costs?

$$\approx 84\% \quad (r^2 = .8417)$$

12.3 Correlation and Regression: A Cautionary Tale

Recall: Correlation and regression are powerful tools for describing associations and making predictions. However, these tools can be abused as well. Keep the following in mind when constructing (and reading) the results of this type of analysis:

1. Correlation and regression only describe *linear* relationships.
2. Correlation and regression are *not* resistant to outliers.
3. Regression lines should not be extrapolated far outside the range of observed data.

Definition: extrapolation

Extrapolation is the use of a regression line to predict values of the response variable for values of the explanatory variable that are far outside the observed range of the data.

Why is this a problem?

We have no assurance that the linear trend continues beyond the observed range of the data.

EXAMPLE 12.2 CONTINUED

We previously fit the following least squares regression line that describes the linear relationship between Olympic winning time of the 200m dash (Time) and the year in which the Olympics took place (Year):

$$\widehat{\text{Time}} = 22.355 - 0.0303 \text{ Year}$$

To convince yourself that extrapolating is the wrong thing to do, use the fitted line to predict the winning dash time in the 2016 and 2640 Olympics.

For the 2016 Olympics:

$$\widehat{\text{Time}} = 22.355 - 0.0303 \text{ Year} = 22.355 - 0.0303(2016 - 1900) = 18.84$$

For the 2640 Olympics:

$$\widehat{\text{Time}} = 22.355 - 0.0303 \text{ Year} = 22.355 - 0.0303(2640 - 1900) = -0.07$$

The winning time cannot be negative!

4. Correlation does not necessarily imply causation

Suppose r is the correlation between two quantitative variables x and y . A value of r close to -1 or 1 suggests these two variables are *associated*. It does **NOT** necessarily mean that a change in x *causes* a change in y (or vice versa)

When there is an association but no causal relationship between two variables, there is often an alternative explanation for the association:

Definition: lurking variable

A *lurking variable* is a variable which is not included in the study, but strongly affects the relationship among the variables of primary interest. It may either falsely suggest a strong relationship or hide an important existing relationship.

EXAMPLE 12.3

Based on United Nations data, it can be shown that there is a strong, negative correlation between a nation's infant mortality rate (IMR) and the per capita television ownership. How can this be explained?

Nobody actually thinks that watching more TV *causes* an improvement in infant health.

Lurking variable: National GDP

People in wealthier nations tend to have healthier babies and own more TV's.

Association does not imply causation!

In general:

The effects of lurking variables cannot be ruled out in observational studies.

How can we establish causation?

1. **Set up a carefully designed experiment that controls for potential lurking variables.**

2. Evaluate criteria for establishing causation:

- (a) **Association is strong.**
- (b) **Association is consistent across many studies.**
- (c) **Alleged cause is plausible.**
- (d) **Alleged cause preceded the effect (in time).**
- (e) **Higher doses evoke stronger responses.**

CHAPTER 13: MULTIPLE REGRESSION

EXAMPLE 13.1

In an experiment at Ohio State University, 16 student volunteers each drank a randomly assigned number of beers and had their blood alcohol content (BAC) measured 30 minutes later.

Gender	Weight	Beers	BAC
female	132	5	0.1
female	128	2	0.03
male	192	8	0.12
.	.	.	.
.	.	.	.

A full data set can be found at <http://www.stat.umn.edu/~wuxxx725/data/BAC.txt>.

An obvious factor in one's BAC is how many beers they've had to drink. However, there are other factors that may influence BAC. For instance, drinking 3 beers may have a stronger affect on a person who weighs 140 pounds than a person who weighs 240 pounds.

In general, there are often several explanatory variables that may be good predictors of a single response variable. In Chapter 12 we learned how to perform regression analysis using one explanatory variable at a time.

GOAL:

simultaneously use multiple explanatory variables to predict a single response variable

Why?

