Assumptions

Gary W. Oehlert

School of Statistics
University of Minnesota

February 7, 2016

Background

Our inference tools make assumptions. We assume that

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where the ϵ_{ij}s are independent with distribution $N(0, \sigma^2)$.[1]

If the model is correct, our inference is good and matches with randomization inference.

Unfortunately, wishing doesn't make it so.

[1] Equivalently, we can say that y_{ij} follows $N(\mu_i, \sigma^2)$.

If the assumptions are not true, our inferences might not be valid, for example,

- A confidence interval might not cover with the stated error rate.
- A test with nominal type I error of $\mathcal{E}$ could actually have a larger or smaller type I error rate.

This is obviously bad news and can be the source of controversy and disagreement over how the analysis was done and the validity of the results.

(But if you did a randomization, your randomization inference is still valid.)

Some procedures work reasonably well (e.g., actual interval coverage rate is near to nominal, or actual p-value is close to nominal p-value) even when some assumptions are violated.

This is called robustness of validity.

Generally these procedures work better when violations are mild and work less well as violations become more extreme.

A procedure that has robustness of validity can be inefficient, so we might not want to use it even if it is robust.

The basic assumptions are

- Independence (most important)
- Constant variance
- Normality (least important)

Many ways that data can fail to be independent; we will learn to check for one.

In this course we will not generally try to fix or accommodate dependence. We leave that for other courses (e.g., time series, multivariate analysis, etc.).

Residuals

To make matters interesting, our assumptions are about the ϵ_{ij}, but we never get to see them. They are unobservable, so we must guide our analysis using something else.

What we do have are residuals.

The basic raw residual is

$$r_{ij} = y_{ij} - \text{fitted value}$$

In our simple models to date that is

$$r_{ij} = y_{ij} - (\widehat{\mu} + \widehat{\alpha}_i) = y_{ij} - \overline{y}_{i\bullet}$$

The raw residual is useful for many purposes, and is often good enough in balanced designed experiments. However, we can do better.

The standardized residual (sometimes called internally Studentized) adjusts r_{ij} for its estimated standard deviation:

$$s_{ij} = \frac{r_{ij}}{\sqrt{MS_E(1 - H_{ij})}}$$

The H_{ij} value is called the leverage; it is a diagonal element of the "Hat" matrix, which is why we call it H. Use `hatvalues()` in R.

Roughly speaking, the s_{ij} should look like standard normals, particularly in large samples.

One further step is the Studentized residual (or the externally Studentized residual if you like calling standardized by internally Studentized):

$$t_{ij} = s_{ij}\sqrt{\frac{\nu - 1}{\nu - s_{ij}^2}}$$

where ν is the df in the MS_E.

If model is correct, t_{ij} follows a t distribution with $\nu - 1$ df. A t with reasonable df will look pretty much like a normal.

Studentized residuals are especially good in looking for outliers.

Think of adding a dummy variable to a model that is 1 for point i,j and 0 otherwise. The t-test for the coefficient of that dummy variable is the Studentized residual in the original model.

Studentized residuals say how well the data value fits the model estimated from the rest of the data.

Assessing assumptions

I don't like to test for normality or constant variance etc.:

- With small sample sizes, you'll never be able to reject the null that there are no problems.
- With large sample sizes, you'll constantly detect little problems that have no practical effect.

It's really all shades of gray (at least 50), and we would like to know where we are on the scale from mild issues to severe issues.

So assess assumptions qualitatively; don't just rely on a test.

Residual plots

Our principal tools for assessing assumptions are various plots of residuals:

- Normal probability plot
- Residuals versus predicted plot
- Residuals in time order

The first two are the basic plots for assessing normality and constant variance; the last one is just one of many potential plots for assessing independence.

The NPP plots the residual against its corresponding normal score. The smallest residual plots against the smallest normal score for a sample of N; the second smallest residual against the second smallest normal score, and so on.

Normal scores depend on N. Think about an independent sample of N standard normals. They all have mean 0, but if you just consider the smallest one, it has a negative expectation. That expectation is its normal score.

