

Beiser Field Station Transect Analysis

Due Date: _____

The goals of this assignment are to:

1. Introduce or reinforce basic statistical and data analysis skills.
2. Apply critical thinking to the refined data to answer questions based on the evidence collected.

One of the goals of a scientist is to be able to answer questions with the greatest possible reliance on observable facts, and the least reliance on intuition. While intuition has great importance in finding the right questions to ask, and in finding ways of investigation, once the data is gathered the scientist should rely on the facts at hand. Patterns in the data may be revealed through good graphical analysis, and the patterns should then be tested with statistics to see if they are “real” – or simply the result of the scientist looking at the data and “seeing” a preconceived result. This is an example of bias; one shields oneself from bias by using commonly agreed upon statistical tests as impartial arbitrators of what is “real”

Sometimes the results are unambiguous. Every time you drop a penny it falls to the ground. No one needs statistical analysis to prove the existence of gravity. On the other hand, sometimes the penny lies heads up, sometimes heads down. Determining if this is a random event or something influenced by other factors may require the application of statistics; statistics are also useful to draw conclusions about a larger population by sampling a smaller portion of it.

Results in biology are seldom so clear-cut as to eliminate the need for statistics. There are several basic tests and graphical analyses that should be in every biologist’s “toolkit”. Among the graphing techniques are:

1. The **scatterplot**, which is used to look for correlation between two variables, or to track a variable over time.
2. The **trendline**, which is the superposition of a line drawn from a mathematical model over a scatterplot.
3. The **histogram**, which is used to look for patterns in abundance.

Any pattern that is revealed by the graphical analysis should be examined by statistical tests to see if the pattern is “real”. In most cases, this means determining if the pattern is different enough from what might be expected in a random world. For instance, flipping 51 heads out of 100 tosses would not be unexpected; flipping 80 heads out of 100 tosses, or flipping 20 heads in a row might be unexpected and suggest that something else is at work. The statistical tests that will be of the most use to you in testing apparent patterns are:

1. The **t-test**, which is used to tell if two averages (**means**) (the composite of many measurements), differ in a statistically significant way.
2. The **ANOVA** test, which is kind of a “super” t-test to tell if any of a group of mean values differs from the rest. If the results are positive, you then have to go back with multiple t-test and see which mean or means is different
3. The **correlation coefficient**, which is used to test the statistical significance of a trendline.
4. The **Chi-square** test, which is used to determine if experimental results differ enough from expected results to suggest “real” difference.

In this introductory exercise, you will use data you generated in an exercise at the field station, as well as some data from transects done by Marietta College students in Costa Rica. We will be concentrating on the Correlation coefficient, ANOVA and the t-test to analyze our data from the transects.

First, a reminder of the sampling technique we employed:

Point-Quarter Method

Today's goal: Learn how to accurately determine location using GPS. Learn how to use a point-quarter technique. Learn the principles of transect techniques as an alternative to plot techniques.

Background: Many of the techniques for sampling in the field deal with various plot techniques in which a series of plots (areas) are sampled. Plot techniques are powerful, and have extensive applications in ecology. Still, there are limitations. For one, it is difficult to locate plots randomly, as obstructions in the field often prevent accurate placement of the sampling grid, or at the least may hinder sampling. To overcome this difficulty, many plots must be sampled. Secondly, the intensive nature of the sampling involved means that relatively little overall area can be covered. If you have to sample a lot of plots (to overcome the difficulty of random plot sampling) it will take a lot of time, and you won't cover much ground.

For certain applications, another type of technique can be used. Transect techniques involve sampling along a line. There are many variations, but most involve stopping at regular (or random) intervals and collecting data. For instance, to measure light, temperature, soil pH or moisture, or other conditions, one might extend several transects through a forest, field (or even a body of water), and simply sample every 10 feet or so.

