ANOVA Study Guide and Contrast Examples

The Data and ANOVA Table

This handout gives the basic interpretations and formulas for ANOVA and also two examples of contrasts based an example with smoking data. The response variable, Y, is lung capacity as measure by FEF or forced expiratory flow in Liters per second (higher is better), and the group variable is smoking status. There are k=6 groups, NS = nonsmokers, PS = passive smokers, NI = non-inhalers, LS = light smokers, MS = medium smokers, and HS = heavy smokers. The group means, standard deviations, and an ANOVA table are given below:

Group	Mean FEF	SD FEF	Group Size
NS	3.78	.79	200
PS	3.30	.77	200
NI	3.32	.86	50
LS	3.23	.78	200
MS	2.73	.81	200
HS	2.59	.82	200

Source	df	SS	MS	F
Between (Treatment)	5	184.38	36.875	58.0
Within (Error)	1044	663.87	.636	
Total	1049	848.25		

Interpreting Elements of the ANOVA Table

The ANOVA table for comparing means looks just like the ANOVA table in a regression. This is because ANOVA is actually a special case of regression where all the predictor variables are indicators saying what group you are in. Instead of a row for regression and a row for errors we have a row labeled "between" or "treatment" and a row labeled "within". Between is shorthand for variability **between groups**–it measures how far apart the group means are from each other. Within is shorthand for variation **within groups**–it measures how much points vary about their individual group mean. The basic idea is that if there is more between group variation than within group variation then you can tell the means are different. Let n be the total number of data points and k be the number of groups. Then the degrees of freedom for "between groups" is k-1. As we will see later, this is the number of indicator variables you would need, thinking of this as a regression problem, to describe the k groups (you use the remaining group as the reference). The degrees of freedom for within the groups is n-k–in order to compute the variation of points around their individual group means you first have to calculate the k group means. FInally the total degrees of freedom is, as usual, n-1. In order to compute variation about the overall mean you first have to estimate the overall mean. In the column for sums of squares we have SSB, SSW and SST:

SST is the total variability in Y. In our example it is thte total variability in lung capacity (as measured by forced expiratory flow) for the people in our sample. It will serve as our reference point for how good a job the model does of explaining lung capcity. The value 848.25 is hard to intperpret in practical terms.

SSB is the total variability in Y that is explained by knowing which group you belong to. It is the equivalent of SSR in a regression. Essentially it measures how far the group means are from the overall mean. The bigger it is the more sure we can be that the means are different. In our example, SSB is the variability in lung capacity account for by knowing whether you are a non-smoker, passive smoker, non-inhaler, etc. Our

value of 184 means we have explained a little under 25% of the variability in lung capacity by knowing how much a person smokes. This is pretty good considering how many things there are that affect your lung capacity including size, age, and gender.

SSW is the total variability in Y that is not explained by knowing which group you are in. It is the equivalent of SSE. It describes how much the points vary about their individual means. In our example it tells us how much of the variability in lung capcity is due to something other than a person's smoking status. Our value of 663 tells us there is a lot of variability in lung capcity due to other factors which is not a surprise.

Next comes the column for mean squares which as usual are obtained by dividing the sums of squares by their degrees of freedom, and F which is obtained by taking the ratio of the mean squares.

MSB is the mean squares between groups. Basically this is the average amount of variability explained per group. In our example it is the average amount of variation in lung capcity explained per smoking status category. It is the equivalent of MSR.

MSW is the mean squares within groups. It estimates the variance of points about their individual group means. Since we assume in an ANOVA that all the groups have the same variabnce, it is also the pooled estimate of the variance using all the groups. The square root of this value appears in every standard error formula in ANOVA. It is the equivalent of MSE and its square root is the equivalent or RMSE. In our example it is the variance of lung capacity within the different smoking groups. You can also think of it as the average squared error we would make when using the smoking group means to predict peoples' lung capacities. The square root puts this in better units. Your value of .636 has a square root of around .8, meaning we make an average error of about .8 liters per second when using the smoking group means to predict people's FEF score. This may not seem that good but we wouldn't really expect to get accurate predictions absed on just smoking status since there are som many other factors that affect lung capacity.

F = MSB/MSW is the ratio of explained to unexplained variability. In general, the bigger it is the better the job the model does of explaining variability in Y. Here the special meaning is that the bigger F is, the more widely separated the group means are relative to the within group variability and hence the more sure we can be that at least some of the group means differ. In our example F = 58 which is enormous–the variation between groups (adjusted for degrees of freedom) is 58 times that of the variation within groups (adjusted for its degrees of freedom). Smoking group status definitely explains some of the variation in lung capcity. Formally we measure how big F is by looking at the corresponding p-value. The smaller the p-value, the more sure we are the group means are different. In this example the p-value is less than .001.

