Ecology Lab 3. Statistical Analysis and Testing in Ecology

A few weeks ago, we measured trees and compared morphological traits between two different species. By now, you've written up a lab report in which you have determined whether the two species differed in these traits or not. How did you come to that determination? How different did the averages for the two species have to be before you decided that the species were different? This is the kind of question that statistical analysis gives us the answer to.

Here's another example. Suppose that we were interested in the population density of a mayfly nymph in two different streams. We go out to the streams and collect all the nymphs in a 1 m^2 plot in each stream. One stream (X) yields 76 nymphs, and the other stream (Y) yields 84 nymphs. Do these streams have different densities of nymphs? A statistical test might be able to tell us, but with only one sample from each stream, any conclusions would be very tenuous. So, let's say that we actually collected 5 times from each stream, at five similar though not identical locations (or on five different days). If we did that, we might have a data table like the one below:

Stream	Collection 1	Collection 2	Collection 3	Collection 4	Collection 5	Mean
X	76	111	43	12	138	76
Y	84	119	51	20	146	84

Descriptive Statistics

Our first task is to describe these data. That is, we want to produce a smaller set of values that summarizes the data. We want to summarize not only the "central tendency" of the data (i.e. the mean or median), but also the variability. The mean is simply the average, and is valid if the data are more or less equally distributed around the mean (i.e. a normal distribution). The median is the "middle" value obtained during data collection – thus, for the five collections above, the median for X is 76 and the median for Y is 84 because, for each stream, there are two values lower and two values higher than these medians. The fact that, in this case, the median is exactly also the mean is coincidence. For example, if Collection 5 were 139 and 147 for X and Y respectively, then the Means would be higher than they are (by 0.2), but the medians would still be 76 and 84.

The variability can be given as the Range (the highest value minus the lowest) for each stream (i.e. 126 for each stream), more often though, variability is presented as the Variance or the Standard Deviation (squareroot of the variance). Sometimes, variation is shown as the Standard Error, which is the standard deviation divided by the squareroot of the sample size. Standard errors are often used with figures, while standard deviations are often given in tables.

For the data above, this summary table would result:

Creek	Mean	Variance	Standard Dev.	Standard Error
X	76	2563.500	50.63102	22.64288
Y	84	2563.500	50.63102	22.64288

Now – the fact that the variance, standard deviation and standard error for each stream is the same is because I set the data up that way. Generally, these values will be different for each group of samples. Now – back to our original question…Do the streams differ in the number of mayfly nymphs? It certainly seems like the streams might differ – by about 8 nymphs, with stream Y having a higher density, right? This is where statistical testing comes in. In statistical tests, we set up a null hypothesis and test whether our data fit that null hypothesis. Actually, what we get from a statistical test is a probability value (or P-Value). What this P-value tells us is this: **The P-value is the probability that we could get data like we did, just by chance, if the null hypothesis were TRUE.** In science, because our null hypothesis is generally one of no difference or no effect, what we are really trying to show is that the null hypothesis is NOT TRUE (or not supported). Thus, we are looking for P-values that are very low (i.e. there is a very small chance that we could get data like we did by chance if the null were true…therefore, perhaps the null hypothesis is NOT TRUE - and we can accept one of our alternative and more interesting hypotheses). In science, it is accepted that if the P-value is less than or equal to 0.05 (i.e. 5%), then we reject the null hypothesis and embrace an alternative – and we say that the data are statistically significant. Think about this…if we run a statistical test and get a P-value that is very low (less than 0.05), then the probability that we could have come up with data like ours, by chance, if the null were really true is very low. Thus, it is likely that the null is false, and an alternative hypothesis is true.