The *rankit* approximates the normal score:

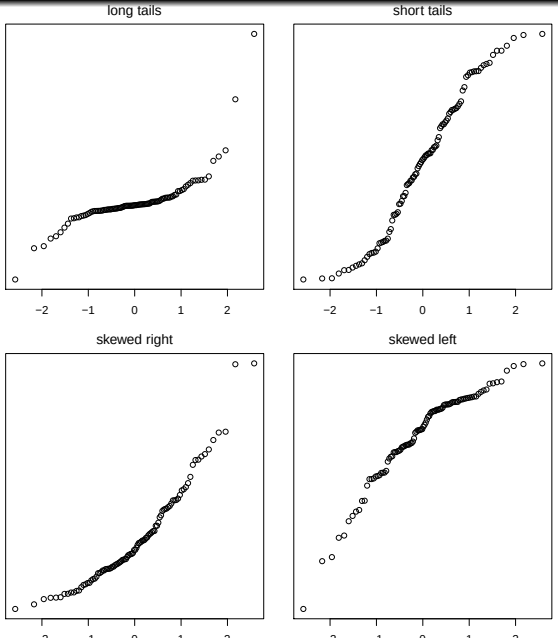
$$\text{rankit}_{i,N} = \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right)$$

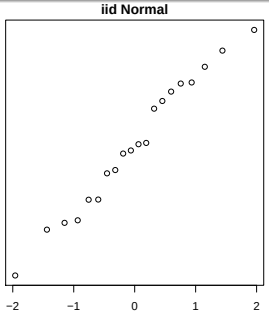
where Φ^{-1} gives normal percent points.

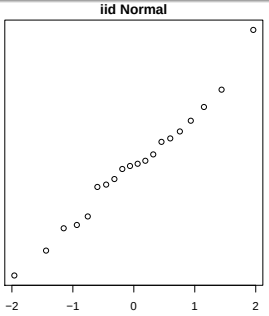
It's probably best to use the Studentized residuals, but the others also work fine in most situations.

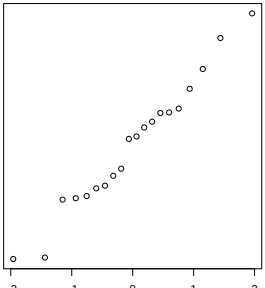
Normally distributed data (and, we hope, residuals from iid normally distributed errors) should have a roughly linear shape, although even normal data can look crooked in small samples.

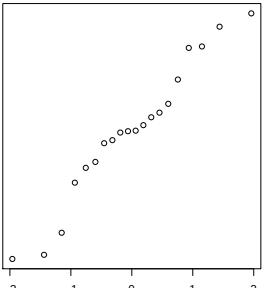
You can tell the shape of the data from the shape of the plot, but you need to practice (and you will).


long tails
short tails
skewed right
skewed left
−2
−1
0
1
2


iid Normal
−2
−1
0
1
2


iid Normal
−2
−1
0
1
2


iid Normal


iid Normal

You can also test for outliers using the Studentized residuals (the t-residuals).

These are a one-at-a-time test. Look at the largest absolute t-residual and then do the test by making a Bonferroni adjustment (i.e., multiply p-value by N, if it still looks small, then you have an outlier).

You can do this sequentially, but the test is only exact for the first one.

The diagnostic plot for non-constant variance is to plot each residual against its corresponding predicted/fitted value.

We are hoping to see no pattern in the vertical dispersion.

The most common problem occurs when larger means go with larger variances. In this case we see a "right opening megaphone."

We sometimes see the reverse, particularly when there is an upper bound on the response.

There are several variations on this, including box plots of residuals and plots of square root absolute residuals against fitted values.

If you have to/want to test for equality of variances, your best bet is Levene's test. This makes a new response as the absolute value of the deviations of the original data from the predicted value, and then does an ANOVA test for the separate means model on the absolute deviations.

There are several variations on this where you might take absolute deviations from the median of each group, or the absolute deviations to some power, etc.

There are several classical tests of equality of variance including Barlett's test and Hartley's test; **avoid them like the plague!** They are incredibly hyper-sensitive to normality.

