Lecture 6

One categorical predictor at k levels

Disclaimer: *These lecture notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students of stat 305A at Stanford University. Copyright: A. B. Owen 2026.*

Here we look at the regression model in the case where there are $k \geqslant 2$ groups. We have a single predictor $w \in \{1, 2, \dots, k\}$ and a response $y \in \mathbb{R}$. The linear model issues in this lecture are about the analysis of variance or ***ANOVA***. We also touch on statistical issues of multiple comparisons and random effects. In the most basic settings the ANOVA can be understood as regression on a collection of indicator variables. We will see some additional complexity brought on by random effects.

6.1 Notation

For observation i we get a categorical variable $w_i \in \{1, 2, \dots, k\}$ and a response $y_i \in \mathbb{R}$. In this setting, our derivations are easier if we recode the data as y_{ij} for $i = 1, \dots, k$ and $j = 1, \dots, n_i$. So $y_{i1}, \dots, y_{in_i}$ are the y values for the case of $w = i$.

The total number of observations is $N = \sum_{i=1}^{k} n_i$. We have promoted the sample size to a bigger font so that it is easily distinguishable from the n_i. We will almost always have $\min_i n_i \geqslant 2$ which allows us to estimate the

variance for data in group i. Things are simplest in the balanced case where n_i take a common value n. That's another reason to use N for the total sample size.

An early use was in agriculture. For k different field treatments, y_{ij} is the j'th yield measurement for treatment type i. The treatment types might represent different fertilizers or cultivation. In medicine, there could be k different therapies and y_{ij} is a response for the j'th person who got therapy i. In an industrial setting there could be k different manufacturing processes. In environmental science there could be k different lakes. In commerce, there might be $k = 3$ levels of suggested tip. In surveys, one question could have k levels and we might want to relate the answer to some other question to that k level question.

6.2 Models

The cell mean model for grouped data has

$$y_{ij} = \mu_i + \varepsilon_{ij}, \tag{6.1}$$

where $\varepsilon_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2)$ are independent. Then $y_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$. We will study the model using Gaussian errors. That gives exact distributions for our tests and confidence intervals.

Sometimes we re-write equation (6.1) as

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}. \tag{6.2}$$

Here μ is the 'grand mean' of the data and α_i is the 'effect' of level i as an offset relative to μ. The effects model (6.2) is over-parametrized. It has $k + 1$ parameters to describe k means. To break the tie, we impose a side condition $\sum_{i=1}^{k} n_i \alpha_i = 0$. That makes the weighted average effect equal to zero.

This setting is sometimes called the one way layout.

We have not said how the treatments have been assigned. One possibility is that we sample n_i observations from distribution F_i for $i = 1, \ldots, k$. Another possibility is that we have intervened and assigned treatment level i to n_i of the observations. That possibility also includes the case where we have made that assignment at random. A randomized allocation would give us more confidence that our statistical findings have a causal interpretation.

6.3 The analysis of variance

The first statistical problem is to decide whether the k groups really differ. That is usually about whether their means differ. We can formulate a null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

under which the mean responses are the same in every group. Another way to write it is

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0.$$

The cell mean model (6.1) is written in our regression notation as $Y = X\beta + \boldsymbol{\varepsilon}$ where

$$Y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} \tag{6.3}$$

and $\boldsymbol{\varepsilon}$ is partitioned similarly to Y. Each row of X above is a vector with one component equal to 1 and the other $k-1$ all equal to 0. This representation for the cell mean model is called the ***one-hot*** encoding. If we were to prepend an intercept column $\mathbf{1}_N$ into X, then we would have a rank deficient design matrix. The first column would be the sum of the subsequent k columns.

We easily find that

$$\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y = \begin{pmatrix} \bar{y}_{1\bullet} \\ \bar{y}_{2\bullet} \\ \vdots \\ \bar{y}_{k\bullet} \end{pmatrix},$$

where

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

The least squares estimates for the cell mean model are $\hat{\mu}_i = \bar{y}_{i\bullet}$.

For the effects model (6.2) we can fit the cell mean model first and then take $\hat{\alpha}_i = \bar{y}_{i\bullet} - \hat{\mu}$ where

$$\hat{\mu} = \bar{y}_{\bullet\bullet} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{N}\sum_{i=1}^{k} n_i \bar{y}_{i\bullet}.$$

Then $\sum_i n_i \hat{\alpha}_i = 0$.