So – what kind of tests are we going to do? And how are we going to do them? First, there are many, many different kinds of statistical tests, and more are being developed all the time. However, this is not a statistics course, so we'll just touch on some of the basic tests, and we'll even ignore most of the assumptions necessary to make these tests truly valid. One simple test for determining whether two groups of samples are from the same distribution or not (i.e. whether they differ or not) is the t-test. The t-test uses the variances to see whether or not the means differ. For our data above, the t-test results in a P-value of 0.809. Put another way, if the null hypothesis is true (i.e. the streams do not differ in nymph density), then 80.9% of the time that we sampled these streams, we could get data like we did just by chance. This would suggest that perhaps the null IS true. The reason is that the variation within streams is much greater than the variation between streams – so the data are not statistically significant.

A second type of test that we might use (the Mann-Whitney U test) requires fewer assumptions about the underlying data. This test is called a nonparametric test, because the data values are transformed into ranks prior to analysis. For our data, the M-W U test results in a P-value of 0.602. Again, the results are not statistically significant, and we would have to say that our data support the null hypothesis.

However, let's consider a similar dataset:

Stream	Collection 1	Collection 2	Collection 3	Collection 4	Collection 5	Mean
X	76	77	75	74	78	76
Y	84	85	83	82	86	84

Again, we have the same means and the same medians as before. However, now the variation has decreased substantially. The new summary table is:

Creek	Mean	Variance	Standard Dev.	Standard Error
X	76	2.50000	1.581139	0.707107
Y	84	2.50000	1.581139	0.707107

A second t-test on these data results in a P-value of 0.00044, clearly much lower than 0.05 and clearly statistically significant. Likewise, a Mann-Whitney U test results in a significant P-value (P=0.009). Thus, although the means of the two data sets were identical, because the variances were lower in the second data set, we have statistical evidence that the two streams do differ in mayfly nymph density. Generally, differences like this are shown graphically with bar graphs (of means), with standard error “bars”.

However, what if we were more interested in the relationship between two variables – say, nymph density and water depth? In this case, we would perform a correlation analysis to see if there was a correlation between the variables. Correlation coefficients (r) can range from -1.0 (perfect negative correlation) to 1.0 (perfect positive correlation). The statistical package will let you know whether the correlation coefficient is significantly different from zero (i.e. whether there is a statistically significant correlation).
Consider the data below:

Coll.#	1	2	3	4	5	6	7	8	9
nymphs	47	33	98	25	55	47	63	69	11
Depth (cm)	16	15	9	16	17	15	10	12	18

For these data, r = -0.84, and P = 0.0045. Thus, the correlation is negative and statistically significant. This means that larger numbers of nymphs are associated with shallower water. Generally, correlations are presented either as tables alone or accompanied by scatterplots with one variable on the X-axis and the other on the Y-axis.

For the data above, it may be that water depth directly influences nymph density, or perhaps water depth doesn’t matter at all, but other factors that do matter to nymph density are correlated with water depth. For example, food supply would be expected to influence nymph density. Perhaps food supply is correlated with water depth (or with water current velocity – which itself is correlated with depth). If we could measure food supply, something we feel definitely influences density, then we could do a regression analysis and draw a line (regression line) representing the influence of food supply on nymph density. Let’s say we could measure food supply as mg of carbon/ml of water. Perhaps the following table results:

Coll.#	1	2	3	4	5	6	7	8	9
nymphs	47	33	98	25	55	47	63	69	11
mgC/ml	66	45	79	41	68	61	70	75	32

If we perform a regression analysis, what we’re doing is fitting a straight line through these nine points. In this case, our **independent** variable (X-axis) is the amount of carbon present, and the **dependent** variable (Y-axis) is the number of nymphs (because we expect that the number of nymphs **depends** on the amount of carbon, not vice versa). A regression analysis is usually presented as a scatterplot with the associated regression line (in the form Y=mx + b), where m is the slope of the line, and b is the Y-intercept. Generally, there are the results of a statistical test on the slope (m) to determine whether it is significantly different from zero. For the data above, the correlation coefficient is 0.951, the slope is 1.51, and the intercept is -39.18. The statistical test on the slope is highly significant (P = 0.00008).