1. **using more information improves prediction**
2. **allows us to analyze the association between 2 variables while controlling for other factors**

13.1 The Multiple Regression Model

Let y denote the response variable and x_1, x_2, \dots, x_k denote k different explanatory variables. Then the multiple population regression model is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

where $\alpha, \beta_1, \beta_2, \dots, \beta_k$ are unknown regression parameters

Interpretation of α :

α is the expected value of y when $x_1 = x_2 = \dots = x_k = 0$

This may not be interpretable in the context of the problem

Interpretation of β_j :

β_j is the expected change in y per unit change in x_j when all other x 's are held constant.

Model Assumptions:

The assumptions about the multiple regression model are similar to those for the regression model with only one explanatory variable:

1. **Linearity:** There is a linear relationship between y and *each* of the explanatory variables.

Check:

a scatterplot matrix displays scatterplots of y versus each of the x 's

2. **Normality:** The distribution of the y values at a given set of (x_1, x_2, \dots, x_k) values is

$$N(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k, \sigma)$$

Check:

normal quantile plot of residuals

3. **Constant variance:** The standard deviation of y , σ , is the same at each set of (x_1, x_2, \dots, x_k) values.

Check:

residual plot of residuals versus fitted values

13.2 Estimation of the Multiple Regression Model

Least Squares Regression

As with simple linear regression with one explanatory variable, we use the least squares procedure to obtain an estimate of the population regression model. The resulting sample prediction equation is

$$\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

EXAMPLE 13.1 CONTINUED

Recall: We want to explore how one's BAC is affected by their weight and the number of beers they drink. Let the true model be represented as

$$\text{BAC} = \alpha + \beta_1 \text{ Beers} + \beta_2 \text{ Weight} + \varepsilon$$

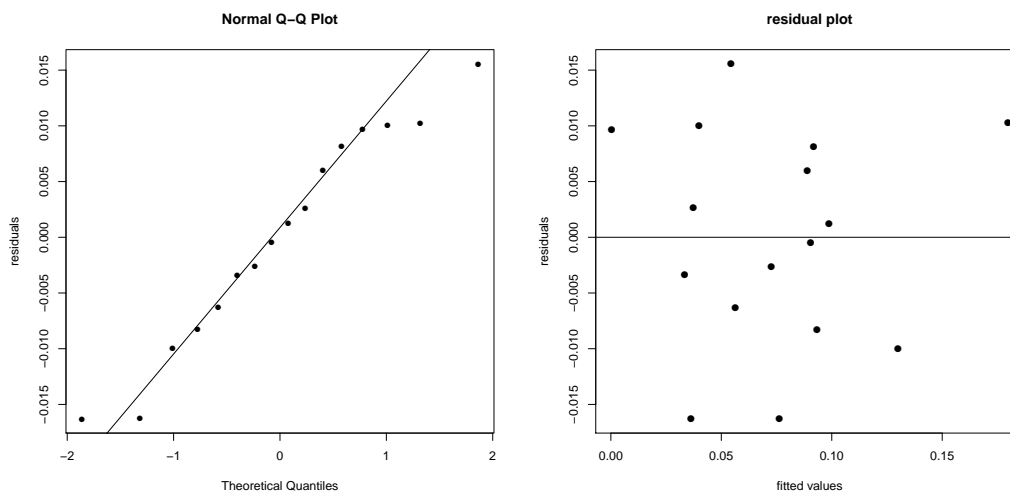
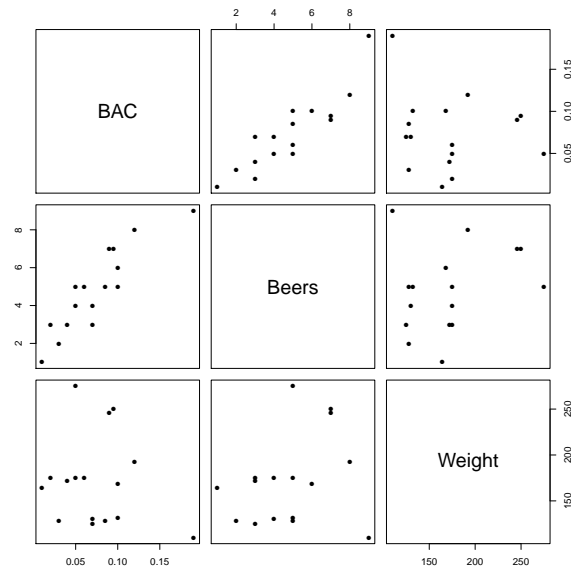
Use R to fit a least squares line for this model:

```
> myfit <- lm(BAC ~ Beers + Weight)
> summary(myfit)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.986e-02  1.043e-02   3.821  0.00212 **
Beers        1.998e-02  1.263e-03  15.817  7.16e-10 ***
Weight      -3.628e-04  5.668e-05  -6.401  2.34e-05 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.01041 on 13 degrees of freedom
Multiple R-squared:  0.9518, Adjusted R-squared:  0.9444
F-statistic: 128.3 on 2 and 13 DF,  p-value: 2.756e-09
```