Another way to estimate the effects model is to use

$$X = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{N\times k},$$

dropping the indicator of the first group from X. The corresponding regression parameter is $\beta = (\mu_1, \mu_2 - \mu_1, \ldots, \mu_k - \mu_1)^{\mathsf{T}}$. For $r > 1$, the interpretation of β_r depends not only on which group has the ones in column r. It also depends on which group was left out to become the intercept. There may be a principled choice such as a level that is much more common than the others, or a treatment that is a natural default such as a placebo in medicine. NB: we are indexing with r here, just to avoid using the letter i to index a column of X!

In this encoding we solve $\hat{\mu} + \hat{\alpha}_1 = \hat{\beta}_1$ along with $\hat{\beta}_r = \hat{\alpha}_r - \hat{\alpha}_1$ and $\sum_{i=1}^{k} n_i \hat{\alpha}_i = 0$ for $r = 2, \ldots, k$ to get $\hat{\mu}$ and $\hat{\alpha}_i$. **[Exercise]**

The sum of squared errors from the cell mean model is

$$\mathrm{SSE} = \mathrm{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\bullet})^2,$$

where SSW is mnemonic for the sum of squares *within* groups. The sum of squared errors under the null hypothesis of a common group mean is simply

$$\mathrm{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\bullet\bullet})^2.$$

The term SST refers to the *total* sum of squares.

Using the extra sum of squares principle leads us to test H_0 with

$$F = \frac{\frac{1}{k-1}\,(\mathrm{SST} - \mathrm{SSW})}{\frac{1}{N-k}\,\mathrm{SSW}}, \tag{6.4}$$

rejecting H_0 at level α if $F \geqslant F^{1-\alpha}_{k-1,N-k}$.

The quantity SST − SSW in the numerator of the F statistic can be shown to be

$$\mathrm{SSB} \equiv \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2, \tag{6.5}$$

the sum of squares *between* groups. Equation (6.5) is equivalent to the analysis of variance (ANOVA) identity:

$$\mathrm{SST} = \mathrm{SSB} + \mathrm{SSW}. \tag{6.6}$$

The ANOVA identity (6.6) can be understood in several ways. Algebraically we get it by expanding $\mathrm{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\bullet} + \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$, and watching the cross terms vanish. We can also divide the three terms by N and interpret it as an empirical version of the identity

$$\mathrm{var}(y) = \mathbb{E}(\mathrm{var}(y\,|\,x)) + \mathrm{var}(\mathbb{E}(y\,|\,x))$$

where x represents i and y represents y_{ij}.

The name ANOVA derives from this partition, or analysis, of the variance of Y into parts. Finally we can view it as an example of the Pythagorean theorem using the right angle triangle with vertices

$$Y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{pmatrix}, \quad \widehat{Y}_{\mathrm{FULL}} = \begin{pmatrix} \bar{y}_{1\bullet} \\ \vdots \\ \bar{y}_{1\bullet} \\ \bar{y}_{2\bullet} \\ \vdots \\ \bar{y}_{2\bullet} \\ \vdots \\ \bar{y}_{k\bullet} \\ \vdots \\ \bar{y}_{k\bullet} \end{pmatrix}, \quad \text{and} \quad \widehat{Y}_{\mathrm{SUB}} = \begin{pmatrix} \bar{y}_{\bullet\bullet} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \bar{y}_{\bullet\bullet} \end{pmatrix}.$$

Source	df	SS	MS	F
Groups	$k-1$	SSB	$\mathrm{MSB} = \mathrm{SSB}/(k-1)$	MSB/MSW
Error	$N-k$	SSW	$\mathrm{MSW} = \mathrm{SSW}/(N-k)$	
Total	$N-1$	SST		

Table 6.1: This is a generic ANOVA table for the one way layout.

6.4 Mean squares and the ANOVA table

The denominator of the F statistic is $\mathrm{SSW}/(N-k)$. This is also a sum of squared residuals divided by its degrees of freedom, which we have seen before. The numerator of the F statistic is $\mathrm{SSB}/(k-1)$. We divide the sum of squares between the k treatment means by $k-1$ which is the number of degrees of freedom among that many means. Intuitively we might consider a list of k effects α_i where we can set the first $k-1$ (hence $k-1$ dimensions or degrees of freedom) of them arbitrarily and then solve for the k'th one via $\sum_i n_i\alpha_i = 0$. Less intuitively we could write SSB as a quadratic form and see that the matrix in it has rank $k-1$.

The F statistic is a ratio of quantities called mean squares, namely $\mathrm{MSB} = \mathrm{SSB}/(k-1)$ and $\mathrm{MSW} = \mathrm{SSW}/(N-k)$. We take a sum of squares and divide it by its degrees of freedom.

The quantities underlying the F test are commonly arranged in an ANOVA table, as shown in Table 6.1. The degrees of freedom are the dimensions of some model spaces.

