Summary of Maxwell and Delaney's presentation of ANOVA and linear models

Summarized by Shravan Vasishth

September 19, 2010

Having understood some of the foundational ideas behind ANOVA and linear models in our book (The foundations of statistics: A simulation-based approach, Vasishth & Broe, 2010), it may be worth looking at ANOVA from a model comparison perspective. This additional material is based on the presentation of Maxwell & Delaney, 2000, a book that we highly recommend to readers interested in more detail on ANOVA and linear models. This chapter is a bit heavy on mathematics, but the mathematics does not go beyond the 10th grade level (no calculus or linear algebra). So don't be put off; it's worth the effort to go through this chapter because it will solidify your understanding of how linear regression models and ANOVA fit together.

We have already seen in several settings now that we can think of hypothesis testing as comparing two linear models, and we have seen the relationship of ANOVA to the linear models in several ways.

Consider now the general case of the linear model:

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \epsilon_i \tag{1}$$

Here, Y_i is the score of participant i on the dependent variable, the observed values, β_0 is the *population* mean μ, β_1 and β_2 are slopes for factors of interest, PARAMETERS that have to be estimated. The slopes signify the contributions of the factors. The VARIABLES X_{1_i} and X_{2_i} refer to the scores from the relevant group (recall the discussion of the three-group data from chapter 5). ϵ_i is the residual (everything that we cannot account for based on the factors A and B).

The equation is completely general: if you have k factors, the equation is:

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \cdots + \beta_k X_{k_i} + \epsilon_i \tag{2}$$

In the following discussion, we build up the underlying details by beginning with a simple example, and then successively look at increasingly complex situations.

0.0.1 One group example

Consider first the simple situation when you have only one group. I.e., instead of two factors affecting the dependent variable Y_i, we consider the case where there is only one.

$$Y_i = \mu + \epsilon_i, i = 1, \ldots, n \text{ where } n = \text{no. of participants} \tag{3}$$

The equation asserts that the variable Y has some unknown typical value μ (the population mean), and that deviations from this typical value are due to random, uncontrolled factors: ϵ_i for each subject i.

Notice that one could have written (3) as

$$Y_i = \beta_0 + \epsilon_i \tag{4}$$

Here, $\beta_0 = \mu$. Equation (3) is actually a series of equations, one for each i. So you have n equations, and $n+1$ unknowns (n unknowns are the ϵ_i's, and there is one unknown population mean μ). We could use *any* of a number of possible values of μ and ϵ_i, but we want a unique solution.

To get the unique, optimal solution, we can treat equation (3) as a prediction equation, where one tries to guess a value for μ that is *as close as possible* to the observed values Y_i. We quantify "as close as possible" with the constraint: minimize ϵ_i.

Suppose we take the mean of our observed values as our predicted (or estimated) value of μ; call this $\hat{\mu}$ (pronounced "μ hat"). Then we want to minimize e_i, which is defined as follows:

$$e_i = \hat{\epsilon_i} = Y_i - \hat{\mu} \tag{5}$$

Obviously, what this means is:

$$Y_i = \hat{\mu} + \hat{\epsilon_i} \tag{6}$$

If you try to get your guess $\hat{\mu}$ as close as possible to the observed Y_i, the extent to which you "missed" is $\hat{\epsilon_i}$. You want to minimize this miss. $\hat{\epsilon_i}$ is thus the error of prediction for each subject, and is estimated by e_i. We can use $e_i^2 = (Y_i - \hat{\mu})^2$ as a measure of our (in)accuracy (or lack thereof), since squaring it gets rid of negative signs, and emphasizes large errors.

So, minimizing e_i is the same as stipulating that the sum or average of e_i^2 is as small as possible. Choosing parameter estimates to minimize squared errors of prediction is the familiar the LEAST SQUARES CRITERION.

In least squares estimation (LSE) we minimize

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\mu})^2 \tag{7}$$

by choosing the appropriate $\hat{\mu}$. Recall that the *sample* mean $\bar{Y}$ has the property that the sum of squared deviations from it are smaller than around any other value. So $\bar{Y}$ is really the best estimator for $\hat{\mu}$.

An important by-product of using LSE to estimate parameters is that it yields a measure of the ACCURACY of the model that is as fair as possible. That is, $\sum_{i=1}^{n} e_i^2$ is as small as it could be for this model.

