Notes on ANOVA

Dr. McIntyre
McDaniel College

Revised: August 2005

1. One-way ANOVA

1.1. The ad-hoc method

1. We have previously been performing hypothesis tests that concern the difference between two means. We now wish to generalize this type of test to one involving $k > 2$ different means. For example, we might want to ask if Fords, Chevys, and Jeeps all have the same fuel effieiency in a statistical sense, or if five GAP employees sell the same amount of merchandise over a given period of time, etc.

2. The mechanics of answering this question involves decomposing the variance that exists in a combined data set into two components: the variance that exists across samples and the variance that exists within samples. If the former is larger (in a statistical sense), it is evidence that the underlying population means are equal. The hypothesis test is:

    1. H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$
       H_A: not all μ_is are equal.
    2. $\alpha = \ldots$usually .10, .05, .01.
    3. When our k samples are of uniform size, n, constructing the test statistic is fairly easy.
        1. Suppose each sample i has sample mean $\overline{x}_i$ and sample variance s_i^2. Define the grand mean, $\overline{\overline{x}}$, as:

$$\overline{\overline{x}} = \frac{1}{k}\sum_{i=1}^{k}\overline{x}_i.$$

Note that since all sample sizes are equal, we can ignore weighting. Next, define the variance of the sample means, $s_{\overline{x}}^2$ as:

$$s_{\overline{x}}^2 = \frac{1}{k-1} \sum_{i=1}^{k} \left(\overline{x}_i - \overline{\overline{x}} \right)^2 .$$

This is our measure of the variance that exists across samples (.i.e variance due to differences in sample/population means).

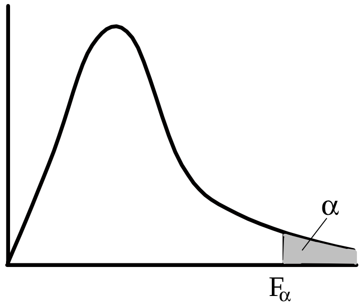
2. Our measure of within-sample variance is just the average sample variances, $\overline{s}^2$:

$$\overline{s}^2 = \frac{1}{k} \sum_{i=1}^{k} s_i^2 .$$

3. Now all we need to do is construct our test statistic, F:

$$F = \frac{n s_{\overline{x}}^2}{\overline{s}^2} .$$

Under H_0, this statistic follows an F-distribution. The F-distribution is a skewed distribtion that looks like this:



There are two parameters—again called degrees of freedom—we need to be aware of when working with the F-distribtion. For this example, F has $k-1$ df associated with the numerator and and $k\left(n-1\right)$ df associated with the denomiator.

4. The rule is to reject H_0 when $F > F_\alpha$ with $k-1$ and $k(n-1)$ degrees of freedom.

1.2. The formal one-way paradigm, equal sample sizes

1. Suppose again we have k samples of size n, so there are $N = kn$ total data points. From here, the formal ANOVA procedure begins with a measure of the the total variance within our set of samples, called the total sum of squares or SSS:

$$SSS = \sum_{j=1}^{k} \sum_{i=1}^{n} \left(x_{ij} - \overline{\overline{x}}\right)^2 .$$

In the above expression, $\overline{\overline{x}}$ denotes the grand mean. This formula looks much nastier than it is; all it is telling us to do is to calculate the grand mean, which is easy enough, subtract it from every observation we have, square them, and add them up.

2. We want to split this "total variance" up into the variance across and within samples. Some brutal algebra allows us to do this. Begin by adding and subtracting $\overline{x}_j$ within the squared term, as follows:

$$SSS = \sum_{j=1}^{k} \sum_{i=1}^{n} \left(x_{ij} - \overline{\overline{x}}\right)^2 = \sum_{j=1}^{k} \sum_{i=1}^{n} \left[\left(x_{ij} - \overline{x}_j\right) + \left(\overline{x}_j - \overline{\overline{x}}\right)\right]^2 .$$

Next, expand the squared term:

$$SSS = \sum_{j=1}^{k} \sum_{i=1}^{n} \left[\left(x_{ij} - \overline{x}_j\right)^2 + 2\left(x_{ij} - \overline{x}_j\right)\left(\overline{x}_j - \overline{\overline{x}}\right) + \left(\overline{x}_j - \overline{\overline{x}}\right)^2\right] .$$

Distribute the summation operators:

$$SSS = \sum_{j=1}^{k} \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2 + 2\sum_{j=1}^{k} \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)\left(\overline{x}_j - \overline{\overline{x}}\right) + \sum_{j=1}^{k} \sum_{i=1}^{n} \left(\overline{x}_j - \overline{\overline{x}}\right)^2 .$$

Now, recall that deviations from the mean always sum to zero, so the middle term in the above expression is equal to zero. Also, note that j does not index anything in the third term above. Making these simplifications yields:

$$SSS = \sum_{j=1}^{k} \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2 + n\sum_{j=1}^{k} \left(\overline{x}_j - \overline{\overline{x}}\right)^2 = SSE + SST,$$

and we are done.