In today's lab we will use a point-quarter technique to estimate the growth of forest in two situations. In a point-quarter technique, the transect line serves as a base for sampling. At every sampling interval, the world is divided into 4 quarters, using the transect line and a line drawn perpendicular to the transect line to divide up the quarters. Light, pH, and soil moisture are taken at the transect point; the tree closest to the transect point in each of the quarters is measured as is the canopy cover in each quadrant.

For our purposes, we will use a simple technique to determine distance. From a starting point, follow a compass line to form the transect line. Stop after a random number of meters (determined in advance) to take the measurements. First mark the location with a piece of tape so if you have to come back to take a measurement you can locate each point. To form the quarters, one student should face in the direction of the transect line and spread his or her arms out straight to either side. Take and record all measurements. Tree diameter and/or circumference should be measured 4 feet off the ground. Record the tree species if possible. If one of these measurements cannot be made, it can be estimated later. Record all data in pencil (pen ink can run if it gets wet). If a sampling point falls on a tree, skip that point (record a blank) and go on to the next sampling point.

In terms of data analysis, we will look to see if average sizes and canopy cover differ in each of the two situations; or if pH, soil moisture, and if other readings vary in a predictable pattern.

Sampling Rules:

1. If there is no tree within _____ meters of the point in a quadrant, skip the tree measurement for that quadrant.
2. A tree must be at least _____ meters tall to be measured.
3. The random number used between points is read off the random number sheet. If it is smaller than _____ or bigger than _____, skip it and use the next number.
4. Sample for _____ meters or _____ points, whichever comes _____.

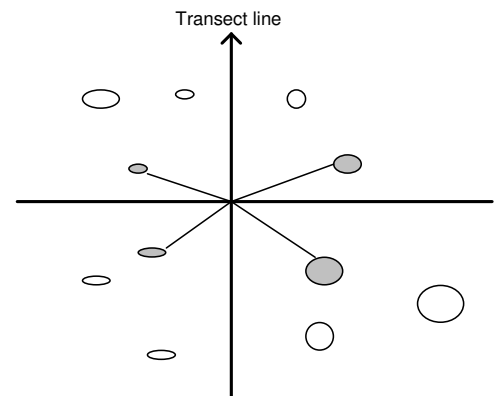


Figure 1 - Diagram of Point-Quarter Technique. Trees to be measured are shown in gray.

Background - The ANOVA and t-test

ANOVA

Let's say we are trying to determine if there is any difference in the average temperature of 4 transects taken in two different forests.

Our raw data looks like this:

Temperature - Degrees F				
	Dry			
	Rainforest 1	Rainforest 2	Forest 1	Dry Forest 2
	82.2	85.3	78.7	76.9
	80.5	80.4	75.4	75.9
	79.7	78.6	77.3	76.9
	81.3	79.6	76.9	77.9
	80.7	81.5	78.3	77.1
	79	85.4	77.7	77.2
	79.6	81.3	72.3	77.3
	80.6	81.4	75.8	76.8
	81.3	82.6	75.1	76.9
	81.9	82.6	76.3	76.9
Average	80.68	81.87	76.38	76.98

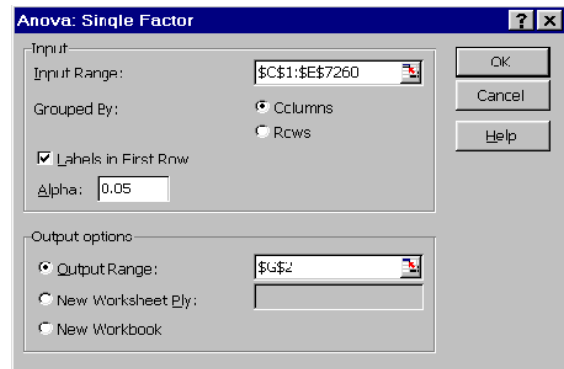
Of course, that question was pretty easy to answer even without doing the statistics. There is a difference between each of these transects; none are the same. A more difficult question is: "are any (or all) of these means **significantly**¹ different from each other?" Think of it this way – minor fluctuations, electrical glitches, software errors, etc. could all lead to apparently random differences in temperature. Looking at the data, we would guess that there is a statistical difference between the means, but we really should test to be sure (and allow us to determine if the difference is **significant**). The first test we will run is the **ANOVA**² test. The ANOVA test allows us to quickly test multiple samples to see if **any** of them are significantly different. If so, then we must run multiple *t*-tests to determine which means are different – a *t*-test can only be run on two sets of data at a time.