Suppose now that we know (from previous experiments) that the mean of a population is known to be μ_0, and we wonder whether it is (close to) μ_0 for a particular sample of the population as well. To give a concrete example, suppose we know that the mean height of all children of a particular age group is μ_0, and we want to know if the mean height μ_i of a specific group of i children (of the same age group) is also μ_0.

The hypothesis that $\mu_0 = \mu_i$ is the NULL HYPOTHESIS H_0. This null hypothesis can be written as:

$$H_0 : Y_i = \mu_0 + \epsilon_i \tag{8}$$

Note that we're estimating *no* parameters here; Maxwell and Delaney (2000) call this the RESTRICTED MODEL, and we will adopt their terminology. Compare equation (8) with the earlier one (6), repeated below:

$$Y_i = \hat{\mu} + \hat{\epsilon}_i \tag{9}$$

where we're estimating $\hat{\mu}$. Call this the UNRESTRICTED MODEL.

Now, for the restricted model,

$$e_i = \epsilon_i = Y_i - \mu_0 \tag{10}$$

which means that

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \mu_0)^2 \tag{11}$$

With some algebraic manipulation (exercise) for the restricted model you get

$$\sum_{i=1}^{n} (Y_i - \mu_0)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2 \tag{12}$$

Now, the minimal error made in the *un*restricted model is:

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{13}$$

Suppose our null hypothesis were true (i.e., $\mu_0 = \mu_i = \bar{Y}$). Then, there would be no difference between the restricted model's error e_{i_R} and the unrestricted model's error e_{i_U}:

$$e_{i_R} - e_{i_U} = 0 \tag{14}$$

This is obvious since $\mu_0 = \bar{Y}$:

$$\left(\sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2 \right) - \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \tag{15}$$

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2 = \tag{16}$$

$$n(\bar{Y} - \mu_0)^2 = \tag{17}$$

$$n(\mu_0 - \mu_0)^2 = 0 \tag{18}$$

It also follows that if the null hypothesis is not true, then

$$e_{i_R} - e_{i_U} \neq 0 \tag{19}$$

This is also obvious since $\mu_0 \neq \bar{Y}$:

$$\left(\sum_{i=1}^{n}(Y_i-\bar{Y})^2+n(\bar{Y}-\mu_0)^2\right)-\sum_{i=1}^{n}(Y_i-\bar{Y})^2= \tag{20}$$

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2-\sum_{i=1}^{n}(Y_i-\bar{Y})^2+n(\bar{Y}-\mu_0)^2= \tag{21}$$

$$=n(\bar{Y}-\mu_0)^2 \tag{22}$$

In other words, the further away $\bar{Y}$ is from our hypothesized value μ_0, the larger the difference in errors.

The key inferential step comes at this point. How much must the error increase for our assumption to be false that μ_0 is the mean of the subset we're interested in? We can take proportional increase in error:

$$\text{Prop. increase in error} = \frac{\text{increase in error}}{\text{minimal error}} \tag{23}$$

This leads to our familiar idea of a TEST STATISTIC. First, let's fix some terminology. Call the unrestricted model the FULL MODEL $\mathcal{F}$ because it contains parameters to be estimated (the group means). Call the restricted model $\mathcal{R}$; recall that in $\mathcal{R}$ we've placed restrictions on the parameters of $\mathcal{F}$. For example, we've deleted a parameter (in the above one-group example). This restriction is our null hypothesis (specifically, in our example, the hypothesis that μ_0 is the subject's mean).

To summarize:

	Model	LSE	Errors
$\mathcal{F}$	$Y_i=\hat{\mu}+\epsilon_{i\mathcal{F}}$	$\hat{\mu}=\bar{Y}$	$E_{\mathcal{F}}=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(Y_i-\bar{Y})^2$
$\mathcal{R}$	$Y_i=\mu_0+\epsilon_{i\mathcal{R}}$	No parameters estimated	$E_{\mathcal{R}}=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(Y_i-\mu_0)^2$

It follows that

$$\text{Prop. increase in error} = \frac{(E_{\mathcal{R}}-E_{\mathcal{F}})}{E_{\mathcal{F}}}=\frac{n(\bar{Y}-\mu_0)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} \tag{24}$$

Proportional increase in error compares the *adequacy* of the models, but ignores their *relative* complexity. We know already that $\mathcal{R}$ has to be less adequate than our $\mathcal{F}$. This is because if the restricted model is less adequate than the full model, $E_{\mathcal{F}} < E_{\mathcal{R}}$). Notice that we're gaining simplicity but losing adequacy in moving from $\mathcal{F}$ to $\mathcal{R}$ (the restricted model has fewer parameters, so it's simpler, in a sense). If we could find out what the loss in adequacy was *per additional unit of simplicity*, we have a measure of the relative adequacy of the models $\mathcal{F}$ and $\mathcal{R}$, taking their relative simplicity into account.

