Mac Guide-Advanced Excel

This document contains instructions for macs where they deviate from PC instructions. Changes are in red. The numbers for each set of instructions correspond to the slide number in the ppt. If nothing is in red, you can follow along using PC instructions.

To right click on a mac: click the track pad with 2 fingers

1. Welcome to the EASE workshop series, part of the STEM Gateway program.

If you haven't already, please download the dataset from the EASE website. Don't worry about opening it right now, I'll make sure you know when to do so.

As with the Basic Excel course, this workshop will just begin to introduce you to additional things you can accomplish in Excel. This will focus on a few basic statistical analysis using the Data Analysis plug-in. The images are from the newest version of Excel, 2013 (which you can get through IT for free), for PC's. Older versions, or those in Macs may look slightly different, but the overall concepts should still apply. I want to make sure we are clear that this is by no means meant to be an all inclusive class in Excel. At each step, there are many options that I won't have time to go over, so I encourage you to spend some time on your own playing with them.

If you are using a Mac today, know that is perfectly ok, but I may not be able to help you with all the issues you encounter, so please help each other. Make sure you download the Mac Guide from the website, too, so you have it to follow along.

Lastly, as a heads up, at the end of the presentation you'll have to complete an exit survey in order for me to accept your assessment sheet.

2. If you don't remember, the Basic Excel workshop that you took in 203 is available through the EASE workshop page. We are not going to spend time going over topics covered in that workshop, but many of them are beneficial for you here in 204, so make sure you go back and look through that presentation and supplemental information. Topics that were covered include graphing, as well as the Filter and Sort functions, which will really help you this semester when it comes to cleaning up your dirty data before you can move forward with any analysis for your homework's.

3. Some additional resources that are available to you through the EASE website are:
Instructions on how you can get Microsoft Office Pro for free for being a UNM student.
Just like with Basic Excel, this presentation is available as a PDF on the EASE website if you need to refer back to it. It includes the script, so you don't have to memorize everything I say to be able to interpret the slides.
There is also a file that guides you through the differences between PC and Mac versions of what we are about to go through, including the need for StatPlus, rather than the Data Analysis Ad-In that we'll use today.

Lastly, there is a document with links to videos on basic Statistic concepts, if you want to learn more about what is underlying each of the tests we learned how to perform.

4. This workshop will first cover a few basic vocabulary terms that are important to understand when doing statistical analysis. I'll walk you through how to add and activate the Analysis ToolPack plug-in. Then we'll go over HOW to run three different statistical analysis Excel, but only briefly go into how to interpret the results.

5. Let's go over mean, median and mode. If we have this example data set.

(*) To get the mean in Excel we type in =average(first cell of the data: last cell of the data). The actual calculation is adding each data value together, then dividing by the number of data points.

(*) To calculate the median by hand, you have to re-write the data set in numerical order to find the central value. Since there are 9 values, the 5^{th} value is the central point, and is 14 in this case. Again, I show you the Excel equation.

(*) Lastly, to calculate mode, it is the value to occurs the most in the data set. You can see from the Median example that we have 13 appear 4 times, 14 twice, and the other values only once. That means that 13 is the mode value.

6. Within the concept of Variability, the range is breadth of the data, that is, the largest value minus the smallest value. The equation in orange is what you would type into an Excel cell to calculate the range of a data set. The Maximum value within data in cell x through cell y, minus the minimum value within the data set.

(*) The variance is calculated as the squared difference of each value from the mean. So with this image, we would take the difference of each value (red lines) from the mean (green line) and square it, then divide it by the number of measurements, in this case 5. Again, the equation you'd enter into Excel is shown in orange. This is for variance of a sample set, rather than a population. We don't have time to go into what that means, but make sure you are using an appropriate calculation of variance within Excel, based on what is appropriate for your data set and test.

(*) Lastly, the Standard Deviation measures how spread out the values are from each other, it is calculated by obtaining the square root of the variance. As with variance, the standard deviation equation is for a sample set of values.

