Overview of the ANOVA Table for Comparing Means

Corresponding to Section 10.2 - *B. Habing, 11/20/03*

Say we are comparing the means of p different groups treatment group where each group has n_i observations.
We'll let the total number of observations be $n = n_1 + n_2 + \ldots + n_p$ then.
Because we have to number both the treatment group and the observation number, we need to use two subscripts.
The first subscript (i) is for the group number and the second (j) is for which observation it is.
So that the notation matches that for regression we'll use y_{ij} to name each observation then... but you could use x_{ij} like the book does if you wanted to.

The data could be laid out as follows:

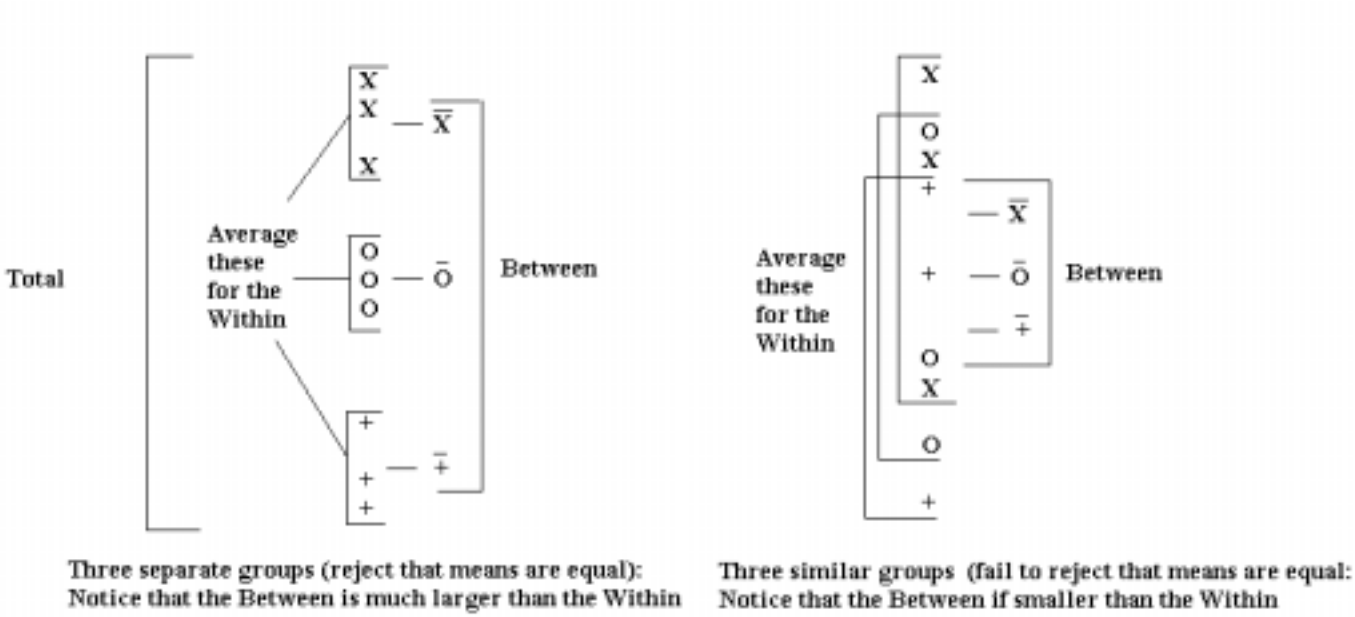
Treatment

	$i=1$	$i=2$	•••	$i=p$	
	y_{11}	y_{21}		y_{p1}	
	y_{12}	y_{22}		y_{p2}	Overall Mean ▼
	⋮	⋮	•••	⋮	
	y_{1n_1}	y_{2n_2}		y_{pn_p}	
Means for ► Each Group	$\bar{y}_1$	$\bar{y}_2$	•••	$\bar{y}_p$	$\bar{y}$

If we looked at Example 10.3 on page 452 we would have p=4 (there are four brands). n_1=10 because 10 different balls were used with the first brand. Similarly n_2, n_3, and n_4 are all 10 in this example (they don't all need to be equal in general) and the total sample size n=40. Just as we have laid out in the table above, y_{11}=251.2 (first group, first observation), y_{23}=265.0 (second group, third observation), etc... Similarly $\bar{y}_1$=250.8 (the mean of the first group), etc... To find $\bar{y}$ we would have to take the average of all 40 observations.

The goal of Analysis of Variance is to test the hypothesis that the means of the *p*-groups are equal (see the top of page 448). We cannot simply use a t-test because there is no way to "take the difference" of more than two different means. It just doesn't make sense.

To get around this problem we need to use a different method. Instead of focusing on means we will focus on variances. In particular we will construct two different variances for our situation: a variance between the groups (due to the treatments) and a variance within the groups (due to random error). If we had three groups x's, o's, and +'s we might get a picture like the following.



Three separate groups (reject that means are equal): Notice that the Between is much larger than the Within

Three similar groups (fail to reject that means are equal: Notice that the Between if smaller than the Within

Notice in the case where the groups are clearly different, the between group variance is clearly larger than the within group variance. If we took the variance between divided by the variance within we would get a large number.

In the second case where the groups are similar, the between group variances is clearly smaller than the within group variance. If we took the between group variance divided by the variance within we would get a small number.

Why divide the variances? Because in Supplement 6 we saw that dividing two variances gives us an F distribution!

So what we will do is use $F = \dfrac{\text{variance between groups}}{\text{variance within groups}}$, and will reject the null hypothesis when F is large.

The goal of the ANOVA table is to calculate this F statistic.

The ANOVA Table for Comparing Means

Source	SS *(Sum of Squares, the numerator of the variance)*	DF *(the denominator)*	MS *(Mean Square, the variance)*	F
Treatment *(or Between or Model)*	$SST = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2$	*p-1*	$MST = \frac{SST}{p-1}$	$F = \frac{MST}{MSE}$
Error *(or Within)*	$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	*n-p*	$MSE = \frac{SSE}{n-p}$	
Total	$TSS = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$	*n-1*		

Notice that if we took the TSS/*n*-1 we would end up with just the overall variance of all the observations. (Compare the formula you get here to our usual formula for s^2... it only looks different because we had to use two subscripts to make sure to include all of the observations.)

The five keys to remember about the Analysis of Variance table are:

1) The sum of squares add up: *SST* + *SSE* = *TSS*

2) The degrees of freedom can be calculated from the sum of squares formulas. Looking at the SST notice that there are *p* different $\bar{y}_i$ and one $\bar{y}$ so we get *p*-1 degrees of freedom. Looking at the SSE we have *n* different y_{ij} and *p* different $\bar{y}_i$ and so the degrees of freedom are *n-p*. Finally, for TSS there are *n* different y_{ij} and one $\bar{y}$ so there are *n-1* degrees of freedom.

3) The degrees of freedom add up: (*p*-1) + (*n*-*p*) = *n*-1

4) The mean squares (the variances) are found by taking the sum of squares (the numerator) and dividing by the degrees of freedom (the denominator).

5) The *F* statistic is calculated by dividing the two MS.