The vector Y varies in one less interesting direction, along $(\mu, \mu, \ldots, \mu) \in \mathbb{R}^N$ and in $N-1$ non-trivial directions orthogonal to this line. Of those $N-1$ non-trivial directions there are $k-1$ dimensions along which the group means can vary about the overall mean. Within group i there are $n_i - 1$ dimensions along which the n_i observations can vary about their common mean. The total degrees of freedom for variation within groups is $\sum_{i=1}^{k}(n_i-1) = N-k$.

Under H_0 we have $Y \sim \mathcal{N}(\mu\mathbf{1}_N, \sigma^2 I_N)$ where $\mathbf{1}_N$ is a column vector of N ones and $\mu \in \mathbb{R}$. The error $\boldsymbol{\varepsilon} = Y - \mu\mathbf{1}_N \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N)$ has a spherically symmetric distribution. Its squared projection on the $k-1$ dimensional space spanned by the cell mean model and orthogonal to $\mathbf{1}_N$ has a $\sigma^2\chi^2_{(k-1)}$ distribution. Its squared projection on the $N-k$ dimensional space orthogonal to the cell mean model has a $\sigma^2\chi^2_{(N-k)}$ distribution. These are independent of each other. Under H_0 the F statistic $\mathrm{MSB}/\mathrm{MSW} \sim F_{k-1,N-k}$, giving geometric intuition to what we got earlier from the extra sum of squares

method.

6.5 Power

The distribution of the F statistic for the one way layout is

$$F'_{k-1,N-k}\left(\frac{1}{\sigma^2}\sum_{i=1}^{k} n_i\alpha_i^2\right).$$

Then the test has power

$$\Pr\bigl(F'_{k-1,N-k}(\lambda) \geqslant F^{1-\alpha}_{k-1,N-k}\bigr) \quad \text{for} \quad \lambda = \frac{1}{\sigma^2}\sum_{i=1}^{k} n_i\alpha_i^2. \tag{6.7}$$

Let $\omega_i = n_i/N$. Then we can write the noncentrality parameter as

$$\lambda = N \times \frac{\sum_{i=1}^{k} \omega_i\alpha_i^2}{\sigma^2}.$$

The noncentrality and hence the power goes up with N, down with σ^2 and up with $\sum_{i=1}^{k} \omega_i\alpha_i^2$ which we can interpret as the variance of the treatment effects over the sample. So the power is driven by the sample size and a signal to noise ratio. Once N is large, increasing it does not much change the threshold $F^{1-\alpha}_{k-1,N-k}$ in (6.7), so extra denominator degrees of freedom don't make much difference.

Given a hypothetical pattern in the group means as well as an idea of σ^2 we can pick sample sizes to get the desired power. One way to reduce the number of choices is to take all $n_i = n$ and then increase n until the desired power has been attained.

6.6 Contrasts

In class I showed an example from Rice (2007). Seven different laboratories had measured the level of an antihistamine (chlorpheniramine maleate) in some pills. They each measured 10 pills. It is then interesting to study how much the mean measurement varies from laboratory to laboratory. The F test clearly rejected H_0 giving evidence that $\sum_i \alpha_i^2 > 0$. What is then more interesting is to consider how much variance comes from repeats within one of those labs (on average) and how much variance comes from variance

between labs. The ANOVA table gives us an estimate of the ratio of these variances.

What is usually even more interesting is to learn something about which pairs i and i' have significantly different values of $\mathbb{E}(y)$. We might want to know which i maximizes or minimizes μ_i.

When $k = 2$, then MSB/MSW $\sim F_{1,N-2}$. This F statistic is just the square of the usual t statistic. If we reject H_0 then we conclude that the two groups have different means. The reason that $k = 2$ is so simple is that there is just one interesting comparison $\mu_2 - \mu_1$ among the cell means. For larger k there is a whole $k-1$ dimensional family of differences. When $k \geqslant 3$ then rejecting H_0 tells us that there are some differences among the cell means μ_i but it does not say which of them are different.

If we're interested in the differences between groups i and i' then we could use a t test based on

$$t = \frac{\bar{y}_{i\bullet} - \bar{y}_{i'\bullet}}{s\sqrt{1/n_i + 1/n_{i'}}} \tag{6.8}$$

and rejecting when $|t| \geqslant t^{1-\alpha/2}_{(N-k)}$. The degrees of freedom in the t test are the same as in the estimate s^2 of σ^2. Data from groups $i'' \notin \{i, i'\}$ affect our comparison of groups i and i' because that data is used in s.