Data Collection

OK – today we'll collect data and use all of these techniques to analyze them. Once again, we're going to do trees. We are going to analyze the morphology of cottonwood leaves from two samples of trees: old trees near water (the irrigation ditch), and young trees away from water (though irrigated – the trees we used last week). We'll break into groups of four. Each group will go out and collect 10 leaves from trees near the ditch, and 10 leaves from trees up at Peaks to Plains. Just pick the leaves off with the stems attached, and we'll cut the stems off in the lab. Bring the leaves back to the lab and measure (to the nearest mm) the length along the mid-rib from the bottom of the leaf to the tip (don't use the stem), the maximum width at the widest point and the mass of the leaf. Be sure to keep all these measurements together – so that for a single leaf you know the length, width, and mass. After you've measured all the leaves and recorded the data on a data table, I want you to cut identical-sized samples from the leaves and weigh those pieces (you don't need to know which leaf each piece came from). To cut these samples – stack 5 leaves from one species on top of each other, lay a plastic ruler across the stack, and cut with a scalpel down both sides of the ruler. Discard the pieces that aren't under the ruler. Then, turn the ruler 90-degrees so that it runs across the "strip" that you have left, and cut again on both sides. This should leave you with a stack of leaf pieces that are more or less square and all the same size. Do this for a second stack of five from that type of tree, and then do it for two stacks from the other type. Weigh these squares and record their weights. After this, you'll be done with the data-collection portion of the lab. Now you get to analyze these data, and write up a report to turn in next time at the beginning of lab.

Data Analysis

You will do your data analysis with a statistical program called SPSS. SPSS is loaded on all the computers accessible to students on campus (or at least, that's what I was told). The main level of the library has several that have it, and I believe that the lower level does too. To access SPSS, go to the "start" button (lower left) and click it. Choose "all programs" and find "SPSS" on that list – click it. You will get a box up that asks you to click in a circle to tell SPSS what you want to do – you want to "Type in Data" so click that circle, and you'll get a spreadsheet. There are two tabs at the bottom of the spreadsheet: Data View and Variable View. Click on the "Variable View" tab, and under "Name" for column 1 type "TYPE". (note – on this spreadsheet, each row will represent a column on the "data view" sheet – so in the first cell under the "name" column, you will type "TYPE"). Under Name for column 2 type "LENGTH" (this will be the first cell in row two). Under Name for column 3 type "WIDTH" (first cell in row 3, etc.). Under Name for column 4 type "MASS". Under Name for column 5 type "SQMASS". Click on the "Data View" tab, and the names that you just typed should be above the appropriate columns. You are now ready to enter your data. We will code the two tree species as "1" and "2" for old (near ditch) and young (from park) respectively – so the first column will have ten 1s and then ten 2s representing the ten leaves from each type of tree (old vs. young). Now, keeping all the data from a single leaf together in a row (excepting the SQMASS data), enter your lengths, widths, and masses in the appropriate columns. All that matters when you enter the data for SQMASS is that you keep the type of tree correct. Once you have entered all your data, you are ready to begin the statistical analysis.

The first thing that we want to do is describe these data. Go up and click on "Analyze" from the toolbar at top. Then, click on "Reports", and then "Case Summaries". Now, highlight your grouping variable (TYPE) and move it into the lower box by clicking on the arrow icon. Now highlight and move all of your dependent variables (LENGTH, WIDTH, MASS, SQMASS) into the upper box. Go to the bottom of the window and click "Statistics". From the list on the left, move the mean, median, standard error, standard deviation, and variance into the box. Click "Continue", then click "OK" and the computer should compute and show you the values for each of the dependent variables separated by tree type (1s and 2s). This will come up in a brand new window. Either copy these numbers down somewhere, or wait until you are finished and print the entire file out (SPSS will keep adding new results to the bottom of this page each time you run a new analysis – so at the end of the entire thing, you could still scroll back up and see these first results, or print it all out).

