Chapter 12: One-Way, Independent Analysis of Variance

Objectives for This Chapter

- Understand when and why ANOVA is used.
- Distinguish among the different types of ANOVA tests.
- Identify the major assumptions of using ANOVA.
- Complete all of the hypothesis-testing steps for an ANOVA.
- Complete and interpret an ANOVA summary table.
- Interpret a Fisher LSD, post-hoc test.

Signals, Noises and Things that go *Bump* in our data

We encounter many situations in our everyday lives in which we are trying to detect something under adverse conditions. For example, you've almost certainly had the experience at a party of trying to understand a conversation and having difficulty doing so because of the loud music and other people talking. You may have experienced so much static on your cell phone that you could not understand the person you were calling. You may even have experienced a pouting rain so loud that you couldn't hear the thunder. In these and similar situations, the background conditions can be considered "static" or "noise."

Likewise, what you're trying to detect--a nearby conversation, a voice on the other end of the phone, or thunder in a rainstorm--can be considered the "message" or "signal." In order for you to hear the message correctly, the signal must be louder than the noise. As you walk through this chapter, think about how signals and noises fit into the type of statistical analysis in this chapter.

Keep these ideas in mind and you will be fine.

- ANOVA is used for comparing the averages across three or more groups.
- The ANOVA statistic is *F*.
- The *F* table to determine critical values for ANOVA.
- A significant *F* is always followed by a little more analysis.

About ANOVA

When and Why we use ANOVA

First off, ANOVA stands for *analysis of variance*. We use ANOVA when we wish to compare more than two group averages under the same hypothesis. How many groups? As many as you want. Hopefully, you will have a very good theoretical rationale for comparing 20 or 30 group averages in a single ANOVA. Most of the time, you won't be comparing that many groups.

The implication of this is that ANOVA is always at some level a *nondirectional test*. Think about it--"direction" implies some prediction as to which group average will be highest. And if we had some previous, theoretical expectation that one group would have a higher average than another group, then we should modify our hypothosis and compare just those two groups--perhaps using a t-test. When we run an ANOVA, we're admitting that we're not sure which of our groups will emerge with the highest group average--and for that matter, which will have the lowest average. It's fair to say that ANOVA is an *exploratory procedure*. This will make more sense when we get into an example.

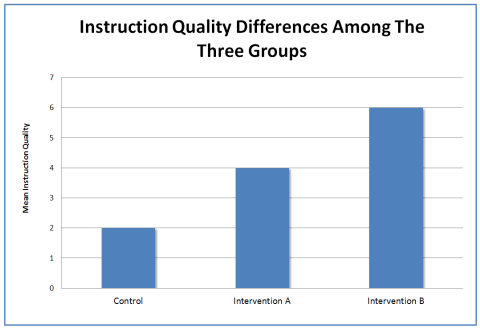
Why don't we do a bunch of t tests? One reason for preferring ANOVA over multiple t tests is that this would be a very tedious exercise. Doing t tests over and over again just aint no fun!

A more important reason is that the more tests we run under the umbrella of the same hypothesis, the more likely we are to get a spurious significant difference. In other words, we're increasing the probability of Type I error. For example, if we set alpha at .05 and run three t tests for three variables that are part of the same hypothesis, our probability of Type I error is actually 15%--5% for each test. If we didn't adjust our alpha level accordingly (dividing it by 3), we would not be holding our data to a sufficiently high significance-testing standard.

Another reason for using ANOVA is that there are complex types of analyses that can be done with ANOVA and not with the t tests. Additionally, ANOVA is by far the most commonly-used technique for comparing means, and it is important to understand ANOVA in order to understand research reports.

Different types of ANOVA Tests

ANOVA is a type of model that can be applied to a variety of data configurations. The ANOVA that we are going to study is the **one-way, independent ANOVA**. Specifically, a one-way ANOVA is used when there is a single independent variable that has three or more categories. The One-Way ANOVA tells us if the three (or more) groups differ from one another on a dependent variable.

