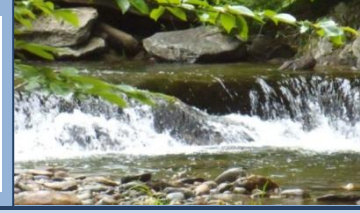
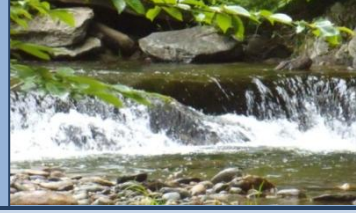


Module 5: Statistical Analysis

Module 5



Statistical Analysis

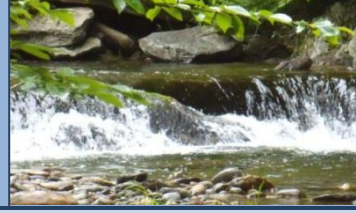
To answer more complex questions using your data, or in statistical terms, to test your hypothesis, you need to use more advanced statistical tests.

This module reviews the formulation of a central question, or hypothesis, and then describes three major categories of statistical tests:

1) **Questions and Hypotheses**
2) **Differences**
3) **Correlations**
4) **Regressions**

For each category, examples of the types of questions/hypothesis the test might help answer are given, along with directions on how to compute these statistical tests and create graphs and figures to illustrate your results.

Module 5



Statistical Analysis

1. Questions and Hypothesis

Central to any scientific research is a question that the research is trying to address. Scientific literature transforms this question into the form of a statement called a hypothesis which will be tested by your research.

Throughout this module we will use the term **"hypothesis"** to refer to your question that has been rephrased to make a statement. In statistics, a hypothesis is really composed of two hypotheses: a "**null hypothesis (H_o)**" and an "**alternative hypothesis (Ha).**" Take the following question as an example:

Are the levels of phosphorus recorded for my forested and urban sites different?

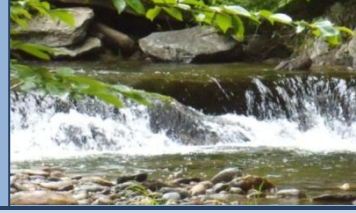
For this question we would write our hypothesis as the following:

H_o = *There is no difference between the levels of phosphorus at my forested site compared to my urban site.*

Ha = *There is a difference in the level of phosphorus at my forested site compared to my urban site.*

Continued…

Module 5



Statistical Analysis

1. Questions and Hypothesis

H_o = *There is no difference between the levels of phosphorus at my forested site compared to my urban site.*

Ha = *There is a difference in the level of phosphorus at my forested site compared to my urban site.*

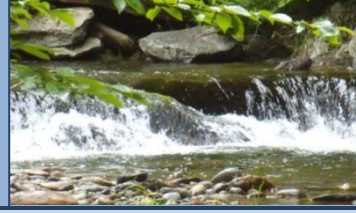
To test your hypothesis you will chose an appropriate statistical test which this module will walk you through. The results of this test will either be ***significant*** enough so that you will "reject your null hypothesis in support of your alternative hypothesis" or ***insignificant*** such that you "cannot reject your null hypothesis in favor of your alternative hypothesis."

Translated in terms of our example question that means:

Insignificant test result = *Your data does not provide enough evidence to show that there might be a difference between the two sites.*

Significant test result = *The results support the idea that there is a difference in the level of phosphorus between your two sites.*

Module 5



Statistical Analysis

2. Differences

Testing for differences allows us to statistically determine if the distributions, means or variances of multiple datasets are different.

Our example question about phosphorus is a question of differences:

Are the levels of phosphorus recorded for my forested and urban sites different?

And our hypotheses were as follows:

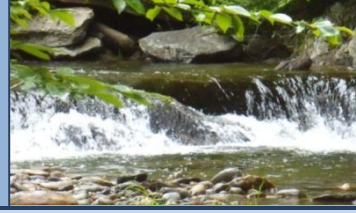
H_o = *There is no difference between the levels of phosphorus at my forested site compared to my urban site.*

Ha = *There is a difference in the level of phosphorus at my forested site compared to my urban site.*

The following statistical test can be used to test your hypothesis:

- **Two-sample t-test**

Module 5



Statistical Analysis

2. Differences

Two-sample t-test

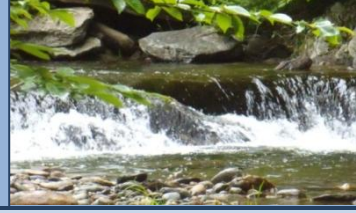
- *What it tests:*
  - The two-sample t-test is a statistical test that allows you to determine if the mean of two datasets are statistically different. It does this by using the mean and variance in a complex equation to produce a test statistic, known as "t." The value for this test statistic is compared to critical value of "t" which shows how likely the relationship between your two datasets is to occur under normal circumstances.