If, in transitioning from $\mathcal{F}$ to $\mathcal{R}$, the loss in adequacy *per unit gain in simplicity* is large, then we have some reason to believe that our null hypothesis was false.

Quantifying simplicity of a model is the key problem now. The fewer the parameters, the simpler the model. Conversely, the more the number of parameters, the more complex the model.

To quantify simplicity, we want a number that should increase as the number of parameters decrease. This is our familiar "degrees of freedom", *df*, which we define as follows:

$$\text{df} = (\text{no. of independent observations}) - (\text{no. of indep. parameters estimated}) \tag{25}$$

We can use *df* as our index of simplicity. So now we have the right measure – proportional increase in error relativized to simplicity:

$$\text{Prop. increase in error} = \frac{(E_{\mathcal{R}} - E_{\mathcal{F}})/(df_{\mathcal{R}} - df_{\mathcal{F}})}{E_{\mathcal{F}}/df_{\mathcal{F}}} = F \tag{26}$$

The interesting thing with this presentation is that *all* tests in ANOVA, ANCOVA, bivariate and multiple regression can be computed using this formula. Every new setup discussed after this point depends on the above result.

Notice that if the adequacy of $\mathcal{R}$ and $\mathcal{F}$ per degree of freedom is the same, $F = 1$. In that case, we'd prefer the simpler model $\mathcal{R}$. On the other hand, if the error per *df* of $\mathcal{R}$ is larger, the simpler model is inadequate; this amounts to saying that there is a significant difference between the population mean and the mean of the subset we're interested in. For example, if $F = 9$, that means that the additional error of the simpler, restricted model per its additional *df* is nine times larger than we would expect it to be on the basis of the error for the full model per degree of freedom. That is, the restricted model is considerably worse per extra degree of freedom in describing the data than is the full model relative to its *df*.

This can be re-stated as follows:

Hypothesis	Model
$H_1 : \mu \neq \mu_0$	Full : $Y_i = \mu + \epsilon_{i_{\mathcal{F}}}$
$H_0 : \mu = \mu_0$	Restricted: $Y_i = \mu_0 + \epsilon_{i_{\mathcal{R}}}$

1 Extending Linear Models to two groups

The above can be extended to two groups. The situation now is summarized as follows. Let μ_1 be the population mean for one group, μ_2 the population mean for the other group. More generally, let μ_j be the population mean for the jth group (i.e., $j = 1, 2$), and $i = 1, \ldots, n_j$.

Hypothesis	Model
$H_1 = \mu_1 \neq \mu_2$	Full = $Y_{ij} = \mu_j + \epsilon_{ij_{\mathcal{F}}}$
$H_0 = \mu_1 = \mu_2$	Restricted = $Y_{ij} = \mu + \epsilon_{ij_{\mathcal{R}}}$

To take a concrete example, we could be comparing the IQs of two groups of children, one "normal", and the other hyperactive. The normal group of n_1 children ($j = 1$) would have IQ μ_1, and the hyperactive group consisting of n_2 children ($j = 2$) would have IQ μ_2. We want to know if $\mu_1 = \mu_2$ (the null hypothesis).

After a little bit of mathematics (interesting, but we can skip it; see (Maxwell & Delaney, 2000, 77-80) for details), we get the following for the two-group situation:

$$PIE = \frac{(E_{\mathcal{R}} - E_{\mathcal{F}})/(df_{\mathcal{R}} - df_{\mathcal{F}})}{E_{\mathcal{F}}/df_{\mathcal{F}}} = \frac{\sum\limits_j n_j(\bar{Y}_j - \bar{Y})^2}{\sum\limits_j\sum\limits_i (Y_{ij} - \bar{Y}_j)^2/(N-2)} \tag{27}$$

1.1 Traditional terminology of ANOVA

Traditionally, F tests are supposed to indicate whether between-group variability is greater than within-group variability:

$$F = \frac{\text{Variability between groups}}{\text{Variability within groups}} \tag{28}$$

$$= \frac{\text{Mean square error between groups}}{\text{Mean square error within groups}} \tag{29}$$

$$= \frac{MS_b}{MS_w} \tag{30}$$

Intuitively, the logic is as follows. Given two groups with means μ_1 and μ_2, it is almost certain that $\mu_1 \neq \mu_2$, because of sampling variability. So the question really is: is the difference between treatment groups greater than within each group? The latter would be due to sampling variability. The equation in (28) reflects this.

