Analysis of Variance (ANOVA)

Alan B. Gelder
06E:071, The University of Iowa[1]

1 Theoretical Introduction

An unlikely name: The analysis of variance (or simply ANOVA) may seem like an unlikely name for a statistical technique that is used for comparing several means with each other. Why would we want to analyze variance when we are comparing means? Yet if you step back and think about it, we have been incorporating standard deviations into our means problems and our proportions problems all along; so it should not be too out of place that we are using variances to analyze groups of means. However, why were we using standard deviations in the first place when we studied problems with one or two means or proportions? The answer to that question will give us part of the insight for why examining groups of means is associated with variance. **(Draw pictures comparing means with different variances.)**

1.1 Hypotheses

Recall that a *null* hypothesis implies that there is *no* difference. When we are comparing several means from several populations, the null hypothesis is that all of the means are the same (or that no one mean is different from any of the other means). Formally, the null hypothesis for a problem comparing n means is expressed as follows:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_n$$

The alternative hypothesis is the logical negation (or the opposite) of the null hypothesis. In words, the alternative hypothesis states that at least one of the means is different from the others. We will express this as

$$H_A : \text{There exists } \mu i \text{ and } \mu_j \text{ such that } \mu i \neq \mu_j$$

It could be the case that only one of the means is different. Or perhaps two of the means are different from all of the others. We could even have the case where every mean is distinct from every other mean. However, all that we need in order for H_0 to be incorrect if for there to be at least one pair of means that are different from each other.

[1]The notes for this class follow the general format of Blake Whitten's lecture notes for the same course.

1.2 Introducing the F-Statistic

The F-distribution and its corresponding F-statistic are named after Sir Ronald A. Fisher who developed the techniques necessary to compare several means in a statistically meaningful way. It will be useful to present some intuition about the F-statistic upfront before delving into how it is actually calculated.

$$F = \frac{\text{Variance Between Means}}{\text{Variance Within Samples}} = \frac{\text{Estimate of } \sigma^2 \text{ if } H_0 \text{ is true}}{\text{Baseline estimate of } \sigma^2 \text{ independent of } H_0}$$

In the F-statistic, the numerator is the variance that we would expect to find if H_0 is indeed true. The denominator, however, does not depend on H_0. Rather, the denominator is an estimate of the variance that is derived by pooling all of the samples from all of the different populations that we are comparing. Since the estimate of the variance in the denominator is constant regardless of whether or not H_0 is true, it can be thought of as a *gold standard*—it normalizes our estimate of σ^2 that we obtain assuming that H_0 is correct.

The variance between means measures the variance between the average of each group and the overall average across all groups. The variance within samples is an average of the sample variances which is weighted by the sample size in each group.

One as the magic number: The F-statistic is a ratio of variances. If H_0 is indeed correct, then we would expect that the variance in the numerator and the variance in the denominator would be the same. (**Why is that?**) This would give us an F-statistic of one.

Rejecting H_0: The F-distribution is shaped like the χ^2 distribution; it begins at zero and it has a long tail going to the right (since variances are never negative we cannot get a negative F-statistic). Since an F-statistic of one is what we would expect if H_0 is true, we typically need an F-statistic that is quite a bit bigger than one in order to have enough evidence to reject H_0.

There is no cut and dry baseline for when an F-statistic is significant. This is because the F-statistic has both numerator and denominator degrees of freedom, and the critical values change quite a bit depending on which combination of degrees of freedom you have. With 10 or more denominator degrees of freedom and any num-

ber of numerator degrees of freedom, an F-statistic of 5 is significant for $\alpha = 0.05$.

Interpretation of the F-statistic: The denominator in the F-statistic (our gold standard) normalizes our estimate of the variance assuming that H_0 is true. Hence, if F = 2, then our sample has two times as much variance as we would expect if H_0 were true. If F = 10, then our sample has 10 times as much variance as we would expect if H_0 were true. Ten times is quite a bit more variance than we would expect. In fact, for denominator degrees of freedom larger than 4 and any number of numerator degrees of freedom, we would reject H_0 at the 5% level with an F-statistic of 10.