7. Before you can run any type of statistical analysis on your data, you have to first know what type of data you have. There are two major types:

(*) Categorical variables are those that are qualitative, or nominal. In other words, they do not consist of numbers directly, but of categories or names. For example, if we were comparing brands of shoes, the brand name is the categorical variable.

(*) Other attributes associated with shoes could be considered numerical, or quantitative, variables. These are measured on a numerical scale. These are split into two types: discrete and continuous.

Discrete numerical variables consist of only certain fixed values with no intermediate values, partial numbers are not possible, such as shoe size, you can have a 4 or a 5, but not a 4.27.

(*) On the other hand, continuous numerical variables include all possible intermediate values. Variables such foot size, since your feet can have any value within a given shoe size.

8. Now that you know the type of data you have, you need to know what question you're asking. The statistical tests that we'll go through focus on determining if two or more data sets are significantly different from each other, or if the values are correlated.

The null hypothesis is that there no difference between data sets, or that the data sets are not correlated, that is, they are independent of each other.

The alternative hypothesis accounts for all other possible outcomes.

To test the data in light of these hypotheses, we use a probability distribution table, which links each outcome of a statistical experiment with its probability of occurrence.

9. To be able to determine significance, we have to settle on a level of significance. How strong does the relationship have to be to reject the null hypothesis? The alpha is usually 0.01 (1%), 0.05 (5%), or 0.1 (10%). The smaller the value, the stronger the evidence must be to reject the null hypothesis. There are different levels because different disciplines have different requirements.

The probability that the outcome is not statistically supported assuming the null hypothesis is true is known as the **p-value**. P-values that are smaller than your set significance level (α) mean that the outcome is statistically significant. For example, a p-value less than 0.05 means there is less than a 5% chance that the observed outcome is due to chance.

If your p-value is less than the designated alpha, then there IS a statistical difference between the groups in question, at that level of confidence. In other words, you would reject the Null hypothesis, and accept the alternate hypothesis. We have an alpha to determine how acceptable in our results. The significance of the p-value is based on the

chosen alpha. Yes, the p-value has to be small, but it is all relative to the chosen alpha. This is why medical fields use such a small alpha, but ecology allows a larger one.

10. There are four common tests to test these hypotheses:

Chi-Square tests compare two categorical variables. That's as much as we're going to discuss Chi-Square because there is not a "quick" way to use Excel to run this, like there is for the other types of tests we'll talk about.

Regressions compare two non-categorical variables within the same data set to determine if they are correlated or independent of each other.

And, ANOVA (ANalysis Of VAriance) and t-test compare means and distributions across two (t-test) or more (ANOVA) data sets. Comparing one categorical variable to one numerical variable.

11. As I mentioned, when you have Categorical and Numerical data, you have to select between an ANOVA and a t-test. These are used to compare the MEANS of data sets. If you want to compare medians, there are other tests for that.

The main difference between the two tests is that a t-test is used to compare the difference of means and distributions between TWO groups, where an ANOVA is used to compare the means of more than two groups. We'll go into details on each of these in a minute, but first let's look at some examples.

12. Earlier, we went over how to determine the type of data you have, but not how that relates to which analysis you should run. This graphic should help with that choice. We'll go over how you can do each of these in Excel. You'll have to spend time looking through other resources I mentioned earlier to fully understand the mechanics behind each of these tests.

13. So, based on what we've gone over so far, match the hypothesis with the appropriate test and write these down on your assessment sheet, with a description as to WHY you chose that answer.

What did you come up with? What type of test would you run for the first hypothesis?

14. They first hypothesis would need a t-test, since you are comparing categorical (gender) and numerical (height) variables. And, it is a t-test since there are only two groups being tested.

15. When you run a t-test, you'll get a resulting p-value. Remember, the alpha level you've decided on will determine if the p-value indicates the acceptance or rejection of the null hypothesis. In this case the p-value is so small that the null is rejected at any of the standard alpha levels. That means there IS a difference between mean and distributions of male and female height.