¹ The word "**significantly**" in science is used ONLY when statistics have been run on the data. In that sense it is a "code" word, and when a scientist uses it other scientists assume the statistics have been done. When a layperson or politician uses the term "significant" a scientist assumes that the answer has been pulled out of someone's anus.

² ANOVA = Analysis Of VAriance

To do the ANOVA:

1. Select Tools:Data Analysis from the menu.
2. Choose ANOVA: Single Factor.
3. Fill out the form as shown to the right.
4. Click OK



The ANOVA table will be generated; a sample is located below.

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Rainforest 1	10.0000	806.8000	80.6800	1.0618
Rainforest 2	10.0000	818.7000	81.8700	4.8868
Dry Forest 1	10.0000	763.8000	76.3800	3.5018
Dry Forest 2	10.0000	769.8000	76.9800	0.2484

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	220.0208	3.0000	73.3402	30.2472	0.0000	2.8663
Within Groups	87.2890	36.0000	2.4247			
Total	307.3098	39.0000				

In the summary portion, the ANOVA table repeats some of the information of the descriptive statistics, such as the count, the mean, and the variance for each of the columns. The true ANOVA table comes next. The SS column refers to the sum of squares, and is basically the squared difference between (or within) the groups (each data point is subtracted from the mean and the result squared). The *df* refers to the degrees of freedom; with 4 groups there are 3 degrees of freedom, and within a group the degrees of freedom are equal to the number of measurements minus 1. Don't worry about the MS (which is the SS divided by the degrees of freedom and represents the "mean error"). Focus on the *F* value. If the *F*-value is larger than the *F crit*, then there is at least one pair of means with a significant difference. The *P-value* gives the chance of making a Type I mistake, where you assume the means are different when in fact they are the same (and random chance in sampling or measurement makes them appear different). In this example, the *F*-value is much greater than the *F crit*, so we reject the hypothesis that all 4 means are the same. At least one of the means is significantly different from one of the others. We will have to turn to t-tests to ferret (*Mustela nigripes*) out which means are different.

t-test.

The t-test allows us to narrow down which means are different, but in contrast to the ANOVA, the t-test is limited to testing 2 sets of data at a time. The *t*-test helps you answer the question "Are the means of these two data sets the same or not?" Or, to be more precise, the *t*-test allows you to reject the hypothesis that the two data sets have the same mean with a certain chance of making a mistake. The possibility of making a mistake comes about because of the variation within natural populations. If you wanted to compare the heights of people in two different cities, you might watch 100 people pass through a doorway with the heights marked on it. If, by chance, in one city you did your measurements while an elementary school went on a field trip, and in the other city you

caught the athletes at the city basketball tournament, you would conclude (incorrectly) that the two cities had different average heights. To protect against making this type of mistake you set a benchmark – the alpha (α) value at a high level. If you set it at 5%, that means there is only a 5% chance that you might erroneously conclude that the means are different when in fact you just had bad luck in sampling.

Two key things about t -tests. First, if you have more than 2 data sets, do NOT run t -tests unless you first run an ANOVA *and* your ANOVA shows that at least two of your means are different. Why? Consider this case: You set the alpha level to 0.05, which means that 1 time in 20 you expect to see a result showing a significant difference when in fact the difference is due to sampling error or other “noise” in your data. If you do 20 t -tests to compare different sets of data, chances are at least one of those will come up positive! If the ANOVA says none of the means are different, stop there. Second, if the ANOVA tells you there are some differences, start with the most obvious – the smallest and biggest means. That is most likely where significant differences lie. Of course, if you only have two samples you can just go ahead and run the t -test – no need to do an ANOVA .