   1. The first term in the above equation is called the error sum of squares, or SSE. It is our measure of the variance that exists within samples. Note what the SSE term is telling us to do. Subtract the sample mean from each element is sample i, square them, and add them up, a procedure very similar to calculating a sample variance.
   2. The second term is the treatment sum of squares, or SST. It is a measure of the variance across samples. Note again how SST is calculated: subtract the grand mean from each sample mean, square them, and add them up.

3. We are two intermediate definations away from being able to calculate an F-statistic. Define the mean squared error, or MSE, as follows:

   $$MSE = \frac{SSE}{N-k};$$

   the MSE is just the error sum of squares divided by its degrees of freedom, $k(n-1)$. Likewise, if we divide the treatment sum of squares by its degrees of freedom, $k-1$, we have the treatment mean square, MST:

   $$MST = \frac{SST}{k-1}.$$

4. From here, our F-statistic is just the ratio of mean squares:

   $$F = \frac{MST}{MSE}.$$

   This statistic follows an F-distribution with $k-1$ and $N-k$ degrees of freedom, respectively.

5. An easy way to keep everything straight when calculating your F-statistic is to fill in an ANOVA table as you go along:[1]

[1] The ad-hoc and formal ANOVA paradigms are equivalent. To see this note that:

$$\begin{aligned} MST &= \frac{SST}{k-1} = \frac{n\left(\overline{x}_1 - \overline{\overline{x}}\right)^2 + n\left(\overline{x}_2 - \overline{\overline{x}}\right)^2 + \cdots + n\left(\overline{x}_k - \overline{\overline{x}}\right)^2}{k-1} \\ &= n\left[\frac{\left(\overline{x}_1 - \overline{\overline{x}}\right)^2 + \left(\overline{x}_2 - \overline{\overline{x}}\right)^2 + \cdots + \left(\overline{x}_k - \overline{\overline{x}}\right)^2}{k-1}\right] = ns^2_{\overline{X}}, \end{aligned}$$

Source	df	Sum of Squares	Mean Square	F
Treatments	$k-1$	SST	$MST = SST/(k-1)$	MST/MSE
Error	$N-k$	SSE	$MSE = SSE/(N-k)$	
Total	$N-1$	$SST+SSE$		

1.3. The formal one-way paradigm, unequal sample sizes

1. When all sample sizes are not the same, though, ANOVA problems are considerably more labor intensive. The use of a common spreadsheet program like MS Excel, however, makes these problems very manageable. (And in 2216, you will learn how to use statistical software that will perform all calculations for you.) This handout is designed to lead you through this process.

2. Suppose that, as usual, we have k samples. The sample sizes are not equal, however, which forces us to index our sample sizes as follows: n_1, n_2, $n_3, \ldots,$ n_k. There are N total data points; $n_1 + n_2 + \cdots + n_k = N$. With unequal sample sizes, we must use the formal ANOVA formula, modified slightly (note we are now subscripting our n's by the sample j):

$$SSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \overline{\overline{x}}\right)^2 .$$

3. As usual, we can break up our total sum of squares into the treatment sum of squares, SST, and the error sum of squares, SSE. The derivation is as follows:

    1. First, add/subtract $\overline{x}_j$ inside the square term:

$$SSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \overline{\overline{x}}\right)^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left[\left(x_{ij} - \overline{x}_j\right) + \left(\overline{x}_j - \overline{\overline{x}}\right)\right]^2 .$$

and:

$$\begin{aligned} MSE &= \frac{SSE}{N-k} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2}{k\,(n-1)} = \frac{\sum_{j=1}^{k} \left[\frac{1}{n-1} \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2\right]}{k} \\ &= \frac{\sum_{j=1}^{k} \left(s_j^2\right)}{k} = \overline{s}^2. \end{aligned}$$