For the moment, we'll not worry about whether the variances are equal in the groups, because we have enough other things to worry about when they are equal. If $n_i \doteq n_{i'}$ then we would have less to worry about.

It is possible that the test in (6.8) rejects the hypothesis $\mu_i = \mu_{i'}$ on the same data for which the F test based on (6.4) fails to reject $\mu_1 = \mu_2 = \cdots = \mu_k$. It seems illogical that two groups could appear to differ while a superset of k groups appear not to differ. But there are $k(k-1)/2$ different pairs that could be tested via (6.8). Under H_0 we expect $k(k-1)\alpha/2$ false rejections from doing all pairwise t tests, but only α false rejections from the F test. So it is quite reasonable that (6.8) could be significantly large when the F test is not, despite how hard it might be to explain to a non-statistician.

Less intuitively, we could find that the F test rejects while none of the individual t tests reject. For example we might find that both $\bar{y}_{2\bullet} - \bar{y}_{1\bullet}$ and $\bar{y}_{3\bullet} - \bar{y}_{1\bullet}$ are almost but not quite significantly large. Perhaps they get a p-value of 0.06. Either of $\mu_2 - \mu_1$ or $\mu_3 - \mu_1$ could be zero. But it may be really unlikely that both of them are zero. In a case like this we would be confident that some difference exists but not sure which of the possible differences is real.

The two group t test above is just one way to look at group differences. More generally we could look at contrasts. A contrast is a linear combination

$\sum_{i=1}^k c_i\mu_i$ of the cell means where $\sum_{i=1}^k c_i = 0$, and to rule out trivialities, $\sum_{i=1}^k c_i^2 > 0$. We easily find that $\sum_{i=1}^k c_i\mu_i = \sum_{i=1}^k c_i\alpha_i$. The t test for a contrast is

$$t_{\boldsymbol{c}} = \frac{\sum_{i=1}^k c_i\bar{y}_{i\bullet}}{s\sqrt{\sum_{i=1}^k c_i^2/n_i}}. \tag{6.9}$$

Notice that we can replace all the c_i by ac_i for $a > 0$ and it would make no difference to t. That is, $t_{a\boldsymbol{c}} = t_{\boldsymbol{c}}$.

In class I used λ_i for the contrast elements but that clashes with λ for the noncentrality.

The meaningful contrasts depend on context and what the different groups represent. Consider the contrast vectors

$$\begin{aligned}
\boldsymbol{c}_1 &= (4, 4, 4, 0, 0, -3, -3, -3, -3),\\
\boldsymbol{c}_2 &= (k-1, -1, -1, \ldots, -1), \quad \text{and}\\
\boldsymbol{c}_3 &= (0, 0, 1, 0, -1).
\end{aligned}$$

for different problems with varying values of k. Contrast vector $\boldsymbol{c}_1$ tests for a difference between the average of the first 3 groups and the average of the last 4 groups, ignoring the middle two. We can see that by dividing it by 12. So it compares the average of α_i in one set of groups to the average in another set. Treatments within each of those sets might have something in common that we would want to test for.

The second contrast $\boldsymbol{c}_2$ compares the first group to the average of all the others. Maybe the first group represents a batch from a new supplier and the others represent $k-1$ earlier batches. The third contrast $\boldsymbol{c}_3$ is what we use when testing $\mu_3 = \mu_5$.

The square of the t test for a contrast is

$$t_{\boldsymbol{c}}^2 = \frac{(\sum_{i=1}^k c_i\bar{y}_{i\bullet})^2}{s^2\sum_{i=1}^k c_i^2/n_i} \sim F'_{1,N-k}\left(\frac{(\sum_{i=1}^k c_i\mu_i)^2}{\sigma^2\sum_{i=1}^k c_i^2/n_i}\right).$$

We can get a simpler expression for the balanced case. Then the noncentrality is n/σ^2 times

$$\frac{(\sum_{i=1}^k c_i\mu_i)^2}{\sum_{i=1}^k c_i^2} = \frac{(\sum_{i=1}^k c_i\alpha_i)^2}{\sum_{i=1}^k c_i^2} = \left(\frac{\boldsymbol{c}^{\mathsf{T}}\boldsymbol{\alpha}}{\|\boldsymbol{c}\|}\right)^2. \tag{6.10}$$

The noncentrality does not change if we normalize $\boldsymbol{c}$ to make it a unit vector. We maximize it by taking the contrast $\boldsymbol{c}$ to be parallel to the vector

$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$. It might look like we could maximize it by taking a vector $\boldsymbol{c}$ parallel to $(\mu_1, \ldots, \mu_k)$ but such a vector might not sum to zero, so it would not be a contrast.