1.3 Calculating the F-Statistic

Two Parts of Our Data: Consider a problem where we are comparing data from k different groups. Within the ith group are n_i data points. We can think of each data point in group i as having two components: (1) the population mean μ_i and (2) an error which is normally distributed around the mean μ_i with variance σ^2. So for the jth element of group i we have that

$$x_{ij} = \mu_i + \varepsilon_{ij} \text{ where } \varepsilon_{ij} \sim N(0, \sigma^2)$$

Note that σ^2 does not depend on group i. In ANOVA we will assume that each of our samples are drawn from populations that are normally distributed with variance σ^2 (although the means of these distributions can differ from each other). It is also important to note that we do not know μ_i or ε_{ij}. The sample analogue for dividing each datum into two parts is as follows:

$$x_{ij} = \bar{x}_i + e_{ij} \quad \Rightarrow e_{ij} = x_{ij} - \bar{x}_i$$

The *error* or *residual* e_{ij} is how far off a datum is from the group mean. The within group variance focuses on these errors.

Means and the Overall Mean: While the variance within groups concentrates on the errors described above, the variance between means looks at the difference between each group mean $\bar{x}_i$ and the overall mean of all of the data in all of the groups $\bar{x}$:

$$\bar{x}_i - \bar{x}$$

Terminology:
SSG = Sum of Squared difference between Groups
SSE = Sum of Squared Errors within groups

SST = Sum of Squared Total difference from the mean of all the groups

MSG = Mean Squared difference between Groups
MSE = Mean Squared Error within groups

DFG = Degrees of Freedom for the between Group difference
DFE = Degrees of Freedom for the within group Error
DFT = Degrees of Freedom for the Total difference from the mean of all the groups

k = Number of groups
N = Total number of observations in all groups

The ANOVA Table:

Source	DF	Sum of Squares	Mean Square	F
Groups	k-1	$\sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2$	SSG/DFG	MSG/MSE
Error	N-k	$\sum_{i=1}^{k} (n_i - 1) s_i^2$	SSE/DFE	
Total	N-1	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}\right)^2$	SST/DFT	

The ANOVA table developed as a way to both calculate the F-statistic and the check your work along the way. While we often only need the F-statistic and p-value, the complete table has become the standard way of presenting ANOVA calculations. Statistical programs will often provide the p-value in an additional column following the F-statistic.

Sample variance: Let y_1, y_2, ..., y_j be the elements in a sample. Then the sample variance is

$$s^2 = \frac{\sum_{i=i}^{j} \left(y_i - \bar{y}\right)^2}{j - 1}$$

MSG: The MSG is essentially just the sample variance of the means in each of the groups.

$$MSG = \frac{SSG}{DFG} = \frac{\sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2}{k - 1}$$

The only difference between the formula for MSG and the formula for the sample variance s^2 is that each of the squared differences between the group means x_i and

the overall mean $\bar{x}$ is weighted by the number of elements in each group n_i.

MSE: In order to calculate the MSE we first need to find the sample variance for each of the groups. We then pool these groups together as follows:

$$MSE = \frac{SSE}{DFE} = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{N - k} = s_p^2$$

Note that if all of the groups have the same number of elements (i.e. $n_1 = n_2 = ... = n_k = n$), then the MSE simply becomes the average of the sample variances:

$$\frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{N - k} = \frac{(n-1)\sum_{i=1}^{k} s_i^2}{k(n-1)} = \frac{\sum_{i=1}^{k} s_i^2}{k}$$

Connection to Regression: Regression analysis also uses the F-statistic (in fact, there are a couple of different F-statistics that we will study in the context of regression). With some minor changes in terminology, the terms and concepts that we are developing here will continue into our study of regression.

1.4 Tukey Confidence Intervals

Gateway Test: Like χ^2, ANOVA is a gateway test. If we fail to reject H_0 in an ANOVA problem, then there is not sufficient evidence that any one of the means that we are testing is statistically different from any of the other means. We can stop there, go home, and say, "It looks like all of the means are the same." One the other hand, if we reject H_0, then we can safely say that at least one of the means is different from the others. That's all that we can say though. We need to do further investigation to find out which means are different. In addition to finding out which means are different, we may also want to see how the means are ranked. Which mean is highest? Which mean is lowest? Is there only one mean that is highest, or is there a tie between multiple means?