(a) Use the following plots to check the model assumptions.

```
> pairs(cbind(BAC, Beers, Weight), pch=16)           #scatterplot matrix
> qqnorm(myfit$residuals, ylab="residuals", pch=16) #normal quantile plot
> qqline(myfit$residuals)
#Residual plot:
> plot(myfit$fitted, myfit$residuals, xlab="fitted values",
>       ylab="residuals", pch=16)
> abline(h=0)
```



(b) Write down the equation of the predicted regression line.

$$\widehat{\text{BAC}} = a + b_1 \text{Beers} + b_2 \text{Weight} = .0399 + 0.02 \text{Beers} - 0.00036 \text{Weight}$$

(c) Predict the BAC for a 200 pound person after drinking 3 beers.

$$\widehat{\text{BAC}} = .0399 + 0.02(3) - 0.00036(200) = 0.0279$$

- (d) State and interpret the estimated value of β_1 , the parameter associated with ‘Beers’.

$$b_1 = 0.02$$

At any fixed weight, BAC increases by 0.02 percent on average for every one beer drank.

- (e) State and interpret the estimated value of β_2 , the parameter associated with ‘Weight’.

$$b_2 = -0.00036$$

Fix the number of beers one drinks. Then for every 10 pound increase in a person’s weight, BAC decreases by $10(0.00036)=0.0036$ percent on average.

13.3 Inference for the Multiple Regression Model

We can extend the inferential techniques we learned for the simple regression model to multiple regression models.

r^2 : Measuring the Strength of the Multiple Regression Relationship

As in simple linear regression with a single explanatory variable, we can use r^2 to measure the strength of the linear association between y and x_1, x_2, \dots, x_k .

Interpretation:

$R^2 =$ proportion of the variation in y accounted for by its linear relationship with x_1, x_2, \dots, x_k .

Properties of r^2 , the sample estimate of R^2 :

- r^2 is always between 0 and 1, and equals the square of the sample correlation between \hat{y} and y .**

$$r^2 = \left(\frac{1}{n-1} \sum \left(\frac{\hat{y}_i - \bar{y}}{s_{\hat{y}}} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right)^2$$

2. r^2 will almost always increase (and will never decrease) when *any* new explanatory variable is added to the model, even if that variable is not helpful for predicting y in the population.