2. Next, expand the sum:

$$\begin{aligned} SSS &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{\overline{x}}\right)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left[\left(x_{ij}-\overline{x}_j\right)+\left(\overline{x}_j-\overline{\overline{x}}\right)\right]^2 \\ &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left[\left(x_{ij}-\overline{x}_j\right)^2+2\left(x_{ij}-\overline{x}_j\right)\left(\overline{x}_j-\overline{\overline{x}}\right)+\left(\overline{x}_j-\overline{\overline{x}}\right)^2\right] \end{aligned}$$

3. Distribute the summation operators:

$$\begin{aligned} &\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left[\left(x_{ij}-\overline{x}_j\right)^2+2\left(x_{ij}-\overline{x}_j\right)\left(\overline{x}_j-\overline{\overline{x}}\right)+\left(\overline{x}_j-\overline{\overline{x}}\right)^2\right] \\ = &\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x}_j\right)^2+2\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x}_j\right)\left(\overline{x}_j-\overline{\overline{x}}\right)+\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\overline{x}_j-\overline{\overline{x}}\right)^2 \end{aligned}$$

Note the $\left(\overline{x}_j-\overline{\overline{x}}\right)$ term. You may recall this as the deviations from the mean. This term sums to zero, hence the entire middle term drops. We are left with the usual formula:

$$SSS = SST + SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\overline{x}_j-\overline{\overline{x}}\right)^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x}_j\right)^2.$$

As far as simplifying algebra goes, we are done (with equal sample sizes we were able to further simplify SST; refer to the previous section).

4. Plugging real numbers into these expressions is not as scary as it appears to be. Consider the following points:

   - Calculate SSS—I'll explain why we are doing this directly shortly—by first calculating the grand mean $\overline{\overline{x}}$. Then just subtract $\overline{\overline{x}}$ from every data point you have (all N of them), square it, and add them all up.
   - Calculate SSE by calculating each sample variance, s_j^2. Multiply that by its respective sample size less one, $n_j - 1$, and add them up. In formulas,

$$SSE = \sum_{j=1}^{k}\left(n_j-1\right)s_j^2.$$

- Calculate SST by subtracting SSE from SSS. That is, calculate $SST = SSS - SSE$. Of the three sums in the ANOVA formula, SST is the toughest to calculate, which is why we often "back it out" like this. In practice, however, it is not that bad. For example, note that this sum can be expanded as follows:

$$SST = n_1 \left(\overline{x}_1 - \overline{\overline{x}}\right)^2 + n_2 \left(\overline{x}_2 - \overline{\overline{x}}\right)^2 + \cdots + n_k \left(\overline{x}_k - \overline{\overline{x}}\right)^2,$$

so, all you need to do is subtract $\overline{\overline{x}}$ from each sample mean, square it, multiply by the respective sample size, and add them all up.

5. Then, all you need to do is fill in your ANOVA table to find your value for F. Note that $MST = \frac{SST}{k-1}$ and $MSE = \frac{SSE}{N-k}$. F will then equal MST/MSE, and will follow an F-distribution with $k-1$, and $N-k$ degrees of freedom respectively. In symbols,

$$F = \frac{MST}{MSE} \sim F_{k-1,N-k}.$$

6. If $F > F_\alpha$ for your chosen level of significance, reject H_0. Here is another ANOVA table with all of the fomulas in it:

Source	df	Sum of Squares	Mean Square	F
Treatments	$k-1$	SST	$MST = SST/(k-1)$	MST/MSE
Error	$N-k$	SSE	$MSE = SSE/(N-k)$	
Total	$N-1$	$SST + SSE$		

2. Two-way ANOVA

1. This move involved ANOVA technique allows one to consider simultaneously **two** sources of non-chance variation among population means. Let me motivate this notion by extending a familiar example: suppose we again interested in the gas mileage of three different makes of cars, Jeeps, Mustangs, and Camaros. Now, each make of car was driven by two different drivers, and each driver got five tanks of gas per car. In this example, there are four potential sources of variation between gas mileage:

   - Make of car. Call this *factor A*.
   - Driver. Call this *factor B*.

- An *interaction* between car and driver. We can call this *factor AB*.
- Chance/error.