Back to the resin example in R.

There are many ways that data could fail to be independent, but we will only talk about the simplest of these: temporal dependence.

In some data sets, but not all data sets, there is a time order of some kind.

One common failure of independence is when data close in time tend to have similar ϵ_{ij}s and thus similar residuals. This is called positive temporal dependence or positive serial correlation.

The reverse can also happen (near in time tend to be unusually far apart), but it is much more rare.

The simplest diagnostic is to plot the residuals in time order and look for patterns.

Do the residuals seem to be high and low together in patches? That is positive serial correlation.

Do the residuals seem to bounce up and down very roughly and alternately? That could be negative serial correlation.

The stronger the pattern, the stronger the correlation and the greater the problem it will cause with inference.

There are a couple of simple tests for serial correlation. Let r_i be one of the kinds of residuals sorted into time order.

The Durbin-Watson statistic is

$$DW = \frac{\sum_{i=1}^{n-1}(r_i - r_{i+1})^2}{\sum_{i=1}^{n} r_i^2}$$

Independent data tend to have DW around 2; positive correlation makes DW smaller; negative correlation makes DW bigger.

If DW gets as low as 1.5 or as high as 2.5, it's definitely time to start worrying about what is happening to the inference.

There are also a whole variety of "runs" tests, variously defined. These look for things like runs of residuals that are positive (or negative), or runs of data that are increasing (or decreasing).

In any event, there are several runs tests, but they can also be used to assess temporal correlation.

Only assess temporal correlation if your data have a time order!

Accommodating problems

There are two basic approaches to dealing with things when assumptions are not met:

- Alternate methods
- Massaging the data

Developing alternate methods is basically a full-employment act for academic statisticians. The problem is that there are so many things we want to do with our standard approaches, that developing alternatives is also difficult and very time consuming (and life would be really difficult for the non-academics).

I'll mention a few broad areas, but only talk about a couple alternatives.

Robustness is a philosophy and class of techniques that deal with long-tailed, outlier prone data.

Generalized Linear Models (GLM) is a class of techniques for using models with linear predictors but which have non-normal data including count data and various kinds of non-constant variance.

Time series is a class of statistical models for working with serial correlation (among other things).

Spatial statistics includes, among other things, the ability to fit linear models when the data are correlated in space.

Direct replacements are usually developed to solve specific narrow issues without building a whole new class of statistical models.

Many of you are familiar with the version of the t-test that does not use a pooled estimate of variance. Instead, it uses

$$t = \frac{\overline{y}_{i\bullet} - \overline{y}_{j\bullet}}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}}$$

where s_i^2 and s_j^2 are the sample variances in two groups. There is a formula for approximate df, and then you compare with a t-distribution.

This is the direct replacement for ANOVA when g=2 and there is non-constant variance.

In this case, the replacement is so easy and works so well that there is little reason not to use it all the time.

The Brown-Forsythe method generalizes this to $g > 2$ groups, but even this simple problem is getting a bit messy. Let

$$d_i = s_i^2(1 - n_i/N)$$

Then the Brown Forsythe F is

$$BF = \frac{SS_{Trt}}{\sum_{i=1}^{g} d_i}$$

Treat this as F with g-1 and ν df where

$$\nu = \frac{\sum_{i=1}^{g} d_i^2}{\sum_{i=1}^{g} d_i^2/(n_i - 1)}$$

Massaging the data

This sounds like iniquity, but it's really not that bad.

The simplest form of this practice is removing outliers and reanalyzing the data. Ideally, we would like to get the same basic inference with and without the outliers.

If the inference changes substantially, this means that it is dependent on just a handful of the data.

You can't automatically reject a data value simply because it does not fit the model you assume.

Our go-to approach is usually to transform the data, that is, to re-express the data on another scale. Thus we might use

- pH instead of hydrogen ion concentration (log transformation);
- diameter of a bacterial colony rather than area (square root transformation);
- time to distance instead of rate of advance (reciprocal transformation).

In general, any monotone transformation will work, but we concentrate on power family transformations.