Click on the video icon to watch a video on how to use the t-test to calculate a P-value using Microsoft Excel

- *Interpreting the Output:*
  - The output that you will get from running a t-test in excel is the probability ("p-value") of getting the t-statistic calculated for your datasets. As a general rule, if your p-value is less than the critical value of .05 it means your results are significant and therefore support your alternative hypothesis which states that there is a difference in the distributions of your two datasets. The significance of the critical value of .05 is not explained in this tutorial, but we encourage you to explore further outside of what is offered here.

Continued…

Module 5



Statistical Analysis

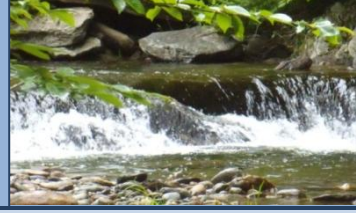
2. Differences

Two-sample t-test

- *Talking About Results*:
  - If you get a significant test statistic ($p<.05$), let's say for our question about the difference in phosphorus levels at your forested and urban sites, the results of your analysis support your alternative hypothesis that there is a difference in the phosphorus levels measured at these two sites.
  - If you get a significant test statistic that is >.05, you cannot reject your null hypothesis that there is no difference in the phosphorus levels measured at these two sites.
  - Your analysis can only say that there is or is not a statistically significant difference; this statistic *does not* explain what is causing the difference between the two datasets.
  - If you establish that there is a difference, you might look at other variables in your datasets such as land use or geology to help you speculate about what might potentially be causing these differences. Be sure to mention these ideas when you are describing your results!

Continued…

Module 5

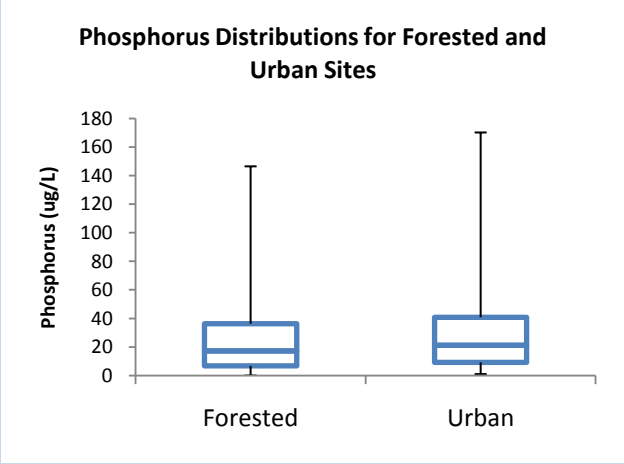


Statistical Analysis

2. Differences

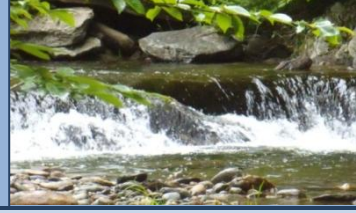
Two-sample t-test

- *Visualizing Results*:
  - A side-by-side box plot can be used to illustrate the results of your two sample t-test. First review how to create a single box plot in Module 4.
  - While a side-by-side box plot is used to compare the distribution of two datasets, it can also help you visually compare the central tendencies of multiple datasets as the middle line of the box represents the median value of the dataset which should be about equal to your mean.



Click on the video icon to watch a video on how to create a box-plot

Module 5



Statistical Analysis

3. Correlations

Testing for correlations allows us to statistically determine if there is a relationship between two variables in a dataset, and if so, the nature of the relationship (positive = they increase together or negative = one decreases while the other increases).

The following is an example of a question of correlation:

Is there a relationship between the level of E.coli in the water and water temperature?

And our hypotheses would be as follows:

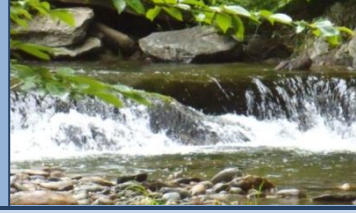
H_o = *There **IS NO** relationship between E.coli and water temperature measured at a stream site.*
Ha = *There **IS** a relationship between E.coli and water temperature at a stream site.*

To test for correlation, the following statistical test would be used:

- **Spearman's Rank Correlation**

Continued…

Module 5



Statistical Analysis

3. Correlations

Spearman's Rank Correlation

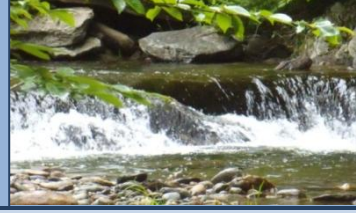
- *What it tests:*
  - Spearman's Rank Correlation is a statistical test that allows you to determine if there is a relationship between two variables in a dataset. It does this by using the mean in a complex equation to produce a correlation coefficient referred to as "R."