This graphic is referred to as a box plot and the whiskers or error bars on the box are the upper and lower quartiles, or 1/4 and 3/4 medians. The error bars could also indicate the standard deviation, or range.
Just as a side note, you can NOT easily make box-plots like this in Excel, but I wanted you to see it in case you encounter it somewhere else, and to give you a visual of what the test is doing.

16. For the second hypothesis, again we have categories of state, and numbers of height, but we have three groups, thus an ANOVA, rather than a t-test, is appropriate.

17. As before, when you run an ANOVA, you'll get a resulting p-value. If the p-value is significant, it indicates that you reject the null that all means and distributions are the same, and that there IS a difference among the means and distributions of heights, but from this you can NOT determine WHICH pair are different, just that some pair is different. You'd have to run additional statistical tests to determine which pairs are different.
From looking at the graphic, it looks like NM and NV may be the same, as is NM and TX, but that NV and TX are different. But again, to statistically determine this, though, you'd have to run additional tests.

18. For the final hypothesis, since both temperature and latitude are numerical, a Regression is the best type of analysis.

19. As with the other tests, you'll get a p-value, but there are other things to look at as well. The R-squared value indicates how well the line explains the data. A 0 indicates a poor explanation and a 1 is a perfect explanation. You can have a straight horizontal line that has a slope of zero, thus a high R-squared, but the p-value will be low. This is because the p-value indicates if the relationship of the two variables is different than 0, different than a straight vertical or horizontal line.
You can also get the regression equation, which is the equation of the trendline.
For this specific example, you can see the difference of a high, middle, and low R-squared values, with all significant p-values.

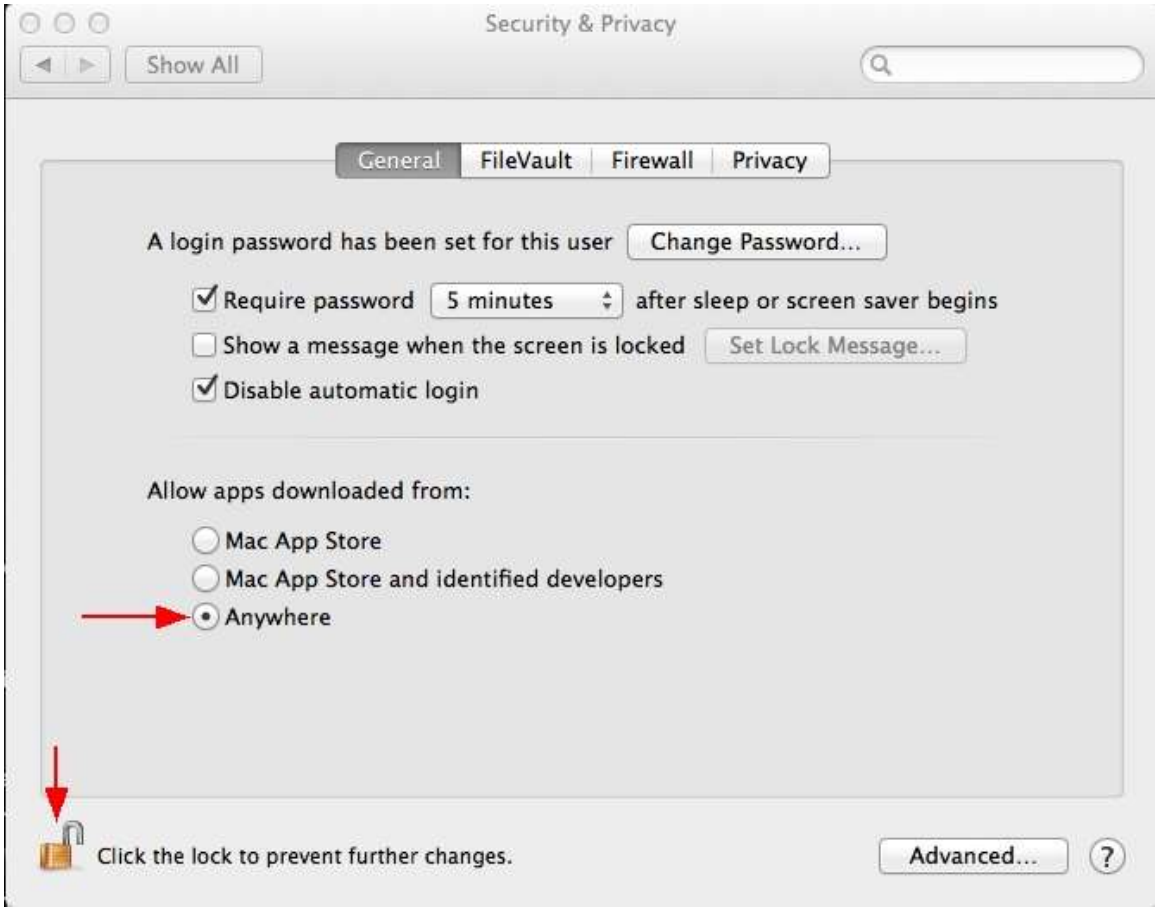
20. Ok, now let's get to you working with some data! You can run all of these tests within Excel without the plug-in by using the pull-down menu of functions under the "formulas" tab, but the ToolPak makes it easier to find and run.
To more easily do any of these analysis in Excel, you have to activate the analysis toolpack. This provides data analysis tools for financial, statistical, and engineering data analysis. There are many more analysis than we are going to cover, so make sure you take some of your own time and look through them.