Power family transformations work for positive data. If you have some zeroes or negatives, you must first add a constant to all data.[2]

So

$$y_{ij} \to y_{ij}^{\lambda}$$

Use a log transformation instead where $\lambda = 0$ would go.

A lower power tends to reduce right-skewness and reduce increasing variance.

A higher power tends to reduce left-skewness and reduce decreasing variance.

[2] This actual produces a more general transformation, because it has two parameters, the power and the addend, and you can change either.

Note: if the data only range over a factor of 2 or 3, then power transformations are of limited utility. As the ratio of largest to smallest increases, power transformations can have more effect.

Serendipity. More often than we have any right to expect, transformations that make variance more constant also improve normality.

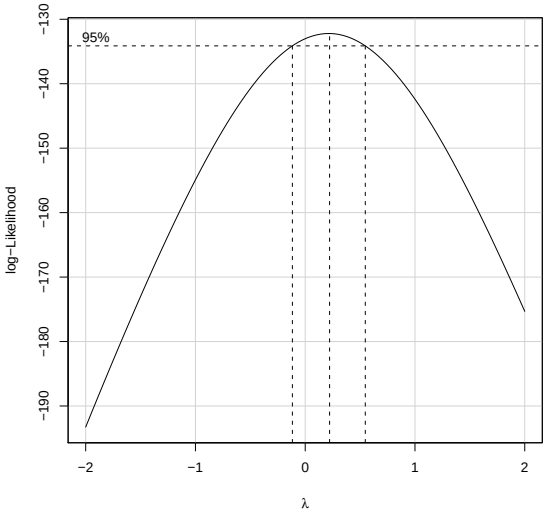
However, if I have to choose between the two, I generally go for more constant variance at the cost of worse normality.

The Box-Cox procedure helps us

- Pick a reasonable range of transformation powers
- Decide whether we need a transformation

I try not to be a slave to the Box-Cox test, and I also try to pick a transformation power that both fixes the problems and is also interpretable. But it is still a very useful guide.

In R, Box-Cox gives us a likelihood profile for λ as well as a 95% confidence interval.



R examples.

Inference

If the null is that distributions in different treatments are the same on one scale, they will also be the same on some other scale.

We might as well use the one where are assumptions are plausible.

We can test equality of means on any scale and get proper inference.

That's the good news . . .

The bad news shows up when you want to make inference on means across scales.

Means do not transform cleanly across power transformations.

That is, you cannot exponentiate the mean of the log data to get the mean of the natural scale data.

A transformed CI for the mean of normal data is a CI for the median on the transformed scale, *not* for the mean.

Land's method helps in the specific case of logs and anti-logs, but in general you either make due with medians or work on the original scale and take you lumps on the quality of inference.

Consequences

So how bad is this, really?

Skewness measures how asymmetric a distribution is.
Kurtosis measures how long-tailed (outlier prone) a distribution is.
The normal has both 0 skewness and 0 kurtosis.

Absent outliers, F-test is only slightly affected by non-normality.

F-test has reasonable robustness of validity, but it is not resistant; individual outliers can change test results.

Often check to see if inference is consistent with and without outliers.

For balanced data (all sample sizes equal),

- Skewness has little effect
- Long tails (positive kurtosis) leads to conservative tests. These tests have nominal p-values larger than they really should be, so fewer rejections than we should have.
- Short tails (negative kurtosis) leads to liberal tests. These tests have nominal p-values smaller than they really should be, so more rejections than we should have.

Table 6.5 gives some numerical results.

Inconsistent results for unbalanced data.

Smaller effects in larger data sets.

Skewness and kurtosis for selected distributions

Distribution	γ_1	γ_2
Normal	0	0
Uniform	0	-1.2
Normal truncated at		
± 1	0	-1.06
± 2	0	-0.63
Student's t (df)		
5	0	6
6	0	3
8	0	1.5
20	0	.38
Chi-square (df)		
1	2.83	12
2	2	6
4	1.41	3
8	1	1.5

Actual Type I error rates for ANOVA F-test with nominal 5% error rate for various sample sizes and values of γ_1 and γ_2 using the methods of Gayen (1950).