If there is a linear trend among the levels, so that $\mu_i = A + Bi$, then the most powerful contrast for testing it has $c_i \propto (i - (k+1)/2)$. This can have surprising consequences. When $k = 3$ we get $\boldsymbol{c} \propto (-1, 0, 1)$. The contrast does not even use the middle group! The middle group is not exactly useless. It contributes to s^2 and we can use it to test whether the true trend might be nonlinear. Whenever k is odd, the middle group in a linear contrast will get weight 0. When k is even, then all groups get nonzero weight. For $k = 4$ the contrast comes out proportional to $(-3, -1, 1, 3)$.

If we really thought that the group means μ_i were nearly a linear function of group indices i, then using that linear contrast with $c_i \propto (i - (k+1)/2)$ may give a t test that is more powerful than the F test for the ANOVA. The F test is sensitive to every possible departure from equality of the means. The linear contrast 'specializes' in just that linear pattern.

Finally we can look at the contrast that gets the most significant t statistic. Again assuming a balanced ANOVA, we can maximize t^2 by taking $c_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}$. This is the *cheater's contrast* because it looks at the actual pattern in the group means and then tests for it.

After some algebra we find that

$$t^2_{\text{cheat}} = \frac{\text{SSB}}{s^2} \sim (k-1)F_{k-1,N-k} \tag{6.11}$$

under H_0. This distribution takes larger values than the $F_{1,N-k}$ distribution appropriate to contrasts that are specified in advance. When k is large it can be far larger.

If some contrast $\boldsymbol{c}$ had $t^2_{\boldsymbol{c}} \geqslant (k-1)F^{1-\alpha}_{k-1,N-k}$ then we can be sure that it is significant even if we picked it after looking at the data. That gives some protection against 'data dredging' or forming hypotheses based on the data we have seen. On the other hand, for large k, that threshold becomes very high. Scheffé's procedure is to declare the contrast $\sum_i c_i\mu_i$ significantly different from zero if

$$|t_{\boldsymbol{c}}| \geqslant \sqrt{(k-1)F^{1-\alpha}_{k-1,N-k}}.$$

If something is significant by this criterion, then no matter how much we scoured the data to come up with our hypothesis, we still have an experiment-wise type I error that is at most α.

6.7 Multiple comparisons

If we run all $k(k-1)/2$ two group hypothesis tests at level α, then we have much greater than α probability of rejecting H_0 when it holds. Our test-wise type I error rate is α but our experiment-wise error rate could be much higher.

It is possible to adjust the test statistics to get control of the experiment-wise type I error rate. The simplest method is Bonferroni's method. We test each of the $k(k-1)/2$ pairs at level $2\alpha/(k(k-1))$ instead of at level α. More generally, if we have a list of m contrasts to test, we can test them at level α/m. Clearly, the higher m is, the more stringent our tests become.

The Bonferroni test becomes quite conservative when m is large. We could use Scheffé's procedure to compensate for doing all those tests. That can be even more conservative than Bonferroni because it protects against all possible contrasts. Interestingly while Scheffé protects against essentially $m = \infty$ different tests, when we use it we do not use $\alpha/\infty = 0$ as our threshold. The reason is that those contrasts in Scheffé's procedure are statistically dependent.

Tukey introduced a test that is more appropriate than Scheffé's when we only want to do the $k(k-1)/2$ pairwise tests, and not all possible contrasts. Suppose that we sort the groups by their group means, so that $\bar{y}_{(1)\bullet} \leqslant \bar{y}_{(2)\bullet} \leqslant \cdots \leqslant \bar{y}_{(k)\bullet}$. For balanced data sets, we will get at least one significant pairwise t-test whenever $\bar{y}_{(k)\bullet} - \bar{y}_{(1)\bullet}$ is large enough. Tukey worked with the distribution of

$$Q_{n,k} \equiv \frac{\max_i \bar{y}_{i\bullet} - \min_i \bar{y}_{i\bullet}}{s\sqrt{2/n}},$$

(for normal data) and his test declares groups i and i' significantly different when the t test to compare those groups has absolute value larger than the $1-\alpha$ quantile of $Q_{n,k}$. The distribution of $Q_{n,k}$ is tabulated. Now we could just simulate it or use `qtukey` in R.

Sometimes we don't need to look at all pairwise comparisons. In a medical setting group 1 might get a placebo while groups 2 through k each get one of $k-1$ experimental treatments. If all of our comparisons are between the placebo and one of the experimental treatments, then our inferences can be based on the distribution of

$$D \equiv \max_{2\leqslant i\leqslant k} \frac{|\bar{y}_{i\bullet} - \bar{y}_{1\bullet}|}{s\sqrt{1/n_1 + 1/n_i}}.$$

Dunnett worked out the distribution of this quantity (for the balanced Gaussian setting) and his test declares treatment i different from placebo when

the absolute value of the t statistic between these groups exceeds the $1-\alpha$ quantile of D.