For Macs, you will need to download StatPlus in order to run the following analyses. For instructions on how to download StatPlus:

Follow this link, and the instructions therein, to get, install, and start using StatPlus:

http://depts.washington.edu/mbaclub/wordpress/wp-content/uploads/2013/07/Tutorial-Adding-Solver-Excel-2011-Mac.pdf

Note: you may need to changes your security settings to allow StatPlus to open. To do so, click on the Apple (•) in the upper left corner of your screen, or go to Applications, to open System Preferences. Go to Security and Privacy, and under the General Tab, select Anywhere from the options under the Allow apps downloaded from:. You may also need to click on the lock in the lower left corner to enable changes.



21. Ok, let's make sure the add-in is activated. Click "file", then select "options". Macs are slightly different, and you have to use StatsPlus, so make sure you go back and access the "Advanced Excel for Macs - Guide" to help you with the differences between platforms. (See link in previous section.)

22. You should have followed the instructions in the above link to activate the Analysis ToolPak.

23. Same as previous step.

24. When you click the “data” tab, you should now see the “Data Analysis” option. (On a mac, you will not see this.)

Once you’ve activated this on your personal computers, you should be set. But, if you are using a university computer, then you may have to repeat this process each time you use it. You an always check before you get in to your analysis by looking for the Ad-In under the data tab.

Analysis within this are many, including descriptive statistics which will give you the Standard Deviation of a data set. Play around with them on your own at a later time, since, as I said already, we’ll only go over 3 major analyses.

25. The first type of analysis we’re going to cover is a linear regression, or correlation. This looks at how strongly two variables are correlated or associated with each other. If the variables are correlated, they form a linear relationship. The Null hypothesis is that there is NO relationship between the two variables, that they are independent of each other. The alternate hypothesis then, is that there IS a relationship between the two variables.

26. This Excel output example is for the frequency of cricket chirps in relation to the ambient temperature. The graph is not made automatically, but I wanted you to have a visual of the data in relation to the output statistics. And, step by step on how to create a graph like this was covered in the Basic Excel workshop.

The R-squared is 0.697. Remember, the closer to 1, the better the line explains the data. For the regression equation, you look at the coefficients of the intercept and the x-variable, with the intercept being the final value, and x variable being the coefficient to x in your standard y=mx+b equation of a line.

To determine if there is a significant relationship between the two variables tested, you look at the “Significance F” value, which assessed the same way you would a P-value. In this case, the results are significant, so you reject the null of no difference and accept the alternative hypothesis that there IS a correlation between cricket chirp rate and temperature. On a side note, remember, just because two variables are correlated, it does not mean that one is causing the other.