The difference between μ_1 and μ_2 depends on the variability of the population, which can be estimated: take either of the groups' variance, or a weighted average of the two (weighted by the number of scores in each group).

Suppose each group's variance is s_j^2:

$$s_j^2 = \frac{\sum\limits_i (Y_{ij} - \bar{Y}_j)^2}{n_j - 1} \tag{31}$$

Then, σ^2, the weighted[1] average (or pooled estimate) of the two variances, s_1^2 and s_2^2, is:

$$\sigma^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)}{n_1 + n_2 - 2} \tag{32}$$

From equation (31) it follows that (**make sure that you see why it follows!**):

$$\sigma^2 = \frac{\sum\limits_i (Y_{i1} - \bar{Y}_1)^2 + \sum\limits_i (Y_{i2} - \bar{Y}_2)^2}{n_1 + n_2 - 2} \tag{33}$$

$$= \frac{\sum\limits_j\sum\limits_i (Y_{ij} - \bar{Y}_j)^2}{\sum\limits_j (n_j - 1)} \tag{34}$$

[1] Weighted by the number of free parameters: this is $n_j - 1$ since we've already "used up" one parameter to estimate s_j^2 – the mean $\bar{Y}_j$.

Since the last result above is an average or mean squared deviation *within* the groups, we have MS_w:

$$MS_w = \frac{\sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2}{\sum_j (n_j - 1)} \tag{35}$$

Notice here that the sum of squares within the groups, call it SS_w, is:

$$SS_w = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 \tag{36}$$

We will be using SS_w a lot in the future, so it's a good idea to internalize what this means.

Next, we calculate the MS_b, the mean standard deviation *between* the two groups.

Suppose the null hypothesis is true: $\mu_1 = \mu_2$. What is the variability between the sample means μ_1 and μ_2? I.e., what is the variance of the means? The answer is:

$$\frac{\sum_j (\bar{Y}_j - \bar{Y})^2}{a - 1} \tag{37}$$

This is the variance of the sample means; call it $\sigma^2_{\bar{Y}}$. We know that (assuming an equal number of scores n in both groups):

$$\sigma^2_Y = n \times \sigma^2_{\bar{Y}}, \text{where } \sigma^2_Y \text{ is the population variance} \tag{38}$$

It follows that

$$\sigma^2_{\bar{Y}} = n \times \frac{\sum_j (\bar{Y}_j - \bar{Y})^2}{a - 1} \tag{39}$$

This is the mean squared deviation between groups:

$$MS_b = n \times \frac{\sum_j (\bar{Y}_j - \bar{Y})^2}{a - 1} \tag{40}$$

Again, here the sum of squares between groups, call it SS_b, is:

$$SS_b = n \times \sum_j (\bar{Y}_j - \bar{Y})^2 \tag{41}$$

We will need SS_b again later.

Regarding MS_b and SS_b, note that when you have $j = a$ groups with n_j subjects in each group, the equation generalizes to:

$$MS_b = \frac{\sum_{j=1}^{a} n_j (\bar{Y}_j - \bar{Y})^2}{a - 1} \tag{42}$$

$$SS_b = \sum_{j=1}^{a} n_j (\bar{Y}_j - \bar{Y})^2 \tag{43}$$

To summarize:

- Sum of squares between groups:

$$SS_b = \sum_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})^2 \tag{44}$$

- Mean square deviation between groups:

$$MS_b = \frac{\sum\limits_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})^2}{a-1} \tag{45}$$

- Sum of squares within groups:

$$SS_w = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 \tag{46}$$

- Mean square deviation within groups:

$$MS_w = \frac{\sum\limits_j \sum\limits_i (Y_{ij} - \bar{Y}_j)^2}{\sum\limits_j (n_j - 1)} \tag{47}$$