Example:

```
> set.seed(123456)
> random.num<-rnorm(16)
> overfit<-lm(BAC~Beers+Weight+random.num)
> summary(overfit)
```

Call:

```
lm(formula = BAC ~ Beers + Weight + random.num)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.093e-02	1.060e-02	3.862	0.00226	**
Beers	2.074e-02	1.546e-03	13.420	1.38e-08	***
Weight	-3.857e-04	6.285e-05	-6.136	5.05e-05	***
random.num	-3.854e-03	4.396e-03	-0.877	0.39784	

--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0105 on 12 degrees of freedom
Multiple R-squared: 0.9547, Adjusted R-squared: 0.9434
F-statistic: 84.29 on 3 and 12 DF, p-value: 2.486e-08

EXAMPLE 13.1 CONTINUED

- (a) Use R to estimate the correlation between BAC and Beers as well as between BAC and Weight:

```
> cor(Beers,BAC)
[1] 0.8943381
> cor(Weight,BAC)
[1] -0.1549634
```

Use the above output to estimate...

- how much of the variability in BAC can be accounted for by its linear relationship with 'Beers' alone (when a person's weight is unknown).

$$r^2 = .894^2 = .799 \text{ (About 80\%)}$$

- how much of the variability in BAC can be accounted for by its linear relationship with 'Weight' alone (when the number of beer a person has had is unknown).

$$r^2 = (-.155)^2 = .024 \text{ (not much at all!)}$$

- (b) Use the output on page 186 to state and interpret the estimated value of R^2 for the linear regression of BAC using predictors 'Beers' and 'Weight'.

.9518

The F -test

Before testing for the existence of a significant effect of the individual explanatory variables, we should perform an F -test for the existence of a significant *collective* effect of the explanatory variables on the response variable y .

Assumptions:

1. Random sample of n sets of $(y, x_1, x_2, \dots, x_k)$ data
2. Model Assumptions:
 - (a) Linear relationships are appropriate between y and each x_j
 - (b) Normality
 - (c) Constant variance

Hypothesis:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

H_a : At least one β parameter is not equal to 0

Interpretation:

H_0 : y is independent of all the x 's

H_a : at least one of the x 's is linearly associated with y

Test Statistic:

$$F = \frac{\text{Mean Square for Regression}}{\text{Mean Square Error}}$$

p-value:

$$\text{p-value} = P(F_{df_1, df_2} > F)$$

Conclusion:

If p-value $< \alpha$, reject H_0

If p-value $\geq \alpha$, fail to reject H_0

- If we fail to reject H_0 of the F test:

Stop. None of the x 's are good predictors of y .

- If we reject H_0 of the F test and conclude that y is significantly associated with at least one of the explanatory variables:

Investigate further. Run individual t-tests for each of the x 's to determine which are significantly associated with y .

Hypothesis Tests for β_j

Assumptions: Same assumptions as for the F test.

Hypothesis:

$$H_0: \beta_j = 0$$

$$H_a: \beta_j \neq 0$$

Interpretation:

H_0 : When all other x 's are known, x_j does not add a significant amount of predictive information about y .

(If we have the other x 's, we don't need x_j .)

Test Statistic:

$$t = \frac{b_j}{se(b_j)} \sim t_{n-(k+1)} \quad \text{when } H_0 \text{ true}$$

p-value:

$$\text{p-value} = 2P(t_{n-(k+1)} > |t|)$$

Conclusion:

If p-value $< \alpha$, reject H_0

If p-value $\geq \alpha$, fail to reject H_0

Confidence Intervals for β_j

Under the same assumptions as the F test, the $1 - \alpha$ confidence level CI for β_j is

$$b_j \pm t_{\alpha/2, df} se(b_j)$$

where $df = n - (\# \text{ of parameters in the mean function}) = n - (k + 1)$

EXAMPLE 13.1 CONTINUED

Perform inference for the multiple regression model for BAC and the explanatory variables ‘Weight’ and ‘Beers’:

$$\text{BAC} = \alpha + \beta_1 \text{ Beers} + \beta_2 \text{ Weight} + \varepsilon$$

- (a) Use the R output from page 186 to perform the F test at the 0.05 level.

Assumptions:

We would need to assume that the data represent a random sample from the population. From the plots on page 187, the linearity, normality, and constant variance assumptions are approximately satisfied.

Hypotheses:

$$H_0: \beta_1 = \beta_2 = 0$$

H_a : At least one of the β 's doesn't equal 0.

Test Statistic:

$$F = 128.3$$

p-value:

$$\text{p-val} = P(F_{2,13} \geq 128.3) = 2.756 \times 10^{-9}$$

Conclusion:

p-value < 0.05: Reject H_0 . We conclude that either the number of beers one drinks or one's weight (or both) has a significant effect on BAC.

- (b) If necessary, perform individual t-tests for the explanatory variables 'Weight' and 'Beers' at the 0.01 level.

As we have rejected H_0 in the F test, it is necessary to perform the individual t-tests.

1. Assumptions:

Same as that of the F test.

Hypotheses:

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

Test statistic:

$$t = 15.817$$

p-value:

$$\text{p-val} = 7.16 \times 10^{-10}$$

Conclusion:

p-value < 0.05 : Reject H_0 . Conclude that even if a person's weight is known, the number of beers one drinks adds a significant amount of predictive information about BAC.

2. Assumptions:

Same as that of the F test.

Hypotheses:

$$H_0: \beta_2 = 0$$

$$H_a: \beta_2 \neq 0$$

Test statistic:

$$t = -6.401$$

p-value:

$$\text{p-val} = 2.34 \times 10^{-5}$$

Conclusion:

p-value < 0.05: Reject H_0 . Conclude that even if we know how much beer a person has had, knowing his/her weight adds a significant amount of predictive information about BAC.

(c) Compute and interpret a 90% confidence interval for β_2 .

$$\text{df} = n - (k + 1) = 16 - (2 + 1) = 13$$

$$\begin{aligned} b_2 \pm t_{\alpha/2, df} se(b_2) &= -0.00036 \pm 1.771(0.000057) \\ &= (-0.00046, -0.00026) \end{aligned}$$

We are 90% confident that when the amount of beer one drinks is held constant, BAC decreases between 0.00026 and 0.00046 on average for every one pound increase in weight.

EXAMPLE 13.2: AN EXAMPLE OF CONFOUNDING

The file <http://www.stat.umn.edu/~wuxxx725/data/sat.txt> contains the following state-level data from 1982:

state: Name of the state

sat: Average SAT score for the state (response variable)

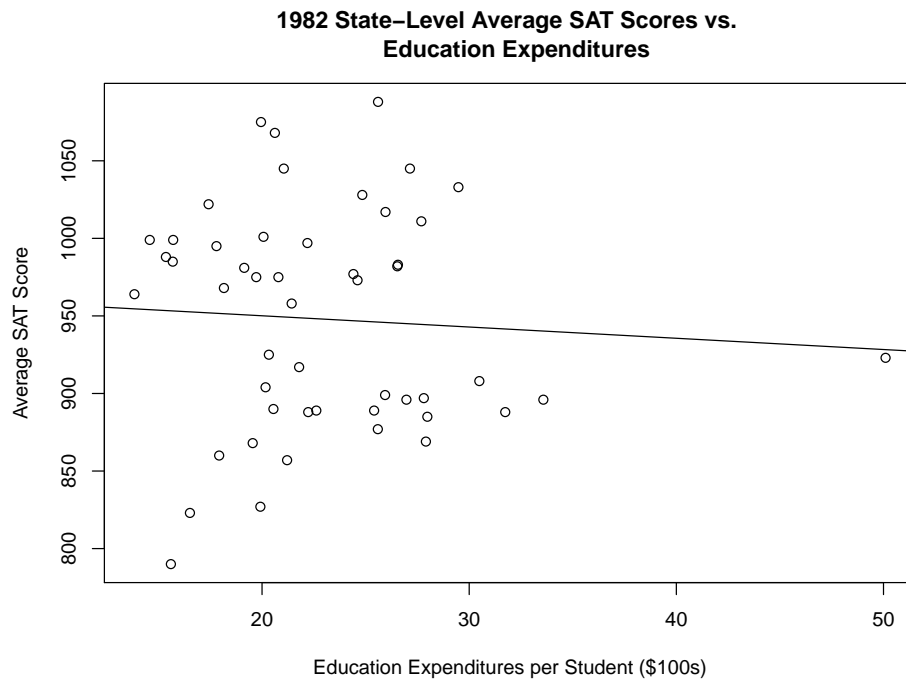
takers: Percentage of total eligible students that took the SAT

expend: money spent on education, per student (in \$100s)

For example, here is the record for Minnesota, where the average score was 1028, 7% of all eligible students took the test, and the amount spent on education per student was \$2,484.