Pairwise Comparison of Means: In order to identify which mean (or set of means) is the highest, we need to do pairwise comparisons of all of the different means that we are looking at. For instance, if we are comparing three means (i.e. $H_0 : \mu_1 = \mu_2 = \mu_3$), then we need to do the following pairwise comparisons: μ_1 with μ_2; μ_1 with μ_3; and μ_2 with μ_3. If instead we are comparing four means, then we have six pairwise comparisons that we need to check. In general, if we have k groups, then we need to make $\sum_{i=1}^{k-1} i = k(k-1)/2$ pairwise comparisons.

Simultaneous Pairwise Comparisons: The trick here is that we want to do all of these pairwise comparisons simultaneously. In other words we want "intervals for all the differences among the population means with confidence (say) 95% that *all the intervals at once* cover the true population differences" (p. 824 of the text). We cannot do this with regular 95% pairwise confidence intervals since the probability that all $k(k-1)/2$ confidence intervals simultanueous contain the actual differences in the population is

$$(.95)(.95)\cdots(.95) = (.95)^{k(k-1)/2} < .95$$

Tukey Confidence Intervals: Fortunately, the statistician John Tukey developed a way to overcome this problem so that we can make simultaneous confidence intervals at the 95% level (or whatever level of confidence we desire).[2] For a pairwise comparison of $\mu_i - \mu_j$, if the Tukey confidence interval contains zero, then there is no difference between μ_i and μ_j. On the other hand, if the Tukey confidence interval is strictly positive, then $\mu_i > \mu_j$. Likewise, if the Tukey confidence interval is strictly negative, then $\mu_i < \mu_j$.

2 Examples

2.1 Steakhouse Dinners

[This example is from Blake Whitten's Topic 6 lecture notes]

The Branson family owns and operates steakhouses at three locations in Des Moines, Iowa: North, West, and South. Branson North is the original Branson family steakhouse (11 years in operation, capacity 50). Branson West is the second-oldest steakhouse (5 years in operation, capacity 80). Branson South is the newest location (2 years in operation, capacity 125). Since the restaurants vary in age, capacity, and location, it is natural to ask which location is the most profitable. In order to account for the differing capacities of the three steakhouses we will formalize our question as follows:

How do total bills for four-person parties differ between steakhouses, *on average*?

[2]Incidentally, John Tukey made some other contributions that are now part of our common vocabulary. According to Wikipedia, John Tukey coined the word "bit" as an abbreviation for "binary digit." He is also credited with being the first person to use the word "software" in a publication (1958).

Parameters:
μ_1 = mean total bill for all four-person parties at Branson North
μ_2 = mean total bill for all four-person parties at Branson West
μ_3 = mean total bill for all four-person parties at Branson South

Hypotheses:
$H_0 : \mu_1 = \mu_2 = \mu_3$
H_A: At least two of the μ_i are not equal

Suppose that five bills of four-person parties were sampled from each restaurant ($n_1 = n_2 = n_3 = 5$). Branson North had an average bill of \$100, Branson West had an average of \$110, and Branson South had an average of \$120. We will consider three possible samples that produce these averages.

Case 1 Data

North	West	South
$x_{11} = 98$	$x_{21} = 108$	$x_{31} = 118$
$x_{12} = 99$	$x_{22} = 109$	$x_{32} = 119$
$x_{13} = 100$	$x_{23} = 110$	$x_{33} = 120$
$x_{14} = 101$	$x_{24} = 111$	$x_{34} = 121$
$x_{15} = 102$	$x_{25} = 112$	$x_{35} = 122$
$\Downarrow$	$\Downarrow$	$\Downarrow$
$\bar{x}_1 = 100$	$\bar{x}_2 = 110$	$\bar{x}_3 = 120$

Case 2 Data

North	West	South
$x_{11} = 80$	$x_{21} = 90$	$x_{31} = 100$
$x_{12} = 90$	$x_{22} = 100$	$x_{32} = 110$
$x_{13} = 100$	$x_{23} = 110$	$x_{33} = 120$
$x_{14} = 110$	$x_{24} = 120$	$x_{34} = 130$
$x_{15} = 120$	$x_{25} = 130$	$x_{35} = 140$
$\Downarrow$	$\Downarrow$	$\Downarrow$
$\bar{x}_1 = 100$	$\bar{x}_2 = 110$	$\bar{x}_3 = 120$