Major Concepts and Formulas

In an ANOVA you are interested in the individual group means and confidence intervals for them, the overall F test to see whether or not all the means are equal, and in doing confidence intervals and hypothesis tests for differences between individual means. We have the following useful formulas. Recall that n is the total number of data points, k is the total number of groups, Y_{ij} is the value of Y for the ith subject in group j, n_j is the number of subjects in group j, $\bar{Y}_j$ is the mean of group j, s_j^2 is the sample variance of group j, and $\bar{Y}$ is the average of all the data points.

$$SSB = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_j - \bar{Y})^2 = \sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2 = \sum_{j=1}^{k} n_j\bar{Y}_j^2 - \frac{(\sum Y_{ij})^2}{n} = \sum_{j=1}^{k} n_j\bar{Y}_j^2 - n(\bar{Y})^2$$

$$SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2 = \sum_{j=1}^{k}(n_j - 1)s_j^2$$

$$MSB = \frac{SSB}{k-1}$$

$$MSW = \frac{SSW}{n-k}$$

To get confidence intervals you use a t distribution with n-G degrees of freedom and a standard error based on MSW. For a single mean it is

$$\bar{Y}_j \pm t_{\alpha/2,n-k}\sqrt{\frac{MSW}{n_j}}$$

For the difference of two means you get

$$\bar{Y}_1 - \bar{Y}_2 \pm t_{\alpha/2,n-k}\sqrt{MSW(\frac{1}{n_1}+\frac{1}{n_2})}$$

The test statistic for a t-test comparing two means is

$$t_{obs} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{MSW(\frac{1}{n_1}+\frac{1}{n_2})}}$$

STATA Printouts

STATA gives two kinds of printouts in ANOVA. The first type, "oneway" gives you the ANOVA table and group means and standard deviations. The second type, "anova" gives you the ANOVA table (though in a slightly different form) and allows you to do tests comparing groups of means.

Linear Combination and Contrast Examples

A **Linear combination** is just an expression involving the group means in which you might be interested. The basic form is

$$L = \sum_{j=1}^{k} c_j \mu_j$$

where the μ_j are the group means and the c_j are constants. For instance, if we are interested in studying the difference in lung capacity between non-smokers and heavy smokers we want

$$L = \mu_{NS} - \mu_{HS} = \mu_1 - \mu_6$$

This is a linear combination with $c_1 = 1$ and $c_6 = -1$. In the case that the c_j's sum to 0, as in this example, the linear combination is called a **contrast** because it contrasts group 1 with group 6. The best estimate for a linear combination is to plug in the sample means for each group:

$$\hat{L} = \sum_{J=1}^{k} c_j \bar{Y}_j$$

It turns out we can also get an expression for the standard error of a contrast:

$$\sqrt{MSW \sum_j \frac{c_j^2}{n_j}}$$

Using these two expressions we can obtain confidence intervals for or perform hypothesis tests about any linear combination of interest. We will use a t-distribution with n-k degrees of freedom, corresponding to MSW which is at the heart of our standard error formula. For example, suppose we want to test whether the smokers who do not inhale (NI) are different from smokers who do inhale (LS, MS,HS). The appropriate contrast to compare the average for non-inhalers versus that for people who do inhale is

$$L = \mu_{NI} - \frac{\mu_{LS} + \mu_{MS} + \mu_{HS}}{3}$$

Note that we have $c_1 = 1$ and $c_4 = c_5 = c_6 = -1/3$. Our best estimate of the contrast is

$$\hat{L} = 3.32 - \frac{3.23 + 2.73 + 2.59}{3} = .47$$

The associated standard error is

$$\sqrt{.636(\frac{1}{50} + \frac{(-1/3)^2}{200} + \frac{(-1/3)^2}{200} + \frac{(-1/3)^2}{200})} = .117$$

Suppose we want to test whether the lung capacity for non-inhalers is higher than that for inhalers. Then we want a one-sided test with

$H_0 : L \leq 0$–non-inhalers have a lower or the same average lung capacity as inhalers.
$H_A : L > 0$–non-inhalers have a higher lung capcity than inhalers.

Our test statistic is

$$t_{obs} = \frac{.47 - 0}{.117} = 4.00$$

This has a t-distribution with 1044 degrees of freedom, but since the degrees of freedom are so high it is essentially the same as a Z distribution. Thus our p-value is

$$P(Z \geq 4.00) = .000032$$

I got the exact p-value from STATA–from the Z table the best you could say is that it is less than .0001. Note that since this is a 1-sided test there is no doubling of the p-value. Since my p-value is miniscule, I reject the null hypothesis and conclude that non-inhalers have a higher average lung capacity than inhalers.