Now we can see the connection between the model-comparison approach and the traditional view of F tests:

$$F = \frac{MS_b}{MS_w} \tag{48}$$

$$= \frac{\sum\limits_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})^2}{a-1} \div \frac{\sum\limits_j \sum\limits_i (Y_{ij} - \bar{Y}_j)^2}{\sum\limits_j (n_j - 1)} \tag{49}$$

$$= PIE \tag{50}$$

This is because (recall (27), repeated below):

$$PIE = \frac{(E_{\mathcal{R}} - E_{\mathcal{F}})/(df_{\mathcal{R}} - df_{\mathcal{F}})}{E_{\mathcal{F}}/df_{\mathcal{F}}} = \frac{\sum\limits_j n_j(\bar{Y}_j - \bar{Y})^2}{\sum\limits_j \sum\limits_i (Y_{ij} - \bar{Y}_j)^2/(N-2)} \tag{51}$$

2 Individual comparisons of means – between subject data

Our null hypothesis (in an a-group study) has so far been:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \tag{52}$$

The full and restricted models were:

$$Y_{ij} = \mu_j + \epsilon_{ij_{\mathcal{F}}} \tag{53}$$

$$Y_{ij} = \mu + \epsilon_{ij_{\mathcal{R}}} \tag{54}$$

Suppose now that our null hypothesis is: Do the means of two of the groups (say groups 1 and 2) differ? I.e.,

$$H_0 : \mu_1 = \mu_2 \tag{55}$$

The *restricted* model now changes to

$$Y_{ij} = \mu_i + \epsilon_{ij_{\mathcal{R}}}, \text{where } \mu_1 = \mu_2 \tag{56}$$

We could re-write this as:

$$Y_{i1} = \mu^* + \epsilon_{i1_{\mathcal{R}}} \tag{57}$$

$$Y_{i2} = \mu^* + \epsilon_{i2_{\mathcal{R}}} \tag{58}$$

$$Y_{ij} = \mu_j + \epsilon_{ij_{\mathcal{R}}}, j = 3, 4, \ldots, a \tag{59}$$

The new means μ^* refers to the mean of the two groups' scores; groups 3 to a can have their own potentially unique means. Since we've identified the restricted and full models, determining the F value is simply a matter of algebraic manipulation.

$$E_F = SS_w = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 \tag{60}$$

$$E_R = \sum_{j=1}^{2} \sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}^*\right)^2 + \sum_{j=3}^{n} \sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}_j\right)^2 \tag{61}$$

$$F = \frac{(E_{\mathcal{R}} - E_{\mathcal{F}})/(df_{\mathcal{R}} - df_{\mathcal{F}})}{E_{\mathcal{F}}/df_{\mathcal{F}}} \tag{62}$$

Recall the definition of df:

$$df = \text{no. of independent observations} - \text{no. of parameters} \tag{63}$$

Since $df_F = N - a$, where N is the total number of scores (across all groups, and $df_R = N-(a-1)$, we can rewrite F as follows (after some algebraic messing around, that is):

$$F = \frac{n_1 n_2 (\bar{Y}_1 - \bar{Y}_2)^2}{(n_1 + n_2) MS_w} \tag{64}$$

3 Complex comparisons

Suppose we administer a blood pressure treatment study. There are four treatments, and one is called a "combination treatment". Suppose our research question was: "Is the combination treatment more effective than the average of the other three?" The corresponding null hypothesis is simply the negation of this statement, and is expressed as shown below:

$$H_0 : \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \mu_4 \tag{65}$$

The full model remains unchanged:

$$Y_{ij} = \mu_j + \epsilon_{ij\mathcal{F}} \tag{66}$$

but the corresponding restricted model is:

$$Y_{ij} = \mu_j + \epsilon_{ij\mathcal{R}}, \text{ where } \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \mu_4 \tag{67}$$

Suppose we re-write the null hypothesis as:

$$H_0 : \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 - \mu_4 = 0 \tag{68}$$

A more general form of this hypothesis would be:

$$H_0 : c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + c_4\mu_4 = 0 \tag{69}$$

In the present case, $c_1 = c_2 = c_3 = \frac{1}{3}$ and $c_4 = -1$. Let us stipulate that the situation in equation (69) is a CONTRAST (or COMPARISON). In other words, let us define contrast ψ as a linear combination of population means in which the coefficients of the means sum to zero.

$$\psi = \sum_j^a c_j\mu_j \text{ where } \sum_j^a c_j = 0 \tag{70}$$

Such a general definition of contrasts allows us to test *any* contrast at all.