Modern computer power now lets us do our own home-cooking for multiple comparisons. If we can make a finite list $\mathcal{L} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m\}$ of the contrasts we're interested in, we can then simulate the distribution of $\max_{1\leqslant\ell\leqslant m} |t_{\boldsymbol{c}_\ell}|$ by sampling millions of normally distributed data sets with the same values of n_i as we have in the real data. We can also simulate from non-normal models if we have some that we would prefer.

Multiple comparison testing is a special case of multiple hypothesis testing. We formulate a family of null hypotheses, and make up a test with a family-wise type I error rate of Pr(any false rejections) $\leqslant \alpha$. The hard part is deciding what constitutes a family of tests. At one extreme, each test is its own family. At the other extreme, you could make up a family using every test that you ever make.

Whether some tests belong together as a family of tests depends on the context, and in particular on what actions you will take based on the outcomes of the tests, and on what the costs are from errors. It is an area where reasonable people can disagree.

Suppose for example that you are about to test $k-1$ unproven new treatments against a placebo in a medical context. Whichever one looks best will be followed up on, assuming that it is statistically significant, and developed. If you're wrong, and none of the new treatments are any good, then you'll lose a lot of time and money following up. Here one clearly wants the family to include all of the new versus placebo tests. You lose the same if you make two wrong rejections, following up one, as you would if you made just one wrong rejection.

In other cases the testing situations are independent and what you win or lose is additive (and small compared to your budget or total tolerance for loss). Then each test ought to be its own family.

6.8 False discovery rates

We did not cover this and it is not part of this year's applied statistics course. Many of you will see it in a theoretical statistics course. I'm actually skeptical of this method because it is derived assuming independence but commonly applied to dependent tests. The FDR generally is reasonably well controlled but in some settings the false discoveries come in bursts giving occasional FDPs far from the FDR that I think could mislead people.

Having to choose between the family-wise error rate Pr(any false rejections)

	H_0 not rejected	H_0 rejected	Total
H_0 True	U	V	m_0
H_0 False	T	S	m_1
Total	$m-R$	R	m

Table 6.2: Table illustrating the false discovery proportion.

and the per-comparison error rate $\mathbb{E}(\text{number of false rejections})/m$ is a false dichotomy. After making m hypothesis tests each test has either been rejected or not. Unknown to us, each of those m null hypotheses is either true or not. The possibilities can be counted in a 2×2 grid illustrated in Table 6.2.

Walking through this table, we have made m hypothesis tests of which m_0 really are null and m_1 are not. The total number of tests rejected is R of which V were true nulls and S were false nulls. Of the $m-R$ tests we do not reject, there are U true and T false. So we have made V type I errors (rejecting true H_0) as well as T type II errors (failing to reject a false H_0).

Put another way, we have made R discoveries of which V are false discoveries and S are true discoveries. So our proportion of false discoveries is $\text{FDP} = V/R$ where by convention $0/0$ is interpreted as 0. The false discovery rate is $\text{FDR} = \mathbb{E}(V/R)$.

Benjamini and Hochberg have developed a way of controlling the false discovery rate. The BH method, and a good many others, are developed in the book by Dudoit and van der Laan. Suppose that we control the false discovery rate to say 20%. Then on average our rejected hypotheses are at least 80% valid and of course up to 20% false discoveries. FDR methods are often used to screen a large list of hypotheses to identify a smaller list worthy of follow up experiments. It may well be acceptable to have a list of leads to follow up on that are mostly, but not completely, reliable.

The derivation of FDR control methods is outside the scope of this work. But operationally it is easy to do. The BH method works as follows. Take m p-values and sort them, getting $p_{(1)} \leqslant p_{(2)} \leqslant p_{(3)} \leqslant \ldots \leqslant p_{(m)}$. Let $L_i = i\alpha/m$. If all $p_{(i)} > L_i$ then reject no hypotheses. Otherwise let $r = \max\{i \,|\, p_{(i)} \leqslant L_i\}$. Finally, reject any hypothesis with $p_i \leqslant p_{(r)}$.

If the m p-values are statistically independent, then the BH procedure gives $\text{FDR} \leqslant \alpha\, m_0/m \leqslant \alpha$. When the p-values are dependent, then we can adjust the BH procedure taking $L_i = i\alpha/(mC_m)$ instead where $C_m = \sum_{i=1}^{m} 1/i \doteq \log(m)$. The result controls FDR for any possible dependence

pattern among the p_i so it may be conservative for a particular dependence pattern.