27. Ok, now it’s your turn. Make sure you are on the Regression worksheet. Select the data tab. Then click on “Data Analysis”.

One thing I want to point out is the “HELP” option, which is great if you want a description of each test to determine if it’s appropriate.

Scroll down and select the Regression analysis.

These data are for the relationship between the selling price of a house and the size of the house.

17. Earlier, we went over how to determine the type of data you have, but not how that relates to which analysis you should run. This graphic should help with that choice. We'll go over how you can do each of these in Excel. You'll have to spend time looking through other resources to fully understand the mechanics behind each of these tests. The difference between a t-test and an ANOVA is that a t-test looks at the difference of means between TWO groups, where an ANOVA compares more than two. We'll go into details on each of these in a minute, but first let's look at some examples.

18. If the hypothesis is that men and women are not, on average, the same height. What type of data would this result in? – Categorical and numerical.

(*) So, what are the type of variables? Qualitative (Sex), and Quantitative (Height)
(*) Knowing that you have categorical and numerical variables, how many groups do you have? – 2
(*) So would you do a t-test or an ANOVA? A t-test.

19. When you run a t-test, you'll get a resulting p-value. Remember, the alpha level you've decided on will determine if the p-value indicates the acceptance or rejection of the null hypothesis. In this case the p-value is so small that the null is rejected at any of the standard alpha levels. That means there IS a difference between mean male and female height.

This graphic is referred to as a box plot and the whiskers or error bars on the box are the upper and lower quartiles, or 1/4 and 3/4 medians. The error bars could also indicate the standard deviation, or range.

Just as a side note, you can NOT easily make box-plots like this in Excel, but I wanted you to see it in case you encounter it somewhere else, and to give you a visual of what the test is doing.

20. So here's an example hypothesis: People from New Mexico, Texas and Nevada do not, on average, have the same height. What type of data would this result in? – Categorical and numerical.

(*) So, what are the type of variables? Qualitative (State of residence), and Quantitative (Height)
(*) Knowing that you have categorical and numerical variables, how many groups do you have? – 3
(*) So would you do a t-test or an ANOVA? ANOVA

21. As before, when you run an ANOVA, you'll get a resulting p-value. If the p-value is significant, it indicates that you reject the null that all means are the same, and that there IS a difference among the mean heights, but from this you can NOT determine WHICH pair are different, just that some pair is different.

From looking at the graphic, it looks like NM and NV may be the same, as is NM and TX, but that NV and TX are different. To statistically determine this, though, you'd have to run t-tests on each of these pairs.

22. Last one: The temperature of the earth decreases with latitude.

(*) So, what are the type of variables? Both Quantitative (temperature and latitude)
(*) So what type of test would you do? Regression

23. As with the other tests, you'll get a p-value, but there are other things to look at as well. The R-squared value indicates how tight the data points to each other. A 0 is no relationship and a 1 is a perfect linear relationship. The p-value indicates if the relationship is significant or not. Sometimes points will seem correlated but they are not significantly so, and visa versa.

You can also see the regression equation, which is the equation of the trendline.

For this specific example, the relationship between latitude and temperature is fairly strong at an R-squared at 0.8, and a highly significant p-value. It is a negative relationship, as indicated by the negative intercept in the equation. The lower the latitude, the warmer the winter.

24. Let's go through the types of analysis in a bit more detail. The first type of analysis is Chi-squared, which looks for deviations from expectations as a result of chance. As I pointed out before, the null is that there is no difference, in this case between the expected and the observed results, and the alternate is that there is a difference between the two.

This is good for comparing proportions and is easy to calculate by hand.

We're not going to go through how to do this particular test in Excel since there is not explicitly this option, but I wanted to point it out as statistical test available for a given data set. To do it, you would set up a series of equations in cells to calculate the necessary values.