t-Test: Two-Sample Assuming Unequal Variances

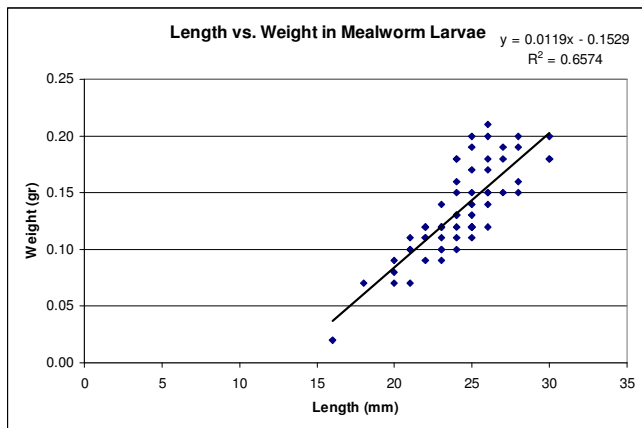
	<i>Rainforest 2</i>	<i>Dry Forest 1</i>
Mean	81.870	76.380
Variance	4.887	3.502
Observations	10.000	10.000
Hypothesized Mean Difference	0.000	
df	18.000	
t Stat	5.994	
P(T<=t) one-tail	0.000006	
t Critical one-tail	1.734	
P(T<=t) two-tail	0.000011	
t Critical two-tail	2.101	

The t -test works by mathematically comparing the variances within the two samples with the difference in their means. The number that results from this is compared to a table of values computed for each possible alpha value. Of course, the computer doesn't have a table to go to; the program generates the value on the fly. In Excel, you get a printout like the one above. The important numbers to look at are the t Stat, the P values, and the t Critical values. The t -stat is the number generated by the computer based on your data. The bigger it is, the greater the significance of the difference between the means. The P values tell you the chance of erroneously saying the means are different. The smaller the number the better; you want it at least to be smaller than your alpha value. The t critical numbers are from the table generated by the computer. If your t Stat is greater than the t critical value then you can assume that the means are different with a chance of being wrong due to unlucky sampling of less than the alpha value you selected. The P values give you the exact chance of making that type of mistake; in the example above it is 0.000011 (not much of a chance). In this case, we reject the hypothesis that the means are the same, and we're pretty confident that the difference is real, not due to chance (there is a 1.1 in 100,000 probability that the difference is due to chance). By the way, work with the absolute value of the t stat; that is ignore any minus signs.

What about the 1 vs. 2 tails? To put it in a nutshell, use the 1 tail test when you can predict the direction of the difference between the means. If you have been feeding one group of mealworms twice as much as another group, you would expect the group being fed to be heavier, and you would use a 1-tail test. On the other hand, if you were just comparing 2 populations of mealworms and knew nothing about their living conditions, you would have no way of knowing which population was eating better and therefore would be heavier. You would use the 2-tailed test. In the case above, it is really silly to compare the temperatures in the two forests (they were miles apart and the measurements were taken at different times on different days), so we have no expectations; I used the 2-tailed test in choosing the P value.

Of course, in this assignment you will be using ANOVA and T-tests to determine if there are any differences in environmental parameters, tree density or size between the plots.

Correlation Coefficient



You learned how to make an xy-scatterplot in Biology 105, and how to insert a trendline into that graph. Basically, in order to perform this analysis you graph two sets of data against each other on a trendline. Usually, you plot what you consider to be the independent variable on the x-axis, and the variable you expect to change with the independent variable on the y-axis. For instance, if you are trying to see if there is a relationship between sunlight and plant growth, you would plot the sunlight on the x-axis and the plant growth on the y-axis. It's only logical to assume that plant growth would be influenced by sunlight, but a stretch to imagine that the sun would be influenced by the growth of plants on Earth. So, plant growth **DEPENDS** on the sun and is thus the **DEPENDENT** variable and sunlight is **INDEPENDENT** of

plant growth. Of course, in some cases, like the one illustrated here, it is not at all clear which variable is independent and you just have to make a choice as to which you graph on the x-axis.