6.9 Random effects

The foregoing analysis covers treatments that are known as fixed effects: we want to learn about the actual k levels used. We also often have treatment levels that are random effects. There, the k levels we got are thought of as a sample from a larger population of levels.

For example, if one group got vitamin C and the other did not, then this two group variable is a fixed effect. Similarly, if we have three types of bandage to use that would be a fixed effect.

If instead we have patients $i = 1, \dots, 24$ with n measurements on each in some clinical trial, then we are almost surely looking at a random effect. We may want to learn something about those 24 patients, but our main interest is in the population of patients as a whole. We might treat the 24 we got as a sample representing that population. Similarly 8 rolls of ethylene vinyl acetate film in a manufacturing plant will constitute a random effect. Those 8 rolls will soon be used up and irrelevant, but we want to learn something about the process from which they (and future rolls) come.

Some variables are not clearly a fixed or a random effect. Perhaps they're both. For example, in a factory with four lathes, we might be interested in the differences among those four specific machines, or we might be interested more generally in how much some product measurements vary from lathe to lathe.

The model for random effects is

$$y_{ij} = \mu + a_i + \varepsilon_{ij}, \quad i = 1, \dots, k,\ j = 1, \dots, n_i, \tag{6.12}$$

where now a_i are random and independent of ε_{ij}. The standard assumptions are that $a_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_A^2)$ independently of $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_E^2)$. Now we have two random quantities a_i and ε_{ij}.

In this model all of the observations have mean μ and variance $\sigma_A^2 + \sigma_E^2$. But they are not all independent. Instead

$$\operatorname{cor}(y_{ij}, y_{i'j'}) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2} \times 1_{i=i'} + \frac{\sigma_E^2}{\sigma_A^2 + \sigma_E^2} \times 1_{i=i'} 1_{j=j'}.$$

So the whole $N \times N$ correlation matrix has k blocks, each of which has 1s on the diagonal and $\rho = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$ off the diagonal. Notice that this model cannot give us negative correlations.

In this simplest random effects setting we're interested in σ_A^2, often through the ratio σ_A^2/σ_E^2. That way we can tell what fraction of the variation in y_{ij} comes from the group variable and what fraction comes from measurement error.

We assume that learning about σ_A^2 will tell us something about the other levels of a_i that were not sampled. This is most straightforward if the levels $i = 1, \ldots, k$ in the data set really are a sample from the possible levels. In practice it is a little more complicated. The patients in a clinical trial may be something like a sample from the kind of patients that a given clinic sees. They won't usually be a random sample from the population (they're probably sicker than average). They might or might not be similar to the patients seen at other clinics.

The statistical analysis of the random effects model still uses the ANOVA decomposition:

$$\begin{aligned}
\mathrm{SST} &= \mathrm{SSA} + \mathrm{SSE} \\
\mathrm{SST} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\bullet\bullet})^2 \\
\mathrm{SSA} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 \\
\mathrm{SSE} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\bullet})^2,
\end{aligned}$$

where now we use SSA for the sum of squares for factor A instead of the SSB for sum of squares between groups.

The fact that the treatment levels are random has not changed the algebra or geometry of the ANOVA decomposition. Those y_{ij} don't know whether they were based on fixed or random effects. The distinction does change the distributions of these sums of squares. To get an intuitive idea of the issues we look at the balanced case where $n_i = n$.

First SSE is the sum of k independent terms that are $\sigma_E^2\,\chi^2_{(n-1)}$ for $i = 1, \ldots, k$. Therefore

$$\mathrm{SSE} \sim \sigma_E^2\,\chi^2_{(N-k)}.$$

Next the group means are IID:

$$\bar{y}_{i\bullet} = \mu + a_i + \frac{1}{n}\sum_{j=1}^{n}\varepsilon_{ij} \sim \mathcal{N}\Big(\mu, \sigma_A^2 + \frac{1}{n}\,\sigma_E^2\Big).$$

Then

$$\frac{1}{n}\mathrm{SSA} = \sum_{i=1}^{k}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 \sim \Big(\sigma_A^2 + \frac{1}{n}\sigma_E^2\Big)\chi^2_{(k-1)},$$

because it is the sample variance formula from k IID normal random variables.