25. The last type of analysis is a linear regression, or correlation. This looks at how strongly two variables are correlated or associated with each other. If the variables are correlated, they form a linear relationship. The Null hypothesis is that there is NO relationship between the two variables, that they are independent of each other. The alternate hypothesis then, is that there IS a relationship between the two variables.

26. This Excel output example is for the frequency of cricket chirps in relation to the ambient temperature.

The R-squared is 0.697. For the regression equation, you look at the coefficients of the intercept and the x-variable.

To determine the significance, you look at the "Significance F" value, which assessed the same way you would a P-value. In this case, the results are significant, so there IS a correlation between cricket chirp rate and temperature.

27. Ok, now it's your turn. Make sure you are on the Regression worksheet. Open StatPlus with your excel sheet in view. Click on the Statistics Tab, then choose Regression, under regression, choose Multiple Linear Regression.

One thing I want to point out is the "HELP" option, which is great if you want a description of each test to determine if it's appropriate.

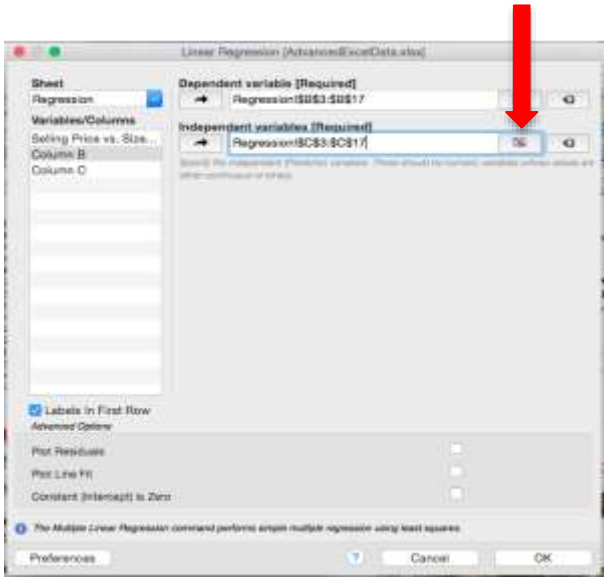
Scroll down and select the Regression analysis.

These data are for the relationship between the selling price of a house and the size of the house.

28. Before you can run your regression, you'll need to determine which variable is the independent (x-axis) and which is the dependent (y-axis) variable.

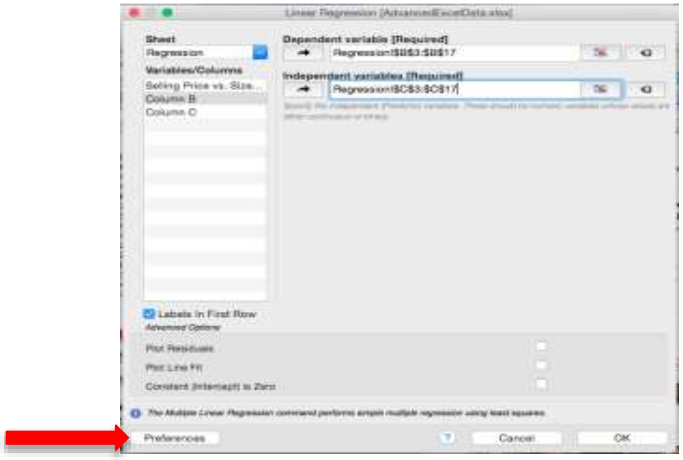
Now, as you did when you selected the desired data in the Basic Excel workshop, you select the desired data for the input Y and X ranges. Make sure "labels" is selected if you want to include the headers, which is convenient for output labeling.

To highlight the dependent and independent variables in StatPlus, choose the grid icon to the right of the popup screen and it will direct you to your excel sheet. Once you have highlighted what you want, click back on the StatPlus popup menu (See below).



Set the confidence, or alpha level to the level you want, and enter the name you want for the output worksheet, then hit OK.

To choose the confidence interval, click prefernces in the bottom left corner of the popup menu for regression in StatPlus (see below).