```
> sat.data[state=="Minnesota", ]
      state  sat takers expend
7 Minnesota 1028      7  24.84
```

First we consider the relationship between SAT scores and education expenditures alone. The scatterplot of these two variables with the least squares regression line looks like this:



Note: The outlier you see with the high value of education expenditures is Alaska. However, the correlation and regression equations are not largely influenced by this outlier.

- (a) Use the scatterplot to describe the association between average SAT score and education expenditures in words. What policy argument could be supported by this scatterplot and the fact that the slope of the least squares line is negative?

The mean SAT score decreases as the per capita expenditure on secondary schools increases.

This could support the argument that the government should reduce its expenditure on secondary schools.

Multiple Regression with `expend` and `takers`

Now we consider the multiple linear regression with average SAT score (`sat`) as the response variable and education expenditures (`expend`) and percentage of eligible students taking the test (`takers`) as explanatory variables.

It turns out that the states with the highest average SAT scores tend to be the states in which relatively few students take the SAT. For example, compare the following sample of midwestern states with a sample of states on the East Coast:

```
> sat.data[state=="Iowa" | state=="Nebraska" | state=="Minnesota",]
      state  sat takers expend
1      Iowa 1088      3  25.60
5 Nebraska 1045      5  21.05
7 Minnesota 1028      7  24.84
> sat.data[state=="Delaware" | state=="NewYork" | state=="NorthCarolina",]
      state  sat takers expend
34 Delaware  897      42  27.81
36 NewYork   896      59  33.58
48 NorthCarolina 827      47  19.92
> cor(sat,takers)
[1] -0.85781
```

- (b) Why do you think `sat` has such a strong negative association with `takers`, the percentage of students taking the test?

The states with low percentages of high school seniors taking SAT are the ones which dominantly use the ACT scores.

In these ACT-dominated states, the high school seniors who take SAT mostly intend to apply for out-of-state colleges, and therefore they might actually represent a sample from the sub-population of high school seniors who are competitive in SAT.

As a result, for these states, the observed mean SAT score might be considerably higher than the true population mean SAT score of all high school seniors.

Now consider the following multiple linear regression model:

$$\text{sat} = \alpha + \beta_1(\text{expend}) + \beta_2(\text{takers}) + \varepsilon$$

Here is the R output for the least squares estimate of this model:

```
> summary(multiple.reg)

