

Statistical Considerations

Research Project 2010-2011

In order to convince others (especially the scientific community) that your research project (and your conclusions) are valid, you need to communicate at least two things:

1. The error, or uncertainty, that you have in your data.
2. How representative your data, or correlation, is of all possible data points that exist (for in your study you only collected a finite number).

The difficulty here is that you really know nothing about the population you are sampling. If you did, you likely wouldn't be doing your investigation in the first place. When you measure something, you are sampling. You might be measuring the calories contained in a nut, the density of an algae culture, the growth of a plant within treated soil, the concentration of arsenic contained in a soil sample, or the concentration of H_2S present in a gas sample. The actual samples you measure and record are only a few of the possible samples that might actually be collected. The body of entire samples is referred to as the "population" that you are sampling from.

For your study, you are to utilize at least one of the two tools described below and demonstrated on the accompanying excel spreadsheet. If what you are doing doesn't "fit" the tools described below, talk with your instructor. We are going to focus on two techniques that will address most group concerns regarding the analysis of their data. In both cases, we will take advantage of Microsoft Excel's built-in statistical tools.

REGRESSIONS: Finding the correlation between two factors. This test is performed to not only measure the strength of a correlation between two sets of data, but also to generate a mathematical function that might then be utilized to calculate or predict unmeasured values. On the spread sheet, the "regressions" worksheet lists two sets of data that has been collected – in this case leaf length and gall mass. The experimenters wish to know if their data actually reflect a correlation between leaf length and gall mass. One approach is to simply plot the pairs of data using a scatter plot on Excel. If there is a linear relationship between leaf length and gall mass, we would expect the data points to make a linear pattern on the resulting plot. In this case, if you look at the "linear regression plot" tab, you can see that the data is scattered all over. When we ask Excel to "fit" a trendline to this data, we are really asking Excel to see if a linear correlation exists. In such a correlation, when the independent variable (the one you are changing) is increased, the dependent variable (the one you are measuring) would also increase (or decrease) in a predictable, consistent manner. Under the Chart menu in Excel, you can ask Excel to try and fit a mathematical model (linear is one option) to your data with the "Add Trendline" command. You can also ask Excel to display the equation and the R^2 factor. The closer R^2 is to 1.0, the better the correlation between your two sets of data. The data in the sample spreadsheet, with an R^2 factor of .1541, does not represent a good correlation (at least to the linear model that Excel has attempted to create for the data set). You are free to attempt other mathematical models – there are several options listed under "Add Trendline". Long story short – you are seeking the model that results in the best R^2 value. If you find a mathematical model that "fits" your data well, in theory one would then be able to use the equation to predict other combinations of data points **EVEN THOUGH YOU DON'T GO OUT AND MEASURE THEM ALL**. The quick and dirty instructions for doing regressions on Excel:

- Place the two sets of paired data into adjacent columns somewhere on the spreadsheet.
- Use the Chart Wizard to generate a scatter plot of your data.
- Once completed, use the "Add Trendline" under the "Chart" menu to add a trendline. Be sure to select the options tab here and have Excel show the equation and the R^2 value.
- Discuss R^2 term – If R^2 is greater than about 0.60 (60%), there is a pretty strong correlation.

T-TESTS: Sometimes an investigation simply attempts to determine if there is a statistically significant difference between two sets of data. Look at the "T-test" worksheet on the accompanying

Excel spreadsheet. Here, 20 samples of soil have been collected – 10 from Sweet Home and 10 from Corvallis. The amount of arsenic in these soils has been measured. When the average arsenic concentration is found for the 10 Sweet Home samples, and then the 10 Corvallis samples, they are found to be different. Is this because there is actually more arsenic everywhere in Sweet Home versus Corvallis? Or is the fact that the two averages are different just due to chance and the fact that only ten samples were taken from each location?

T-testing is a statistical tool that allows one to begin answering these questions. T-testing factors such things as the variance in the data points, the difference between the two average values, and the number of data points that were taken. Smaller variances within each data set, larger differences between average values, and larger numbers of data points usually yield more definitive results. Excel makes this tool rather simple to use. The example on the spreadsheet is fairly self-explanatory. When entering a formula in a cell in Excel, remember to first put an “=” so Excel knows to expect a function or mathematical operation. After you specify your two columns of data, be sure to include the “2” and “3” as shown below. To learn what these do, use the Excel Help menu and query “Ttests”.

The procedure for t-testing on Excel:

- Set up your groups of data in columns.
- Use the AVERAGE function to calculate the average.
- Use the STDEV function to calculate the standard deviation.
- Perform a t-test in some cell by entering something similar to the following (you of course will refer to the cells with your data): =ttest(F5:F14,G5:G14,2,3)
- T-test: If T-test result is less than 0.05, the difference is significant. If the T-test result is 0.05 or greater, there is not enough data to be sure and therefore don't be so quick to make any conclusions.