(*) Answer these questions on your assessment: What is the R-squared? Is the regression significant? What does this tell you about the relationship between house selling price and size?

29. Next, let's start with a t-test first. As I said, this compares two means and distributions. As with many of the topics we're covering, there are various types of t-tests, but we're going to focus on two-tail t-tests.

Again, the null hypothesis is that there is no difference between the means and distributions, while the alternate hypothesis is that there IS a difference between the means or distributions.

30. When you do a t-test in Excel, this is what the output will look like. You can see the mean and variance, as well as the number of data points. The default hypothesis is that the mean difference is zero, but you do have the ability to change this. You can see the degrees of freedom listed, as well as various test statistics. We're focused on the two-tail p-value, which in this case is 0.000474. Assuming our chosen alpha is 0.05, the null hypothesis is rejected because there is a significant difference between life satisfaction in older adults versus younger adults.

31. Make sure you are on the t-test worksheet. This is data for two different groups and their exam scores.

While on the t-test worksheet, open StatPlus and Under Statistics, choose Basic Statistics, then choose Compare Means (T-test). A window will pop up. Change the t-test type to "t-test assuming unequal variances" from the dropdown menu at the bottom of the pop up window.

This is the conservative approach, versus assuming equal variance.

Click OK.

32. You'll have to select the numerical (not categorical) data for each data set, in this case you are looking group 1 and group 2 scores. If you leave the hypothesized difference blank, it will assume the null hypothesis is no difference.

As before, if you have header rows, that's when you'd select tables, and make sure your alpha level is at the level you want it to be.

Lastly, enter a name for the output worksheet, something like t-test results.

Hit OK.

(*) A new worksheet is created with the results. Take a second and answer the question on your assessment sheet. What is the resulting p-value of the t-test? And, based on this, are the means of the two groups significantly different from each other?

33. Lastly, we'll go over ANOVA, which is essentially an extension of the *t*-test. It is used for comparing the means of more than two groups. Unlike the *t*-test, an ANOVA determines if all the means are equal or at least one of the means is different. The test does not tell us which one is different. The simplest way to tell which data sets are different from one another is by using a series of *t*-tests to perform pairwise comparisons.

34. As before, this is what the output will look like for an ANOVA. This is from a comparison of water retention in 5 different brands of concrete. You can see the mean and variance, as well as the number of data points for each group. The p-value between groups is 0.0088, so the null hypothesis is rejected and there IS a significant difference between at least one pair of values. However it doesn't tell you where the significant difference is.

35. Ok, your turn now. Make sure you are on the ANOVA worksheet. These data are for the effects of feed type on animal weight gain. Under Statistics in StatPlus, choose Analysis of Variance (ANOVA). Choose 1 –way ANOVA.

36. The input range will be for the entire data set, and if you have headers (in this case, A-D for type of feed), make sure the correct alpha level is indicated, and make a name for the output worksheet. Hit OK.

(*) on your assessment sheet, what is the resulting p-value of the ANOVA? Are ANY of the groups significantly different? And, are ALL of the groups significantly different?

37. Just as a quick reminder, you can get Microsoft Office Pro for free for being a UNM student. We have instructions for this, along with this presentation is available as a PDF on the EASE website if you need to refer back to it. There is also a file that guides you through the differences between PC and Mac versions of what we went through.

Lastly, there is a document with links to videos on basic Statistic concepts, if you want to learn more about what is underlying each of the tests we learned how to perform.

38. If you don't remember, the Basic Excel workshop that you took in 203 is available through the EASE workshop page. We are not going to spend time going over topics covered in that workshop, but many of them are beneficial for you here in 204, so make sure you go back and look through that presentation and supplemental information.

Topics that were covered include graphing, as well as the Filter and Sort functions, which will really help you this semester when it comes to cleaning up your dirty data before you can move forward with any analysis for your homework's.

39. Any questions?

For the survey, on your data set, there is a tab that says "Exit Survey" that will take you to it. When you are done with that and your assessment, please call me to see the confirmation page for the survey and I'll take your assessment, then you are free to go.