Four Samples of Size 5

	γ_2						
γ_1	-1	-.5	0	.5	1	1.5	2
0	.0527	.0514	.0500	.0486	.0473	.0459	.0446
.5	.0530	.0516	.0503	.0489	.0476	.0462	.0448
1	.0538	.0524	.0511	.0497	.0484	.0470	.0457
1.5	.0552	.0538	.0525	.0511	.0497	.0484	.0470

$\gamma_1 = 0$ and $\gamma_2 = 1.5$

4 groups of k		k groups of 5		(k_1, k_1, k_2, k_2)	
k	Error	k	Error	k_1, k_2	Error
2	.0427	4	.0459	10,10	.0480
10	.0480	8	.0474	8,12	.0483
20	.0490	16	.0485	5,15	.0500
40	.0495	32	.0492	2,18	.0588

Skewness can really mess up one-sided confidence intervals. I mean bad.

Two-sided intervals are less affected by skewness, but the coverage errors may pile up on one side.

Pairwise comparisons with balanced data are generally doing well (the differencing tends to cancel the skewness).

Non-constant variance can have serious effects, although the effects are smaller for balanced designs.

If big n_is go with big σ_i^2s, you get a conservative test that does not reject often enough. (The big variances are "over represented" in our standard MSE.)

If big n_is go with small σ_i^2s, you get a liberal test that rejects too often. (The small variances are "over represented" in our standard MSE.)

Table 6.6 shows some examples of how bad things can get with non-constant variance.

For the settings in that table, nominal 5% tests are actually somewhere between 3% and 20%.

More data does not fix the problem.

For pairwise comparisons, some will be liberal and others will be conservative.

g	σ_i^2	n_i	$\mathcal{E}$
3	1, 1, 1	5, 5, 5	.05
	1, 2, 3	5, 5, 5	.0579
	1, 2, 5	5, 5, 5	.0685
	1, 2, 10	5, 5, 5	.0864
	1, 1, 10	5, 5, 5	.0954
	1, 1, 10	50, 50, 50	.0748
3	1, 2, 5	2, 5, 8	.0202
	1, 2, 5	8, 5, 2	.1833
	1, 2, 10	2, 5, 8	.0178
	1, 2, 10	8, 5, 2	.2831
	1, 2, 10	20, 50, 80	.0116
	1, 2, 10	80, 50, 20	.2384
5	1, 2, 2, 2, 5	5, 5, 5, 5, 5	.0682
	1, 2, 2, 2, 5	2, 2, 5, 8, 8	.0292
	1, 2, 2, 2, 5	8, 8, 5, 2, 2	.1453
	1, 1, 1, 1, 5	5, 5, 5, 5, 5	.0908
	1, 1, 1, 1, 5	2, 2, 5, 8, 8	.0347
	1, 1, 1, 1, 5	8, 8, 5, 2, 2	.2029

Outcomes with dependent data depend extremely delicately on the exact nature of the dependence and the exact nature of the contrast or test.

For example, if data are sequential in time with neighboring ϵs correlated .4, then a nominal 95% confidence interval could have coverage 86% or 99.9% depending on whether the treatments were done in blocks or alternately.

More data does not help.

Randomization does help. If you had randomized the order of the treatments, then the coverage would have been between 95.5% and 94.6%, which is certainly good enough.

Between what we see here and what we saw for non-normality and non-constant variance, it looks like randomized, balanced designs are least susceptible to violations of assumptions.

Error rates $\times 100$ of nominal 95% confidence intervals for $\mu_1 - \mu_2$, when neighboring data values have correlation ρ and data patterns are consecutive or alternate.

	ρ							
	−.3	−.2	−.1	0	.1	.2	.3	.4
Con.	.19	1.1	2.8	5	7.4	9.8	12	14
Alt.	12	9.8	7.4	5	2.8	1.1	.19	.001

All this should make you wonder about people who obsess over whether the p-value is .051 or .049.

Little, undetectable bits of non-normality, non-constant variance, or dependence can easily swing the p-value much more than that.