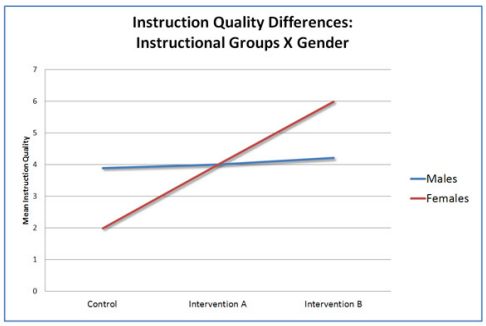


(../../images/anova/instruction_qual_graph.png)Imagine a study in which researchers implemented two separate interventions, one designed to improve skill building among youth (Intervention A) and another designed to increase supportive relationships between instructors and youth (Interventiosn B). In addition, the researchers also employed a control group in their study. The researchers want to know if either of these two interventions improved instruction quality. In this study there are three groups, participants who received Intervention A, those that received Intervention B, and the control group. Below is a graph of what the means for each group might look like. Researchers interested in differences between these three groups in terms of instruction quality would apply a one-way ANOVA. The one-way ANOVA provides information about if there were statistically significant instruction quality differences between these three groups. **Click to enlarge images.**

The result of this one-way ANOVA indicates that there are differences between the three means. However, ANOVA on its own does not provide information about where these differences actually are. In this example there could be a difference in instructional quality between the Control and Intervention A groups, between the Control and Intervention B groups, and/or between the Intervention A and Intervention B group averages. To get at these differences additional analyses must be conducted. More on this, later.

Though we aren't going to walk through full examples of other types of ANOVA, it is important to give you at least a brief introduction to a few other ANOVA models. The first of these is what we will call a **two-way ANOVA**. This one is sometimes referred to as a *factorial* ANOVA. Unlike a one-way ANOVA, a two-way ANOVA is used when there is more than one independent variable. In the previous example there was only one independent variable with three levels (Intervention A, Intervention B, Control). Now, suppose that a researcher also wanted to know if there were *additional group differences between boys and girls* in the youth program. When this second independent variable is added to the analysis, a two-way ANOVA must be used.

The results of a two-way ANOVA consist of several parts. First are called *main effects*. Main effects tell you if there is a difference between groups for each of the independent variables. For example, a main effect of



(../../images/anova/two_way_graph.jpg)intervention type (Intervention A, Intervention B, Control) would indicate that there is a significant difference between these three groups. That is, somewhere among these three averages, at least two of them are significantly different from one another. Like the one-way ANOVA, a two-way ANOVA does not provide information on where these differences are and additional analyses are required.

In addition to the main effects, two-way ANOVA also usually provides information about *interaction effects* as well. Interaction effects provide information about whether an observed group difference in one independent variable varies as a function of another independent variable.

The graph shows that there are no differences for boys (blue line) between the Control, Intervention A, or Intervention B. However, there are clear differences for girls (red line) between these three groups. The implication of an interaction such as this one is that differences between groups are dependent on gender.

Moving right along. The last type of ANOVA is referred to as **repeated-measures ANOVA**, which can be either one-way or two-way. To put it in terms of our t test chapter, we could also call this type of ANOVA a *dependent-groups* ANOVA. As discussed in a previous chapter, a dependent-samples t test is used when the scores between two groups are somehow dependent on each other. One example of such a dependency is when the same people are given the same measure over time to see whether there is change in that measure. The repeated-measures ANOVA takes the dependent samples t test one step further and allows the research to ask the question "*Does the difference between the pre-test and post-test means differ as a function of group membership?*"

Here's an example. Suppose that we are interested in the effect of practice on the ability to solve algebra problems. First we test 20 participants in algebra performance before practice, recording the number of problems they solve correctly out of 10 problems. We then provide the participants with practice on algebra problems and retest their performance after one day and then again after one week. Essentially, we're looking at whether the effects of practice persist. Because we have three groups comprising the same participants, the best analysis would be a repeated-measures ANOVA.

Assumptions of ANOVA

The following are assumptions of one-way, independent ANOVA:

- **Normality:** The data for each group are approximately normally distributed.
- **Homogeneity of variance:** The variances of the group distributions will be statistically similar.
- **Sample size:** per group > 20 is preferred with approximately equal N sizes; aids robustness to violation of the first two assumptions, and a larger sample size

increases power.
- **Independent observations:** scores on one variable or for one group should not be dependent on another variable or group.
- **Interval/ratio scale:** The data comprising the distributions will be interval or ratio level.

Thinking Through an ANOVA Example

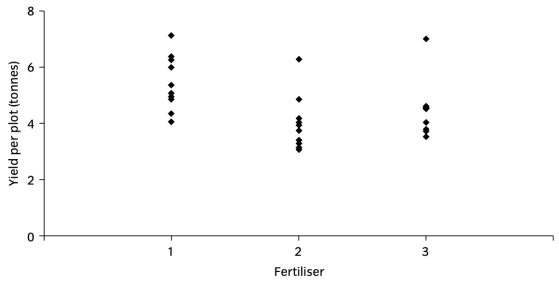
The first and simplest problem to consider is the comparison of three means; though if you're not worried about the math, which we won't be, it doesn't matter how many means we're dealing with. This is done by the analysis of variance (ANOVA). The aim of this chapter is to look at an example in some conceptual detail.