Call:
lm(formula = sat ~ expend + takers)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  973.0426    19.1239   50.881 < 2e-16 ***
expend         2.2624     0.8389    2.697  0.00969 **
takers        -2.9390     0.2341  -12.556 < 2e-16 ***
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.6 on 47 degrees of freedom
Multiple R-squared:  0.7712,    Adjusted R-squared:  0.7615
F-statistic: 79.23 on 2 and 47 DF,  p-value: 8.821e-16
```

- (c) Find and interpret the residual for Minnesota, which in 1982 had an average SAT score of 1028, with education expenditures of \$2,484 per student ($\text{expend}=24.84$) and 7 percent of eligible students taking the SAT.

$$\begin{aligned} \text{residual} &= y - \hat{y} \\ &= 1028 - (973.04 + 2.2624(24.84) - 2.9390(7)) \\ &= 1028 - 1008.67 \\ &= 19.33. \end{aligned}$$

The actual mean SAT score for Minnesota is 19.33 points higher than the value predicted by the linear model based on the expenditure and the percentage of eligible students taking SAT.

- (d) Calculate and interpret a 95% confidence interval for β_1 , the coefficient associated with the `expend` variable in this model. For your interpretation, remember that the `expend` variable is measured in units of \$100's of dollars per student.

The following result from R will be helpful:

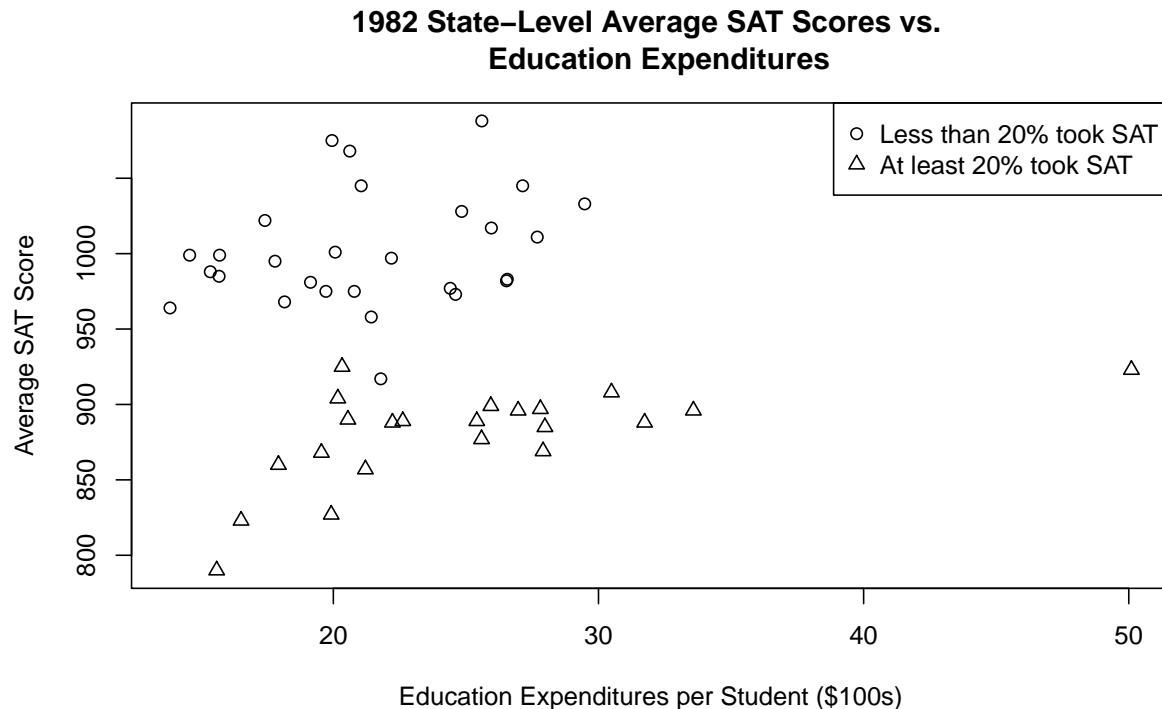
```
> qt(0.975,df=47)
[1] 2.011741
```

$$\begin{aligned} 95\% \text{ CI for } \beta_1 &= \hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-p-1} \cdot \text{se}(\hat{\beta}_1) \\ &= 2.2624 - (2.0117)(0.8389) \\ &= 2.2624 \pm 1.6877 \\ &= (0.5748, 3.9500). \end{aligned}$$

We are 95% confident that when all other conditions are fixed, for each additional \$100 per student a state spent on secondary schools, the state mean SAT score increases between 0.57 and 3.95 points.

It is interesting that our confidence interval suggests that β_1 is almost certainly positive, even though we saw earlier that education expenditures have a *negative* correlation with average SAT scores.

In the plot below, imagine fitting lines through the circle points and the triangle points separately. These lines roughly represent the relationship between SAT scores and education expenditures while adjusting for the percentage of students taking the test.



This apparent paradox occurs because **takers** is a *confounding variable* for the relationship between education expenditures and SAT scores.

Definition: confounding variable

A *confounding variable* in an analysis is an explanatory variable associated with both the response and with one or more other explanatory variables of interest.

In our example, states with a higher percentage of students taking the test have lower SAT scores (for obvious reasons), and they also tend to spend more on education (perhaps due to geographic differences). Once we adjust for this confounding variable, we see that education expenditures are *positively* associated with higher SAT scores when we control for the percentage of takers.

What is the difference between a lurking variable and a confounding variable?