Click on the video icon to watch a video on how to calculate a correlation coefficient using Microsoft Excel

- *Interpreting the Output:*
  - The output that you will get from doing a correlation in excel is the correlation coefficient "R." The closer your correlation coefficient is to 1 or -1 the stronger the relationship between your two variables. If your correlation coefficient is negative than your two variables are inversely related (one increases as the other decreases). If your correlation coefficient is positive, then your two variables are positively correlated (they both increase together).

Continued…

Module 5



Statistical Analysis

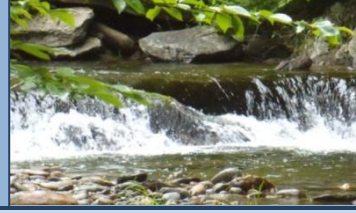
3. Correlations

Spearman's Rank Correlation

- *Talking About Results*:
  - If your correlation coefficient (R) is close to 1 or -1, let's say for our question about E.coli being related to water temperature, the results of your analysis support your hypothesis that there is a relationship between your two variables (E.coli and water temperature).
  - There is no critical threshold that says your correlation coefficient either is or isn't significant; we talk about the results as showing the strength of the relationship on this scale from 0 to 1 and 0 to -1.
  - Be careful: the correlation coefficient *does not* prove with 100% certainty that these two variables are related, and *does not* show cause in effect, though if you suspect that the value of one variables might be dependent on the value changes in the other you should read on about regression analysis!

Continued…

Module 5

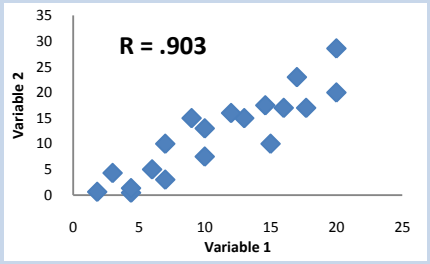


Statistical Analysis

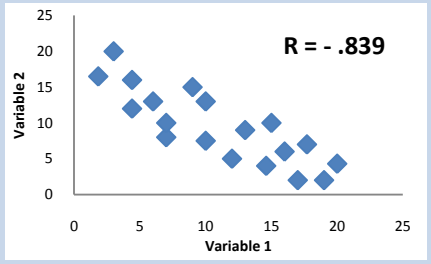
3. Correlations

Spearman's Rank Correlation

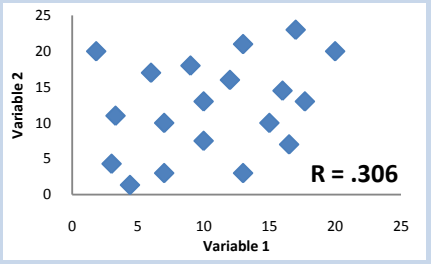
- Visualizing Results:
  - Correlations are best illustrated using a scatter plot
  - You might also use a scatter plot earlier on in your analysis when you are beginning to ask questions of correlations which you then might test using Spearman's Rank Correlation.
  - Scatter plots are made by plotting one variable against the other variable – the following three scatter plots illustrate the types of relationships you might see between two variables who may or may not be correlated:



Significant, positive correlation



Significant, negative correlation

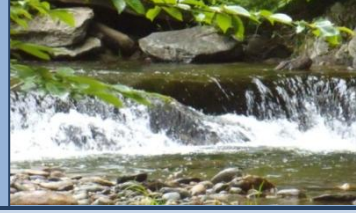


No significant correlation

  - Include your correlation coefficient (R) on the graph.

Click on the video icon to watch a video on how to create a scatter plot using Microsoft Excel

Module 5



Statistical Analysis

4. Regressions

A regression analysis is very similar to a test of correlation. The difference is that with a regression analysis we are looking to see if the values of one variable in our dataset, identified as the dependent variable (Y), increase or decrease as the values of another variable, identified as the independent variable (X), increase or decrease. If a change in ***X*** *does* cause a change ***Y***, the variables would be said to have a linear ***dependent*** relationship.

The following is an example of a question that can be answered through a regression analysis:

Does an increase in agricultural land use cause an increase in the amount of TSS in the water?