In the 105 lab, we used the trendline to calculate the slope of the line formed by a set of data. Often, this slope represents the rate of a reaction, such as photosynthesis or respiration. A trendline does more than this, however. It represents a line drawn through the data in such a way that it minimizes the distance to the line from any of the data points. It is the "best" line that can be drawn through the data. As such, it also gives us some information on **correlation**. Correlation is the relationship between 2 variables. Look at the figure below:

This is a hypothetical graph of the relationship between length and weight in an organism. Logically, we would expect a relationship between length and weight; it is logical to imagine that all things being equal, a longer animal will also weigh more. The graph seems to bear this out. The black line is a trendline; the computer inserted it. There are also two equations. The first:

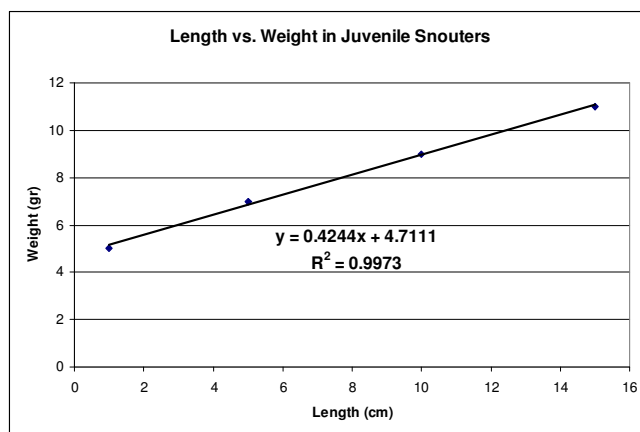
$$y = 0.4244x + 4.7111$$

is the equation of the trendline. It tells us that if you take the x-axis value (length, in this case), multiply it by 0.4244 and add 4.7111 you will get the y-axis value. Let's try it. Suppose we want to know how much a 16cm snouter would weigh. We take 16, multiply it by 0.4244, and add 4.7111 to the result.

$$y = 0.4244 * 16 + 4.7111 = 6.79 + 4.711 = 11.5 \text{ grams}$$

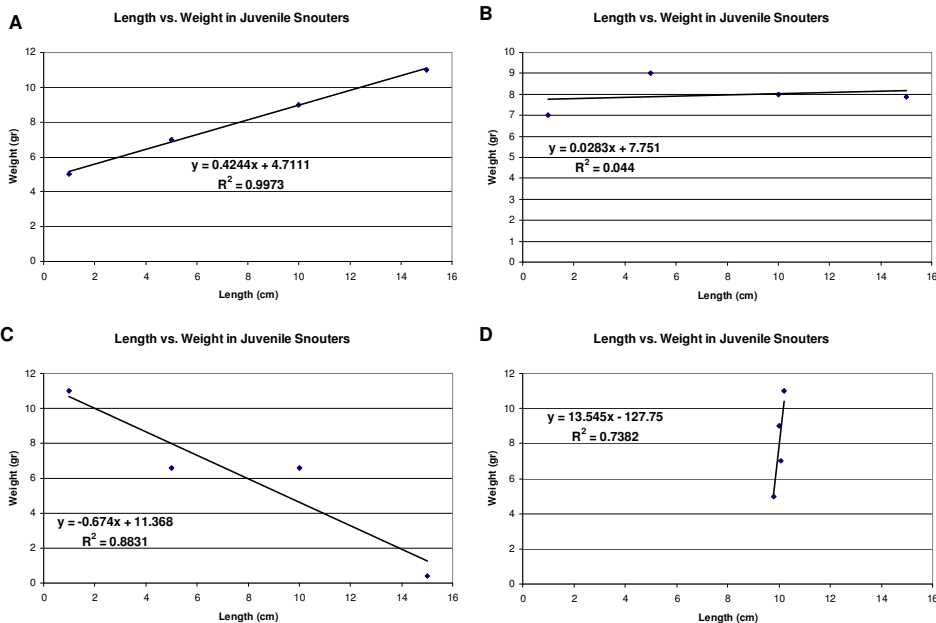
If you extend the trendline in the graph above, you will see that the calculated answer agrees with what you would read off the graph. In this case we have what we call a positive correlation; as the x-values increase so do the y-values.