Case 3 Data

North	West	South
$x_{11} = 93$	$x_{21} = 104$	$x_{31} = 110$
$x_{12} = 95$	$x_{22} = 105$	$x_{32} = 115$
$x_{13} = 100$	$x_{23} = 110$	$x_{33} = 120$
$x_{14} = 105$	$x_{24} = 115$	$x_{34} = 125$
$x_{15} = 107$	$x_{25} = 116$	$x_{35} = 130$
$\Downarrow$	$\Downarrow$	$\Downarrow$
$\bar{x}_1 = 100$	$\bar{x}_2 = 110$	$\bar{x}_3 = 120$

Exercises:
1. Find the F-statistic for each of these sets of data.

2. Interpret the F-statistics.

3. Use Minitab to find the corresponding p-values

4. Finish the hypothesis tests using $\alpha = 0.05$.

5. Use Tukey confidence intervals to rank the restaurants in terms of average four-person bills.

Finding the MSG:
In order to find the F-statistic we will need to calculate MSG and MSE. Fortunately, the MSG is the same for all three cases. This is because the sample sizes remain the same and the average bill at each restaurant is also constant (as shown below).

North	West	South
$\bar{x}_1 = 100$	$\bar{x}_2 = 110$	$\bar{x}_3 = 120$
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$

Here we have three groups, so $k = 3$ (North, West, and South). The overall mean across all three groups is $\bar{x} = (100 + 110 + 120)/3 = 110$. We can now solve for the MSG:

$$MSG = \frac{SSG}{DFG} = \frac{\sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2}{k-1}$$

$$= \frac{5(100-110)^2 + 5(110-110)^2 + 5(120-110)^2}{3-1}$$

$$= \frac{5(100) + 5(0) + 5(100)}{2} = \frac{1000}{2} = 500$$

Finding the MSE:
To find the MSE in each case, we will need the sample variance for each restaurant. The sample variance s^2 is just the square of the sample standard deviation s. In Case 1, the sample restaurant bills at each restaurant all increase by \$1, so $s_N =$

$s_W = s_S = 1.58114$ (can you verify this on your calculator?). The final thing to note for the formula is that $N = n_1 + n_2 + n_3 = 5 + 5 + 5 = 15$.

$$MSE = \frac{SSE}{DFE} = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{N - k}$$

$$= \frac{(5-1)(1.58114^2) + (5-1)(1.58114^2) + (5-1)(1.58114^2)}{15 - 3}$$

$$= \frac{30.0000}{12} = 2.5$$

In Case 2, $s_N = s_W = s_S = 15.81139$ since the bills at each restaurant increase by increments of \$10. However, each restaurant has a different standard deviation in Case 3 ($s_N = 6.08276$, $s_W = 5.52268$, $s_S = 7.90569$). The MSE for Cases 2 and 3 can be found by substituting the relevant standard deviations into the calculations above.

Finding the F-statistic:
Calculating the F-statistic is easy once you have the MSG and the MSE. Here's the F-statistic for Case 1:

$$F = \frac{MSG}{MSE} = \frac{500}{2.5} = 200$$

Interpreting the F-statistic:
An F-statistic of 200 means that our sample has 200 times as much variance as we would expect if the null hypothesis were true (that is, if there is no difference between the average four-person bill at the three restaurants).

2.2 Grade Inflation

A phenomenon that seems to be occurring nationally is grade inflation (i.e. the average GPA of students is increasing). The average GPA of students at six Midwest high schools are shown below for the years 1980, 1990, 2000, and 2010.

School	1980	1990	2000	2010
1	2.9	3.2	3.3	3.4
2	3.1	3.3	3.4	3.6
3	3.0	3.1	3.5	3.5
4	2.8	3.0	3.2	3.3
5	2.7	3.1	3.4	3.5
6	3.2	3.3	3.5	3.6

Questions: Is there a difference between the average GPA of students over time? Test using $\alpha = 0.05$. Between what years has grade inflation clearly occurred?