Note that the coefficients need not sum to zero; this is just a stipulation. However, when they don't, it is often the case that the contrast doesn't really mean anything. For example, in the blood pressure example, one could ask if the combination treatment is four times better than the average of the other three means. Here, $c_1 = c_2 = c_3 = 1/3$ and $c_4 = -4$, and the sum of coefficients is $1 - 4 = -3$. Maybe we do want to know the answer to such a question; it depends on the situation.

By the way, now our null hypothesis can simply be stated in terms of the contrast of interest:

$$H_0 : \psi = 0 \tag{71}$$

It is possible, but difficult, to find $E_{\mathcal{R}}$. But what we really need is only $E_{\mathcal{R}} - E_{\mathcal{F}}$, so that's what we'll derive (take it on trust for now).

$$E_{\mathcal{R}} - E_{\mathcal{F}} = \frac{(\hat{\psi})^2}{\sum_{j=1}^{a}(c_j^2/n_j)} \tag{72}$$

Here, $\hat{\psi}$ is a sample estimate of the population parameter ψ:

$$\hat{\psi} = \sum_{j=1}^{a} c_j \bar{Y}_j \tag{73}$$

Now we're almost ready to replace the terms in the equation for F:

$$F = \frac{(E_{\mathcal{R}} - E_{\mathcal{F}})/(df_{\mathcal{R}} - df_{\mathcal{F}})}{E_{\mathcal{F}}/df_{\mathcal{F}}} \tag{74}$$

Recall again the definition of df:

$$df = \text{no. of independent observations} - \text{no. of parameters} \tag{75}$$

Let's compute the dfs. For a groups, we will have $a-1$ parameters. This is because the ath parameter is predictable if we fix the others: if you have four groups, and your null hypothesis is

$$H_0 : \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 - \mu_4 = 0 \tag{76}$$

then μ_4 is predictable if you know (or have estimated) μ_1 to μ_3. So: df_R is $N-(a-1)$, where N is the number of independent observations. df_F is simply $N-a$, because you have a parameters.

So:

$$df_{\mathcal{R}} - df_{\mathcal{F}} = (N-(a-1)) - (N-a) = 1 \tag{77}$$

Also, recall that:

$$E_F/df_{\mathcal{F}} = MS_w \tag{78}$$

Finally, we're there:

$$F = \frac{\hat{\psi}^2}{MS_w \sum_{j=1}^{a} (c_j^2/n_j)} \tag{79}$$

4 Generalizing the model comparison technique

There is an easier way of talking about the effect a given factor has on the dependent variable (we will use this technique a lot later on, so it's useful to learn it now).

Instead of writing our full model for a groups as

$$Y_{ij} = \mu_j + \epsilon_{ijF} \tag{80}$$

we can write

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij_F} \tag{81}$$

The above reformulation amounts to saying that each factor $j = 1 \dots a$, contributes α_j to the mean: $\alpha_1 \dots \alpha_a$ are the effects of each of the factors.

In order to solve the equation above with a unique solution, we impose (as earlier) a side condition:

$$\sum_{j=1}^{a} \alpha_j = 0 \tag{82}$$

This is not a random choice. Notice that:

$$\mu_j = \mu + \alpha_j \tag{83}$$

That is, α_j is simply the deviation from the mean:

$$\alpha_j = \mu_j - \mu \tag{84}$$

We know that the sum of deviations from the mean sum to zero, and so

$$\sum_{j=1}^{a} \alpha_j = \sum_{j=1}^{a} (\mu_j - \mu) = 0 \tag{85}$$

Also, notice that it follows that the grand mean μ is:

$$\mu = \frac{\sum\limits_{j=1}^{a} \mu_j}{a} \tag{86}$$

Now we estimate parameters in the style of our one group example earlier (1):

Our equation for the full model, i.e.,

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij_F} \tag{87}$$

is a system of linear equations, one for each subject $i = N$. Notice that we have to estimate $N + 1 + (a - 1)$ parameters, where a is the number of groups. The number of parameters to be estimated is $a - 1$, and not a because all the a's sum to zero, so if we know any three the fourth is predictable.

As before, since we want a unique solution, we guess a value of $\mu + \alpha$ that is as close as possible to Y_{ij}. So, as usual, we minimize ϵ_{ij}:

$$\epsilon_{ij\mathcal{F}} = \sum_{j} \sum_{i} [Y_{ij} - (\hat{\mu} + \alpha_j)]^2 \tag{88}$$

How to estimate α? Notice that

$$\hat{\mu} = \frac{\sum\limits_{j=1}^{a} \bar{Y}_j}{a} = \bar{Y}_u \tag{89}$$

where $\bar{Y}_u$ is the unweighted mean.