2. To consider this problem we will again be decomposing our variance according to the ANOVA formula:

$$SSS = SST + SSE,$$

only now the treatment sum of squares needs to be further decomposed:

$$SST = SS_A + SS_B + SS_{AB}.$$

Substituting this in gives us the formal two-way ANOVA formula:

$$SSS = SS_A + SS_B + SS_{AB} + SSE.$$

3. At this point, we need to introduce some additional terminology and notation.

   1. Each factor will have two or more *levels*. In the example above, factor A has three levels (Jeep, Mustang, Camaro), and factor B has two (driver #1, driver #2). In general, factor A will have a levels and factor B will have b levels.
   2. The r data points for each combination of factor A and factor B will be organized in a *cell*.
   3. There will be ab cells in all, and $N = abr$ total data points. (Pages 436 of the Keller text contain examples of how the data tables for two-way ANOVA are laid out.)

4. The computational formula for the total sum of squares is as follows:

$$SST = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{r} \left(x_{ijk} - \overline{\overline{x}} \right)^2,$$

where, i indexes the levels in factor A, j indexes the levels in factor B, and k indexes the date point within a cell.

5. This triple sum is really nasty looking, but it is really not that bad.

1. Calculating SSS is actually pretty tame: just as before calculate the grand mean of your combined data set. From there, subtract $\overline{\overline{x}}$ from each of your N total data points, square it, and add them up.
2. The main-effect sums of squares formulas are:

$$\begin{aligned} SS_A &= br\sum_{i=1}^{a}\left(\overline{x}_i - \overline{\overline{x}}\right)^2, \\ SS_B &= ar\sum_{j=1}^{b}\left(\overline{x}_j - \overline{\overline{x}}\right)^2. \end{aligned}$$

   Evaluating these formulas with actual data is not so bad, either. First, calculate the mean for each factor's levels. For the gas mileage example, the level means for factor A are the average gas mileage for the Jeep, Mustang, and Camaro regardless of driver. For level means for factor B are the average gas mileage for driver #1 and driver #2 regardless of what car each drives.To calculate SS_A, for example, subtract the grand mean from the a level means for factor A, square them, and add them up, and multiply by bn. SS_B is found similarly.
3. Skipping SS_{AB} for a moment, the formula for SSE is:

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(x_{ijk} - \overline{x}_{ij}\right)^2.$$

   Like the total sum of squares formula, this triple sum is looks considerably worse than it is. To calculate this, first calculate each of your cell means; there will be ab of them. (This is the $\overline{x}_{ij}$ part of the above formula.) Then, subtract the cell mean from each of the r elements in that cell, square it, and add them up. Repeat for all cells, and then add all of those squared terms together.
4. Returning to SS_{AB}, thus sum of squares is the most difficult to calculate directly (although it is not that bad). Its formula is:

$$SS_{AB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{x}_{ij} - \overline{x}_i - \overline{x}_j + \overline{\overline{x}}\right)^2.$$

   To evaluate this, subtract from each cell mean its respective level means (two of them) and add the grand mean, square it, and add them up.

For our car/driver example, if $\overline{x}_{ij}$ was the mean for the five observations for the Jeep driven by driver #1, we would subtract from that the mean gas mileage for the Jeep regardless of driver, and the mean gas mileage for driver #1 regardless of vehicle, add the grand mean, and square it. Then, we would do a similar exercise for every other cell, and add them up. We could also "cheat" and back out SS_{AB} as follows:

$$SS_{AB} = SST - SS_A - SS_B - SSE.$$

6. From here, we fill in a two-way ANOVA table:

Source	df	Sum of Squares	Mean Squares	F
Factor A	$a-1$	SS_A	$MS_A = SS_A/(a-1)$	$F_A = MS_A/MSE$
Factor B	$b-1$	SS_B	$MS_B = SS_B/(b-1)$	$F_B = MS_B/MSE$
Factor AB	$(a-1)(b-1)$	SS_{AB}	$MS_{AB} = SS_{AB}/\left[(a-1)(b-1)\right]$	$F_{AB} = MS_{AB}/M$
Error	$ab(r-1)$	SSE	$MSE = SSE/\left[ab(m-1)\right]$	
Total	$N-1$	SSS		

Again, the mean squares are just the sums of squares divided by their degrees of freedom. The F-statistics are the mean squares for each treatment divided by the mean square error. The degrees of freedom for each of the three F-statistics are given by the mean squares. That is, the degrees of freedom for factor B's F-statistic are $b-1$ and $ab(r-1)$ respectively. A significant value for F_A or F_B allows one to reject the null hypothesis H_O: $\mu_1 = \mu_2 = \cdots = \mu_k$, for $k = \{a \text{ or } b\}$, the interpretation being that a statistically significant difference in population means is attributable to the factor in question. Likewise, a significant value of F_{AB} means that the interaction between factors A and B is causing a statistically significant difference in population means.