So, here is the example. If we have three fertilizers, and we wish to compare their effectiveness, this could be done by a field experiment in which each fertilizer is applied to 10 plots, and then the 30 plots are later harvested, with the crop yield (in tons) being calculated for each plot. We have 3 groups with 10 scores in each.

Though we won't be calculating our *F-comp* by hand, it will be helpful to at least know what the raw data are:

Fertiliser	Yields (in tonnes) from the 10 plots allocated to that fertiliser
1	6.27, 5.36, 6.39, 4.85, 5.99, 7.14, 5.08, 4.07, 4.35, 4.95
2	3.07, 3.29, 4.04, 4.19, 3.41, 3.75, 4.87, 3.94, 6.28, 3.15
3	4.04, 3.79, 4.56, 4.55, 4.53, 3.53, 3.71, 7.00, 4.61, 4.55

When these data are plotted on a graph, it appears that the fertilizers do differ in the amount of yield produced (we'll call this *within-groups variation* also known as *error variation*), but there is also a lot of variation between plots (we'll call this *between-groups variation*). Whilst it appears that fertiliser 1 produces the highest yield on average, a number of plots treated with fertiliser 1 did actually yield less than some of the plots treated with fertilisers 2 or 3.



(../../images/anova/fertalizer_graph1.png)We now need to compare these three groups to discover if this apparent difference is statistically significant. When comparing two samples, the first step was to compute the difference between the two sample means (see revision section). However, because we have more than two samples, we do not com- pute the differences between the group means directly. Instead, we focus on the variability in the data. At first this seems slightly counter-intuitive: we are going to ask questions about the means of three groups by analysing the variation in the data. How does this work? **Click images to enlarge.**

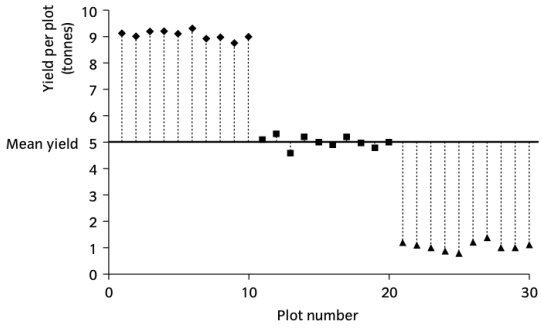
The variability in a set of data quantifies the scatter of the data points around the mean. To calculate a variance, first the mean is calculated, then the deviation of each point from the mean. We know from before that adding up the raw deviations from a mean always yields 0, which is not helpful. If the deviations are squared before

summation then this sum is a useful measure of variability, which will increase the greater the scatter of the data points around the mean. This quantity is referred to as a **sum of squares** (SS), and is central to our analysis.

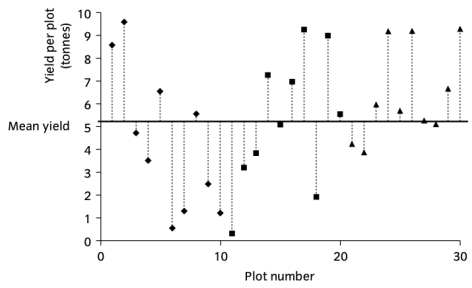
The SS however cannot be used as a comparative measure between groups, because clearly it will be influenced by the number of data points in the group; the more data points, the greater the SS. Instead, this quantity is converted to a variance by dividing by n − 1, where n equals the number of data points in the group. Doint this gives us an "average" SS that can be compared across groups.

Partitioning Variability

In an ANOVA, it is useful to keep the total measure of variability in its two components--within-groups and between-groups; that is, a sum of squares, and the degrees of freedom associated with the sum of squares. Returning to the original question: *Do the three fertilizers produce equal yeilds of crops?* Numerous factors are likely to be involved: e.g. differences in soil nutrients between the plots, differences in moisture content, many other biotic and abiotic factors, and also the fertiliser applied to the plot. It is only the last of these that we are interested in, so we will divide the variability between plots into two parts: that due to applying different fertilisers (between-groups), and that due to all the other factors (within-groups).