- **A confounding variable is a variable included in the study which is associated with both the explanatory and response variables, and therefore would affect the relationship between the explanatory and response variables.**

- A lurking variable is a variable which is not included in the study but could potentially be confounding if it were taken into account.

Including confounding variables in a multiple regression is a *good* thing, because the model can then describe the relationship between a response (here, SAT scores) and an explanatory variable of interest (here, education expenditures) while imagining that the confounding variable (here, percentage of students taking the test) is held fixed.

So, now that we have accounted for this confounding variable (`takers`), does our inference that β_1 is positive mean that higher education expenditures *cause* higher SAT scores?

Since this study is observational, we cannot definitely establish the causal relationship between education expenditure and SAT score from it. However, given that the current model has taken the confounding variable `takers` into account, it is more likely to reflect the true relationship between the variables `expend` and `sat` than the model without `takers`.

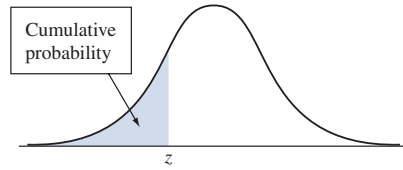
In fact, when Powell and Steelman published this study in 1984, they have considered many other confounding variables: sex composition; racial composition; median income; average number of years studying social sciences, natural sciences and humanities; percentage of students attending public schools; and southern vs. non-southern states. Among these confounding variables, only the effects of `takers`, sex composition, and racial composition are significant. After adjusting for the potential confounding variables, the effect of expenditure remains significant.

With more confounding variables adjusted in the studies and more studies on different populations showing consistent associations, it is less likely that the association between the explanatory variables and response is due to some other lurking variables rather than a causation. In this case, we would conclude that it is plausible, albeit not definitely, that higher expenditure in education causes higher SAT scores.

NOTATION

Sample Statistics		Population Parameters	
n	sample size	p	population proportion
\hat{p}	sample proportion	σ	population standard deviation
s	sample standard deviation	σ^2	population variance
s^2	sample variance	μ	population mean
\bar{x}	sample mean	R	population correlation
r	sample correlation		
IQR	interquartile range		
M	sample median		
Q1	first quartile		
Q3	third quartile		
LSL	least squares (regression) line		

TABLE A

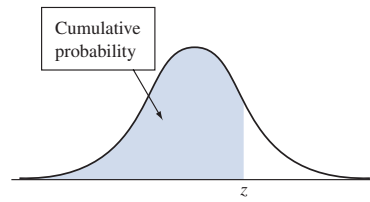


Cumulative probability for z is the area under the standard normal curve to the left of z

Table A Standard Normal Cumulative Probabilities

z	.00
-5.0	.000000287
-4.5	.00000340
-4.0	.0000317
-3.5	.000233

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641



Cumulative probability for z is the area under the standard normal curve to the left of z

Table A Standard Normal Cumulative Probabilities (*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

z	.00
3.5	.999767
4.0	.9999683
4.5	.9999966
5.0	.99999713

ADDITIONAL NOTES