To test the null hypothesis that $\sigma_A^2 = 0$ we use $F = \mathrm{MSA}/\mathrm{MSE}$ as before. But now

$$\begin{aligned} F &= \frac{\frac{1}{k-1}\mathrm{SSA}}{\frac{1}{N-k}\mathrm{SSE}} = \frac{\frac{n}{k-1}(\sigma_A^2 + \sigma_E^2/n)\chi^2_{(k-1)}}{\frac{1}{N-k}\sigma_E^2\chi^2_{(N-k)}} \\ &\sim \frac{n\sigma_A^2 + \sigma_E^2}{\sigma_E^2}\, F_{k-1,N-k} \\ &= \Big(1 + n\frac{\sigma_A^2}{\sigma_E^2}\Big)\, F_{k-1,N-k}. \end{aligned}$$

So under H_0 where $\sigma_A^2 = 0$, we still get the $F_{k-1,N-k}$ distribution. We really had to, because H_0 for the random effects model leaves data with the same distribution as H_0 does for the fixed effects model. But when H_0 does not hold, instead of a noncentral F distribution the F statistic is distributed as a multiple of a central F distribution.

If we reject H_0 we don't have the problem of deciding which means differ, because our model has no mean parameters, except for μ. We may however want to get estimates for σ_A^2 and σ_E^2. The usual way is to work with expected mean squares

$$\begin{aligned} \mathbb{E}(\mathrm{MSE}) &= \sigma_E^2, \quad \text{and,} \\ \mathbb{E}(\mathrm{MSA}) &= n\sigma_A^2 + \sigma_E^2. \end{aligned}$$

Notice how the residual variance σ_E^2 contributes to the mean square for A. Noise from the errors propagates through to inflate MSA. Unbiased estimators are then obtained as

$$\begin{aligned} \hat{\sigma}_E^2 &= \mathrm{MSE}, \quad \text{and} \\ \hat{\sigma}_A^2 &= \frac{1}{n}\Big(\mathrm{MSA} - \mathrm{MSE}\Big). \end{aligned}$$

We can get $\hat{\sigma}_A^2 < 0$. This may happen if the true σ_A^2 is very small. Sometimes one just takes $\hat{\sigma}_A^2 = 0$ instead of a negative value. That will leave an estimate with some bias, because it arises as an unbiased estimate that we sometimes increase.

It is also possible that we can find $\hat{\sigma}_A^2 < 0$ if observations from the same treatment group are negatively correlated with each other. As noted above that could never happen in the model (6.12). But it could happen in real problems. For example if the treatment group were plants in the same planter box (getting the same fertilizer) then competition among the plants in the same box could produce negative correlations among their yields. It is not pre-ordained that this correlation would be negative. It is also possible that the amount of sunshine varies between planters and that may induce positive correlations.

Sometimes we want to get estimates of the actual random effect values a_i. Because a_i are random variables and not parameters, many authors prefer to speak of ***predicting*** a_i instead of estimating it. That is odd terminology since a_i may be a quantity from the past and we ordinarily think of prediction as about a future quantity that we will actually observe. I think that ***estimating*** a_i is just fine.

If we knew μ, σ_A^2, and σ_E^2, then we could work with the joint normal distribution of $(a_1, \ldots, a_k, y_{11}, \ldots, y_{kn_k})^{\mathsf{T}}$ and predict a_i by $\mathbb{E}(a_i \,|\, y_{11}, \ldots, y_{kn_k})$. That is we predict the unmeasured value by its conditional expectation given all of the measured ones. We can get the result under a Gaussian assumption. The result is

$$\hat{a}_i = \frac{n\sigma_A^2}{\sigma_E^2 + n\sigma_A^2}\Big(\bar{y}_{i\bullet} - \mu\Big).$$

Then the corresponding estimator of $\mathbb{E}(y_{ij})$ would be

$$\mu + \hat{a}_i = \frac{n\sigma_A^2 \bar{y}_{i\bullet}}{\sigma_E^2 + n\sigma_A^2} + \frac{\sigma_E^2 \mu}{\sigma_E^2 + n\sigma_A^2}. \tag{6.13}$$

Equation (6.13) can be seen as shrinking the observed group average towards the overall mean μ. The amount of shrinkage increases with the ratio $\sigma_E^2/(n\sigma_A^2)$. In practice we would have to plug in estimates of μ and σ_A^2 and σ_E^2.

The book "Variance Components" by Searle, Casella and McCulloch gives an extensive treatment of random effects models. It includes coverage of maximum likelihood estimation of variance components in models like the one considered here as well as much more general models. It also looks at predictions of random effect components using estimated values of the parameters. There can be many more than two sources of randomness. The full complexity of variance components analysis takes us far beyond regression on binary variables. In this course we will look later at what happens when there are two of these categorical variables. We could have two fixed effects, two random effects, or one of each.

Bibliography

J. A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, third edition, 2007.