(../../images/anova/fertalizer_graph3.png)To illustrate the principle behind partitioning the variability, first consider two extreme datasets. If there was almost no variation between the plots due to any of the other factors, and nearly all variation was due to the application of the three fertilisers, then the data would follow the This pattern. The first step would be to calculate a **grand mean** (the mean of all scores from all groups), and there is considerable variation around this mean. The second step is to calculate the three group means that we wish to compare: that is, the means for the plots given fertilisers A, B and C. It can be seen that once these means are fitted, then little variation is left around the group means. In other words, there is little within-group varability in the groups but large between-groups variability across the groups. When all is said and done, we would be almost certain to reject the null hypothesis for this experiment.



(../../images/anova/fertalizer_graph4.png)Now consider the other extreme, in which the three fertilisers are, in fact, about the same. Once again, the first step is tofit a grand mean and calculate the sum of squares. Second, three group means are fitted, only to find that there is almost as much variability as before. Little variability has been explained. This has happened because the three means are relatively close to each other (compared to the scatter of the data).

The amount of variability that has been explained can be quantified directly by measuring the scatter of the treatment means around the grand mean. In the first of our two examples, the deviations of the group means around the grand mean are considerable, whereas in the second example these deviations are relatively small. When the variability around the grand mean isn't all that much different from the variability around the individiual means, we say that *we've not explained much variance*. Explain it how? Quite simply explain it in terms of our independent variable--type of fertilizer. In this second example, we can't tell whether the variability is coming from the fertilizers, or the soils or the weather, etc. When all is said and done, we would almost certainly fail to reject the null hypothesis for this experiment.

But at what point do we decide that the amount of variation explained by fitting the three means is significant? The word significant, in this context, actually has a technical meaning. It means 'When is the variability between the group means greater than that we would expect by chance alone?' At this point it is useful to define the three measures of variability that have been referred to. These are:

Total sum of squares (SST): Sum of squares of the deviations of the data around the grand mean. This is a measure of the total variability in the dataset.

Within-Groups or Error sum of squares (SSW): Sum of squares of the deviations of the data within each distribution.

Between-groups sum of squares (SSB): Sum of squares of the deviations of the group means from the grand mean. This is a measure of the variation between plots given different fertilisers.

Variability is measured in terms of sums of squares rather than variances because these three quantities have the simple relationship: SST = SSW + SSB. So the total variability has been divided into two components; that due to differences *between* plots given different treatments, and that due to differences *within* plots. Variability must be due to one or other of these two causes. Separating the total SS into its component SS is referred to as partitioning the sums of squares.

Partitioning the degrees of freedom

For ANOVA, there are always two degrees of freedom. This is reflected in the *Table of Critical F Values*. Expand this table (or zoom in) and note the differences between this table and the last one that we used.

Every SS was calculated using a number of independent pieces of information. As with all lists of numbers, some are free to vary and some are not. The first step in any analysis of variance is to calculate SST. It has already been discussed that when looking at the deviations of data around a central grand mean, there are $N - 1$ independent deviations: i.e. in this case $N - 1 = 29$ degrees of freedom (df). The second step is to calculate the three treatment means. When the deviations of two of these treatment means from the grand mean have been calculated, the third is not free to vary. Therefore the df for SSB is 2. Finally, SSW measures variation within each of the three groups that are part of this study.. Within each of these groups, the ten deviations must sum to zero. Given nine deviations within the group, the last is predetermined. Thus SSW has $3 \times 9 = N - 3 = 27$ df associated with it.

Just as the SS are additive, so too are the df. Adding the df for both SSW and SSB equals the df associated with SST. Combining the information on SS and df, we can arrive at a measure of variability per df. This is equivalent to a variance, and in the context of ANOVA is called a *mean square* (MS). The calculations for MS are as follows:

$$\text{Mean Square Between (MSB)} = SSB/df_{between}$$

$$\text{Mean Square Within (MSW)} = SSW/df_{within}$$

Essentially, what we now have is all of the variance in the experiment, properly accounted for. And because we have kept the variability separated into variance within and between groups, we can compute a statistic that will allow us to determine whether there is a significant difference among the average crop yields in our study. How, do we do that? We do what we always do--*we put the signal (between) over the noise (within).*

the ANOVA Summary Table

One of the most important considerations in applying and understanding ANOVA is the summary table. For any ANOVA analysis that you conduct, a summary table will be the center of the output. It is from such a table that we decide whether to accept or reject a given null hypothesis. In the table below, you can see how to calculate *degrees of freedom*, *mean square values* and even *F-comp.*