What about the R^2 number? It's a little more complicated, but you can think of this as a measurement of how well the line fits the data. R^2 can range from 0 to 1; the closer it is to one, the better the line fits the data³. In the case above, it is a pretty good fit; you would expect this since all of the data points touch the trendline.



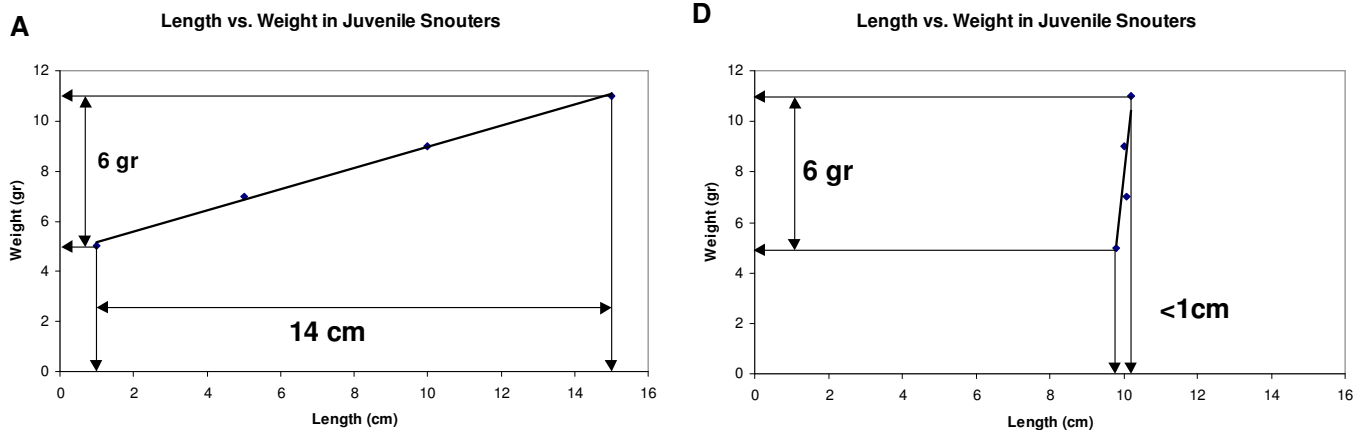
³ In actuality, statisticians prefer to use the R value (the square root of R^2) instead of R^2 to gauge the strength of correlation; like R^2 , R can vary from 0 to 1 and is interpreted the same way. If R^2 is = to 0.81, then $R = 0.9$ and indicates a strong correlation. Excel, however, provides the R^2 value instead of R.

Let's compare 4 different graphs:



Graph A is the graph we were just looking at. Graph B shows a situation with poor correlation. No matter what the length is, the weight remains relatively constant. Thus, knowing the length is of little use in predicting the weight - or vice-versa. In Graph C we have a negative correlation - as the length goes up, the weight comes down. Finally, in Graph D we see another positive correlation.

When looking at trendlines and their equations, there are 3 key things to examine - direction of slope, steepness of slope, and the R^2 value. If the line slants up to the right (Graphs A & D), then you have a **positive correlation**; if it slants down (Graph C), then you have a **negative correlation**. The steepness of the slope is a measurement of the relative strength of the effect. Graph D shows a relationship where a small increase in length leads to a greater increase in weight as compared to Graph A. The figure below shows this; in A it takes a 14 cm increase in length to reach a 6 gram increase in weight, while in D it takes less than a 1 cm increase in length to lead to a 6 gram increase in weight.



Don't confuse a greater **effect** (Graph D) with a greater **correlation** (Graph A), however. The R^2 value measures the predictive value of the correlation; Graph A allows you to be more accurate in your predictions as compared to Graph D.