Since $\bar{Y}_j = \hat{\mu} + \alpha_j$, it follows that

$$\alpha_j = \bar{Y}_j - \hat{\mu} \tag{90}$$

or (replacing $\hat{\mu}$):

$$\alpha_j = \bar{Y}_j - \frac{\sum_{j=1}^{a} \bar{Y}_j}{a} \tag{91}$$

The restricted model's degrees of freedom are $N - 1$, and the full model's are $N - (a - 1) + 1) = N - a$.

When we have equal n in each group,

$$E_{\mathcal{R}} - E_{\mathcal{F}} = \sum_j \sum_i \hat{\alpha}_j{}^2 \tag{92}$$

$$= n \sum_j \hat{\alpha}_j{}^2 \tag{93}$$

With unequal n:

$$E_{\mathcal{R}} - E_{\mathcal{F}} = \sum_j \sum_i (Y_j - \hat{\mu})^2 \tag{94}$$

$$= n_j (Y_j - \hat{\mu})^2 \tag{95}$$

$$= n_j (\hat{\alpha})^2 \tag{96}$$

Now it's straightforward to compute F, but the reason this technique was introduced here is that it's very useful in within subject designs' analyses. That's discussed in the next section.

5 Within subjects, two level designs

Suppose we did an experiment involving one factor with two levels. Recall the model for one way between subjects:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij} \tag{97}$$

A problem here is that for any given subject the errors in group 1 and 2 are likely to be correlated: a subject who gives a high score in one group is likely to give a high score in the other. This is a problem because ANOVA assumes independence of errors, and here they're anything but.

What to do? Notice that the problem is that we have two errors per subject, and they're correlated. If we could get only one error per subject, the correlated-error problem is gone. We can rewrite (97) as:

$$Y_{i1} = \mu + \alpha_1 + \epsilon_{i1} \tag{98}$$

$$Y_{i2} = \mu + \alpha_2 + \epsilon_{i2} \tag{99}$$

Subtracting (99) from (98), we get:

$$Y_{i2} - Y_{i1} = \alpha_2 - \alpha_1 + \epsilon_{i2} - \epsilon_{i1} \tag{100}$$

and this is our full model ($M_{\mathcal{F}}$):

$$D_i = \mu + \epsilon_i \tag{101}$$

Our null hypothesis earlier was stated as $\alpha_1 = \alpha_2 = 0$, but now it is:

$$H_0 = \mu = 0 \tag{102}$$

and our restricted model $M_{\mathcal{R}}$ is:

$$D_i = 0 + \epsilon_i \tag{103}$$

or simply

$$D_i = \epsilon_i \tag{104}$$

Now we compute the terms of our F-statistic. The LSE of μ is $\bar{D}$.

$$E_F = \sum_i (D_i - \bar{D})^2 \tag{105}$$

$$E_R = \sum_i (D_i - 0)^2 = \sum_i D_i^2 \tag{106}$$

What's $E_R - E_F$? Obviously:

$$E_R - E_F = \sum_i (D_i - \bar{D})^2 - \sum_i D_i^2 \tag{107}$$

Perhaps less obviously (as an exercise, try proving the equality below), this reduces to:

$$E_R - E_F = n\bar{D}^2 \tag{108}$$

Perhaps you should think about what the degrees of freedom are for the full and restricted models here.

Now, we're ready to compute F:

$$F = \frac{n\bar{D}^2/n - (n-1)}{\sum_i (D_i - \bar{D})^2/(n-1)} \tag{109}$$

$$= \frac{n\bar{D}^2}{s_D^2} \tag{110}$$

where

$$s_D^2 = \frac{\sum D_i^2 - n\bar{D}^2}{n-1} \tag{111}$$

is the unbiased estimate of the population variance of the D scores.

Notice that F could be rewritten as:

$$t = \frac{\sqrt{n}\bar{D}}{s_D} \tag{112}$$

This is the well-known formula for a dependent t-test. With two levels of the repeated factor, the model-comparisons test reduces to the dependent t-test.