To calculate the t-test, click on "Analyze" then "Compare Means" then "Independent Samples, t-test". Select the grouping variable (TYPE) and move it into the lower box. Then, go down to "Define groups" and click on that. Put in the "1" and "2" in the appropriate box (Group 1 and Group 2) to define your two groups and click "continue". Highlight the "Test" variables you are interested in (your dependent variables), and and move them into the box (move all four in). Click "OK" and you should get your t-test output. You can ignore the results for "Leven's test", and just look at the "2-tailed sig" to find the P-value for this test.

To calculate the Mann-Whitney U test, go to "Analyze", then "NonParametric Tests", then "Two Independent Samples". Be sure that Mann-Whitney U is checked at the bottom (it should already be). Move the grouping variable over as before, and define the groups. Then, move the dependent variables over and click OK – you should get your MW-U test output.

For each of the next two analyses (Correlation and Regression), you'll need to "Select Cases" first. The reason is because you don't want to mix the types together for these analyses. You will have to first select the "older" trees, then run both correlation and regression, then select the "younger" trees, and run the analyses a second time. To select either the younger or older group, go to "Data" on the toolbar, and go almost to the bottom and choose “Select Cases”. Click on the open circle (second from top) “If condition is satisfied”. Then, click on the box immediately below it “IF…”, and you’ll get a new window. Highlight “TYPE” and move it into the box and then type in “=1” after the word “type” in the box. Click on “Continue”, then click “OK”. This should return you to your data, and the trees of “TYPE” 2 should have slashes across them, indicating that they won’t be used in any further analyses (until you change the “select cases” dialog box again). Now do the following two analyses for this tree type, then change the select cases so that you choose tree type #2, and run both analyses again.

To Calculate the correlation, go to “Analyze”, “Correlate”, then “Bivariate Correlation”. Be sure “Pearson” is checked (it should be). Move the variables over that you want to correlate. The variables that you will use for the correlation analysis are “Length, Width, and Sqmass”. Click OK. This should return your correlation matrix for these three variables for this tree type.

To calculate the regression analysis, you will go to “Analyze”, “Regression”, and “Linear”. Move the dependent variable over (MASS), and then move independent variables over singularly and run the analysis – you should do three analyses for each TYPE – one with “Length” as the independent variable, one with “Width”, and one with “Sqmass”.

The Lab Report

Materials and Methods (4 points) – summarize exactly what was done in this lab, how the data was collected and how it was analyzed. You do not need to go through the steps necessary to analyze the data (i.e. what you clicked on to run SPSS), but you do need to mention the tests that you did.

Results (8 points). This section will be fairly extensive. In the text portion, you should simply summarize your results, and refer to the appropriate tables and figures. Do not give each number here and on the tables or figures – just give a text summary of the results and refer to the appropriate table or figure (i.e. “*Older cottonwood leaves averaged longer than younger cottonwood leaves (Table 1). Length and Width were more highly correlated in younger trees’ leaves than in leaves from older trees (Table 2).*” etc.). I expect at least two tables: one with the descriptive statistics and the results of the t-tests and MW-U tests, and one with the correlations. I also expect several figures showing the regressions. It might be most interesting to have three figures, each figure representing the regression of one of the three independent variables on MASS, with two sets of points (and their associated regression lines) on each – one for each TYPE of tree. So, to summarize your results section, I think you need a paragraph or two of text, two tables, and three scatterplot figures with two regression lines plotted on each figure.

Discussion (8 points). I want you to show me that you have thought about your data and results, and you have some feeling for what each of your tables and figures says about the leaves from these trees. This analysis will tell you something about the shapes and masses of these leaves, and how the two types differ in these respects – but it won’t tell you much about the ecology. Thus, I expect you to discuss the morphology and not really hypothesize about what these traits might mean…but you are free to do that if you want to, and if you can support your hypotheses with reasonable assumptions. Also, if you want to propose other analyses, or discuss problems with data collection or some other aspect (especially if you have suggestions for improvement), feel free to do so. As always, be sure that your spelling and grammar are acceptable.