Well, you have all the data you need to do the assignment (which starts on the next page). Good luck on the assignment! Please copy the rest of this file into a new document and use it as a template to type in your answers.

Transect Worksheet

Name _____

Anova: Single Factor - Tree Distance

SUMMARY

Groups	Count	Sum	Average	Variance
Rainforest 1	26.00	84.80	3.26	3.34
Rainforest 2	25.00	69.77	2.79	3.91
Tropical Dry Forest 1	17.00	42.77	2.52	1.83
Tropical Dry Forest 2	34.00	88.05	2.59	1.35

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	8.42	3.00	2.81	1.10	0.36	2.70
Within Groups	251.07	98.00	2.56			
Total	259.48	101.00				

1. Is there a significant difference between any of the transects in terms of tree density? Be sure to defend your answer, including any relevant statistical numbers. (Note: in lab we only did 2 transects, thus we did not generate enough data to do ANOVA on our own data – hence we are using the canned data).

t-Test: Two-Sample Assuming Unequal Variances

	Rainforest	Dry Forest
Mean	3.031	2.565
Variance	3.604	1.475
Observations	51.000	51.000
Hypothesized Mean Difference	0.000	
df	85.000	
t Stat	1.476	
P(T<=t) one-tail	0.072	
t Critical one-tail	1.663	
P(T<=t) two-tail	0.144	
t Critical two-tail	1.988	

2. The *t*-test above was run on data pooled from the rainforest and dry forest sites. Is there a significant difference between tree densities at the two sites? What is the chance that any difference you are seeing is due to chance?
3. Each person in a team should choose two of the variables (pH, Moisture, Light or Temp.) to analyze. Be sure that within the group that all 4 of these variables are analyzed by at least one person. Get the data from the other team, put the data into Excel, analyze the data and paste the data table and the statistical tables here to answer the following questions (for each variable): What is the mean and standard deviation of the variable in each transect? Is there a significant difference in the variable between the two transects?

Variable 1 _____
 [Insert data/analysis of variable 1 here]

Variable 2 _____
 [Insert data/analysis of variable 2 here]

4. Calculate the percent canopy cover for each sampling location on both of the transects. Paste in the data table and statistical table and determine if there is a significant difference in % canopy cover between the two transects.
5. Calculate the average distance of trees from the central point of each sampling site on both of the transects. Paste in the data table and statistical table and determine if there is a significant difference in average distance (tree density) between the two transects.
6. Calculate the average diameter of each tree on both of the transects. Paste in the data table and statistical table and determine if there is a significant difference in tree diameter.
7. Calculate the average circumference of each tree on both of the transects. Paste in the data table and statistical table and determine if there is a significant difference in tree circumference.
8. Prepare a xy scatterplot graph with tree distance on the x-axis and tree circumference on the y-axis. Use all of the data from both transects combined into one data set. Insert a trendline and the equation for the trendline. Paste in the graph, and add a paragraph answering the question: "Is there a significant relationship between tree density and tree circumference?" Your answer should address the nature and strength of the correlation.
9. Summarize your answers from questions 3-10. Is there any data to suggest a significant difference in tree density or size between the two transects? If so, what is the difference and why do you think we see the difference (or, why don't we see a difference)?
10. Lab report (due later): Write a short lab report with the following guidelines: A 1-2 page introduction with background information focused on factors that affect tree size. Include a hypothesis, purpose, etc. – just like in 105. You should have 2 hypotheses on whether or not each of two different physical parameters will have an effect on tree growth. Methods and materials should outline both the field work and subsequent data analysis. Results should analyze 2 of the physical parameters (your choice, try to find one that differs between the sites). The results should include statistical analysis that looks at whether or not the chosen parameters and the tree circumference is different between the transects. The results should also have graphical analysis to demonstrate correlation between the parameters and growth (or lack thereof). The discussion should tie this together with at least 2 references to the primary literature. Don't forget conclusions, future exp's, lit cited, etc. When in doubt, follow the bio 105 instructions. Typed, double-spaced, 1" margins, etc.