6 R example for within-subjects designs

This is an example from Hays' book (1988, Table 13.21.2, p. 518) and was used in the Baron and Li notes on CRAN. A 2×2 within subjects design. We begin by setting up the data:

```
> data1 <- c(49, 47, 46, 47, 48, 47, 41, 46, 43, 47, 46, 45, 48,
+     46, 47, 45, 49, 44, 44, 45, 42, 45, 45, 40, 49, 46, 47, 45,
+     49, 45, 41, 43, 44, 46, 45, 40, 45, 43, 44, 45, 48, 46, 40,
+     45, 40, 45, 47, 40)
> Hays.mul.df <- as.data.frame(matrix(data1, ncol = 4, dimnames = list(paste("subj",
+     1:12), c("Shape1.Color1", "Shape2.Color1", "Shape1.Color2",
+     "Shape2.Color2"))))
> Hays.df <- data.frame(rt = data1, subj = factor(rep(paste("subj",
+     1:12, sep = ""), 4)), shape = factor(rep(rep(c("shape1",
+     "shape2"), c(12, 12)), 2)), color = factor(rep(c("color1",
+     "color2"), c(24, 24))))
```

The ANOVA call in R gives you the following output:

```
> anova.fm <- aov(rt ~ shape * color + Error(subj/(shape * color)),
+     data = Hays.df)
> summary(anova.fm)

Error: subj
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 11  226.5  20.591

Error: subj:shape
          Df Sum Sq Mean Sq F value  Pr(>F)
shape      1   12.0 12.0000  7.5429 0.01901 *
Residuals 11   17.5  1.5909
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: subj:color
          Df Sum Sq Mean Sq F value   Pr(>F)
color      1   12.0 12.0000  13.895 0.003338 **
Residuals 11    9.5  0.8636
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: subj:shape:color
            Df Sum Sq Mean Sq   F value Pr(>F)
shape:color  1    0.0  0.0000 4.495e-28      1
Residuals   11   30.5  2.7727

> coefficients(anova.fm)

(Intercept) :
(Intercept)
```

```
         45

subj :
numeric(0)

subj:shape :
shapeshape2
         -1

subj:color :
colorcolor2
         -1

subj:shape:color :
shapeshape2:colorcolor2
           2.038340e-14
```

Now we compute the ANOVA "by hand", using the equations worked out in this chapter.

First we compute the Sum of Squares within:

```
> c1 <- Hays.mul.df$Shape1.Color1
> c2 <- Hays.mul.df$Shape1.Color2
> c3 <- Hays.mul.df$Shape2.Color1
> c4 <- Hays.mul.df$Shape2.Color2
```

Now we look at the main effect of shape using the formula for F that we just derived in the sections above.

```
> Shape1 <- (c1 + c2) * 0.5
> Shape2 <- (c3 + c4) * 0.5
> DShape <- Shape2 - Shape1
> SumSqShape <- sum((mean(DShape) - DShape)^2)
> sdShape <- sd(DShape)
> n <- 12
> barD <- mean(DShape)
> (F <- (n * (barD^2))/(sdShape^2))
```

```
[1] 7.542857
```

Notice that the F value is exactly what R's ANOVA gives us. Now let's look at the main effect of color:

```
> Color1 <- (c1 + c3) * 0.5
> Color2 <- (c2 + c4) * 0.5
> DColor <- Color2 - Color1
> sum((mean(DColor) - DColor)^2)
```

```
[1] 9.5
```

```
> sdColor <- sd(DColor)
> n <- 12
> barD <- mean(DColor)
> (F <- (n * (barD^2))/(sdColor^2))
```

```
[1] 13.89474
```

Again, we get the F value that R gives us. Finally, look at Color and Shape interaction:

```
> Shapes <- (c1 - c2) * 0.5
> Colors <- (c3 - c4) * 0.5
> DSC <- (Shapes - Colors)
> sum((mean(DSC) - DSC)^2)

[1] 30.5

> sdDSC <- sd(DSC)
> n <- 12
> barD <- mean(DSC)
> (F <- (n * (barD^2))/(sdDSC^2))

[1] 0
```

The F-value for the interaction in the R code is not exactly zero but it's close enough. If you can't guess the reason why R would give a non-zero number, don't worry about it. It's only important to note that R-s F-value, 6.947e-29, is essentially 0.

Problems

Prove that $F = \frac{n\bar{D}^2}{s_D^2}$.

References

Maxwell, S. E., & Delaney, H. D. (2000). Designing experiments and analyzing data. Mahwah, New Jersey: Lawrence Erlbaum Associates.