And our hypotheses would be as follows:

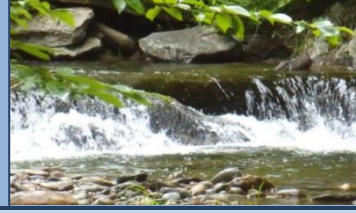
H_o = *The amount of TSS measured **DOES NOT DEPEND** on the amount of upstream agricultural land use.*
Ha = *The amount of TSS measured **DEPENDS** on the amount of upstream agricultural land use*

To test for a linear, dependent relationship the following statistical test would be used:

- **Regression Analysis: simple linear regression**

Continued…

Module 5



Statistical Analysis

4. Regressions

Regression Analysis: Simple Linear Regression

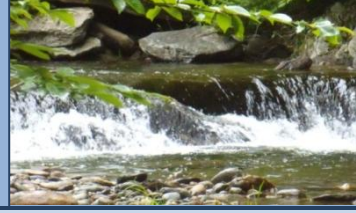
- *What it tests:*
  - A regression analysis is a statistical test that allows you to determine if there is a dependent relationship between two variables in a dataset. First you have to designate one variable as the dependent variable (Y), and the other as the independent (X). To do this, use common sense – would the amount of agricultural land use depend on the amount of TSS in the water? Or is it more likely that the amount of TSS depends on the amount of agricultural land use? Your variables are then organized into X-Y pairs. For example, at site B there is X-amount of agricultural land upstream, and the TSS reading at this site was Y, (etc. for all sites). The relationship between these two variables is represented by the linear equation $Y = aX + b$, and the strength of the relationship measured by the coefficient of determination "R^2."

Click on the video icon to watch a video on how to to calculate R^2 using Microsoft Excel

- *Interpreting the Output:*
  - The output that you will get from doing a regression analysis in excel is the coefficient of determination "R^2." The closer your R^2 value is to 1 the greater the dependent relationship between your two variables. If you read about correlations, R^2 is your R-value squared!

Continued…

Module 5



Statistical Analysis

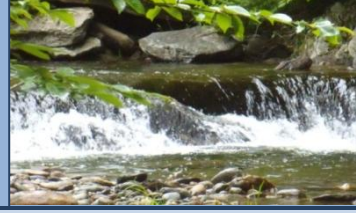
4. Regressions

Regression Analysis: Simple Linear Regression

- Interpreting Results:
  - The closer your R^2 value is to 1, the stronger the linear, dependent relationship between your two variables. This your Y variable being dependent on your X variable.
  - Looking at our question about agricultural land use and TSS, the closer to 1 your R^2 is, the more support there is for our alternative hypothesis that the amount of TSS measured depends on the amount of agricultural land use upstream.
  - Just as with a correlation, this relationship can be either positive or negative depending on the slope (b) of your linear equation: a negative sign means your Y variable decreases in response to an increase in your X variable, and a positive sign means your Y variable increases in response to an increase in your X variable.
  - Be careful: this analysis *does not* show cause in effect, but it does show dependence of one variable on another, and the nature of that dependence (positive of negative).

Continued…

Module 5

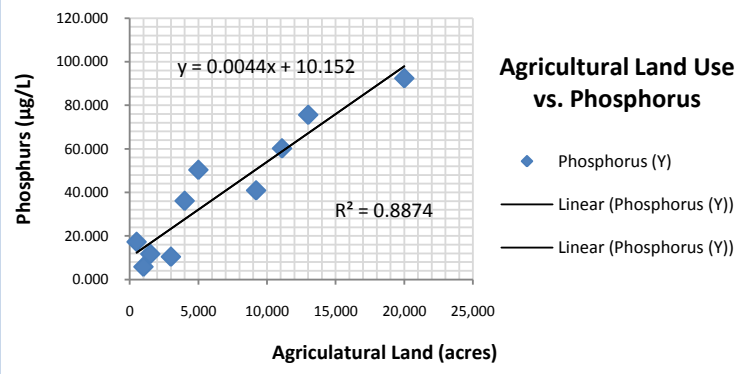


Statistical Analysis

4. Regressions

Regression Analysis: Simple Linear Regression

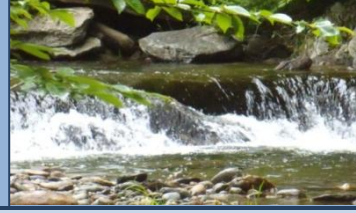
- Visualizing Results:
  - A scatter plot with a "best fit" line is used to illustrate the results of your regression analysis
  - These graphs are made by plotting the independent variable on the X-axis and the dependent variable on the Y-axis. The "best fit" line represents your linear equation y = aX + b.



  - Your equation gives you a line that represents a type of average describing the relationship between your two variables.
  - You should also add your R^2 value to the graph as well.

Click on the video icon to watch a video on how to create a graph of your regression analysis results using Microsoft Excel

Module 5



Statistical Analysis

SUMMARY

- The questions you are trying to answer should be phrased as a hypothesis
- If your hypothesis asks if two datasets are different, then you should use a Two-sample t-test to determine if your two datasets are statistically different
- If your hypothesis asks if two variables in a dataset are correlated, then you should use Spearman's Rank Correlation to determine the strength of the relationship between these two variables.
- If your hypothesis asks is one variable is dependent on another variable, then you should run a linear regression analysis to determine if your dependent variable (X) is dependent on your independent variable (Y).