Source	DF	SS	MS	F	P
Between	K-1	10.823	$\frac{SS_b}{df_b}$	$\frac{MS_b}{MS_w}$	
Within	N-K	25.622	$\frac{SS_w}{df_w}$		
Total	N-1	36.445			

This table helps us keep track of the different parts of our variation. Examination of the summary table reveals some items we have not yet discussed. First, note that the sources of variation (between and within) are indicated on the left side. Next we see places for degrees of freedom. The between-groups *df* is equal to K - 1, where *K* is the number of groups we are comparing. Because we are comparing three groups in our example, this is 3 - 1 = 2. The within-groups *df* equals *N - K*, which means that we subtract the number of groups from the total number of participants in all of our groups. For our example this is 30 - 3 = 27. Finally total df equals *N - 1*, or 30 - 1 = 29, for our example. Notice that *df-between* and *df-within* can be added up equal *df-total*.

The hardest task in constructing a summary table manually is the calculation of the *sums of squares* in the *SS* column. We will not be going into these calculations in this class, but having this information already in the table ensures that we can calculate our *mean squares*. As you can see, calculating MS is simply a matter of dividing SS by its df. *MS-between* comes out to 5.41, and *MS-within* comes out to .949.

And once we have MS-between and MS-within, we can calculate our *F*-comp or our *F* ratio. To do this, we simply *divide MS-Between by MS-Within*. For our example, this is 5.41/.949 = 5.7. And that's it. That's *F*-comp. Remember that *F*-comp will always be positive. *F* is never negative. Now, all we have to do is compare our computed *F* with the critical *F* for our alpha and degrees of freedom.

Finishing up the ANOVA...Almost

Let's go through this using our steps of hypothesis testing. Remember that if we were conducting research, we would go through the first four steps before we even collected data.

1. H_0: $\mu_1 = \mu_2 = \mu_3$. The null hypothesis for ANOVA is called an *omnibus hypothesis*. This type of hypothesis covers multiple groups and assumes that there is equality among all group averages. Stating the null in words: *There is no difference in the amount (measured in tons) of crops yielded by the three fertilizers.*

2. H_1: *The crop yields among the three groups are not all the same*. Another way to say this is: *At least one group average will be significantly different from one other group average.* So, for ANOVA, we will reject the null if only two group averages are significantly different from one another. Even if we run an ANOVA with nine groups, only two of them have to be significantly different from one another in order for our data to yield a significant F-comp.

3. Set $\alpha = .05$.

4. Reject H_0 if *F*-comp ≥ *F*-crit; $F_{.05,\ df\ =\ 2,\ 27} = 3.35$.

5. The fifth step in this chapter is a little different from past chapters. Though you won't be performing any extensive calculations, you do need to be able to fill out a summary ANOVA table, with relevant information. Filling out the summary table correctly will yield the correct F-comp. Here is the completed summary ANOVA table for the fertilizer study. In the previous section, we walked through the steps for filling in the summary table. **Click image to enlarge.**

ANOVA Summary Table For Fertilizer Study

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	10.823	2	5.411	5.702	.009
Within Groups	25.622	27	.949		
Total	36.445	29			

(../../images/anova/fertalizer_ANOVA_summary.png)

6a. *A one-way ANOVA was conducted to determine whether there is no difference in the amount (measured in tons) of crops yielded by the three fertilizers.*

6b. *There was sufficient evidence to reject the null hypothesis; F(2, 27) = 5.7, p < .05.* If our F-comp value was less than F-crit, then we would have said, *There was insufficient evidence to reject the null hypothesis...*

6c. *The average crop yield associated with Fertilizers 1, 2, and 3 are not all the same (5.45, 4.00, and 4.49 tons, respectively).*

At this point we're unable to state exactly which groups have averages that are significantly different from one another, because we have only one computed statistic--a single F value. Based on a single, significant F value we cannot draw any conclusions or talk about the implications of the study. In order to do this, we have to compare pairs of averages using another type of test.

6d. Further post-hoc testing of pairwise differences is necessary to determine which group averages are significantly different from one another.

Post-Hoc Testing after Significant F

The **Fisher LSD (least significant difference)** test, or simply **LSD test**, is easy to compute and is used to make all pairwise comparisons of group means. By pairwise comparisons, we mean that we make all possible comparisons between groups by looking at one pair of groups at a time. One advantage of the LSD is that it does not require equal sample sizes. Another advantages is that it is a powerful test; that is, we are more likely to be

able to reject the null hypothesis with it than other post-hoc tests.

The LSD is sometimes called a **protected *t* test** because it follows a significant *F* test. If we used the t test to do all pairwise comparisons before the F test, the probability of committing a Type I error would be greater than our α. However, by applying a form of the t test, after the *F* test has revealed at least one significant comparison, we say the error rate (probability of Type I error) is "protected."

Without going into the formula and computations, it is sufficient to say that running an LSD test consists of first calculating a critical LSD value, which is based on sample sizes, our MSW value, and a critical *t* value pulled from the *t* table. More than one LSD value needs to be calculated if group sizes are different. Once the LSD value has been calculated, the differences between group averages are compared to the LSD value. Average differences that are equal to or larger than the LSD value are considered statistically significant and would be interpreted as such. Now, in the table below, you won't see the calculated LSD value, because the computer program from which this derives, handles that, behind the scenes. Therefore, interpretation will be a little different than if we were performing all of the calculations manually. Read on.

LSD Table of Mean Differences

Amount of crop yield in tons

LSD

(I) Fertilizer	(J) Fertilizer	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Fertilizer 1	Fertilizer 2	1.446*	.43565	.003	.5521	2.3399
	Fertilizer 3	.958*	.43565	.037	.0641	1.8519
Fertilizer 2	Fertilizer 1	-1.446*	.43565	.003	-2.3399	-.5521
	Fertilizer 3	-.488	.43565	.273	-1.3819	.4059
Fertilizer 3	Fertilizer 1	-.958*	.43565	.037	-1.8519	-.0641
	Fertilizer 2	.488	.43565	.273	-.4059	1.3819

*. The mean difference is significant at the 0.05 level.

(../../images/anova/fertilizer_LSD_Table.jpg)Here is the *LSD Table of Mean Differences*. Note that unimportant columns have been grayed out for easier interpretation. **Click image to enlarge**. This table comprises a few pieces of information, the most important of which are the *Mean Difference* and *Sig*. columns. In the Mean Difference column, we see all of the differences between the group averages (e.g., average of Fertilizer 1 minus the average of Fertilizer 2). The Sig. column just provides the probability associated with each difference. We interpret these group differences as we've always done. If p>.05 (assuming that is the established alpha level for the study), then we accept that the two group averages are not significantly different from one another. If p<.05, then we reject the null hypothesis and conclude that there is a statistically significant difference between two group means.

Now, with the LSD table in front of us, we can finish off this problem. There are basically three steps, here. First, we will clearly state the kind of post-hoc test that we conducted. Second, we will clearly identify all statistically significant, pairwise differences and include the averages for each. Third, we need to talk about the implications of such findings. In other words, we have to answer the "So What?" question. Here it is:

7a. *A Fisher's LSD test was conducted to determine which pairwise differences between group averages (tons of crops yielded for each type of fertalizer) are statistically significant.*

7b. Our LSD test revealed that Fertilizer 1 yielded statistically, significantly more crops (M=5.45 tons) than did Fertilizers 2 and 3 (M=4.00 and 4.49 tons, respectively).

7b. *Given these findings, more extensive usage of Fertalizer 1 is recommended. More resources should now be devoted to assessing the consumer cost of Fertalizer 1, potential ecological impacts of this fertalizer and whether it is ultimately safe for consumers. It would also be advisable to contact farmers for expanded testing--i.e., further testing on other crops in other locations is also advised, so as to improve methodological representativeness.*

That's it! Go forth and harvest!!

Conceptualizing and Visualizing ANOVA

Use this interactive animation to get a better understanding of one-way ANOVA. Notice what happens to *F*-comp when you increase the variability within groups. Notice what happens when you increase the differences between the groups and the grand mean. Also, keep an eye on the bars representing MSB and MSW or the *F* ratio. Can you figure out how to get the largest *F*-comp possible? Definitely one of the coolest applets I've found on the web.

Self Test

(t-tests_test.pdf)

- Self-test for chapter (anova-test.pdf)
- Answers to self-test (anova-answers.pdf)

<< back to top

(index.html)

Some content adapted from other's work. See home page for specifics.

LAST UPDATED: 2014-08-25 6:55 PM

Mesa Community College | 1833 W. Southern Ave. Mesa, AZ 85202 | E-mail Address: dborman@mesacc.edu | Phone: (480) 461-7181 |

DEREK BORMAN: **PSYCHOLOGICAL SCIENCE**
MCC **PSYCHOLOGICAL SCIENCE** HOMEPAGE