A2: Statistics Refresher Lab

CONS 452

January 10th 2019

# Overview of Lab:

Forests can provide multiple ecosystem services that can affect the surrounding land cover. In particular many of these ecosystem services (e.g. microclimate regulation, pollination, nitrogen fixation etc.) can greatly benefit agricultural productivity. This difference in agricultural productivity can sometimes be noticed as one move away from forest edges.

In this lab you will (re)familiarize yourself with a few basic statistical concepts using data from a study area in Southern Ethiopia. Using Vegetation Indices (VIs) as a proxy for agricultural productivity, we measured several different VIs along various distances away from the forest edge. We ask the following questions:

1. Are the vegetation index results significantly different? If so, which ones?
2. Is there a relationship between agricultural productivity (measured via VI) and distance?

We will use data derived from three Vegetation Indices (VIs) that all range from -1 to 1:

* Normalized Difference Vegetation Index (NDVI)
* Soil Adjusted Vegetation Index (SAVI)
* Normalized Difference Red Edge Index (NDRE)

These VIs will be your dependent variables while “Distance” will be your independent variable. The independent variable is the variable that is controlled and unaffected by any other variable in the experiment. The dependent variable is the variable being measured and tested in the experiment.

# Learning Objectives:

* Be able to conduct a t-Test to determine significant difference between two populations
* Be able to conduct an ANOVA analysis to determine significant differences between two or more populations
* Be able to conduct a Regression analysis to determine the relationship between two variables.

# Deliverables:

In a single word document include:

* Short written answers to the two research questions above
* Result output for ANOVA (all three VIs)
* Result output for t-Tests (between all possible combination of VIs)
* Result output for Regression analysis (all three VIs)
* A short results statement for each of the statistical tests that you ran. For example: There was not a significant difference between NDVI and SAVI (F stat, F crit, P-value), therefore, we will accept the null hypothesis
* A scatterplot of each VI vs distance

## Step 0 - Turning on the Data Analysis Tab

Before we begin we first have to turn on the ‘Data’ tab. To do so: open Excel> open the File tab > Options > Customize Ribbon > and check the Data tab.

You should now see the Data tab along the top of the Excel window. Click on it and ensure you see the “Data Analysis” button.

## Step 1 – Comparing among VIs using ANOVA

Our first step is to determine if there is a statistical difference between the three VIs. One common way to determine if there is a significant difference between two or more groups is to use a single factor ANOVA (Analysis of variance). An ANOVA is a type of hypothesis test, in this case the null hypothesis means that the means of several populations are all equal:

H0: μ1 = μ2 = μ3

H1: at least one of the means is different

The ANOVA results will tell us whether we should accept (no statistical difference between means) or reject (at least one of the means is different) the null hypothesis.

Go to Data > Data Analysis > ANOVA: Single Factor

Select the dataset for all the dependent variables (NDVI, SAVI, and NDRE) as the input range. Select an output option and click OK.

\*Note alpha is set to 0.05. This is the default significance value\*

To interpret this result we must look at the **F** and **F crit** columns. If **F** > **F crit** then we must reject the null hypothesis. If **F** < **F crit** then we must accept the null hypothesis. Another important output is the **p-value**. A small p-value that is less than the significance value (alpha) means that there is strong evidence against the null hypothesis while a large p-value indicates that the evidence is not strong enough to reject the null hypothesis.

While an ANOVA is very useful in determining whether or not there is a significant difference among several populations, it does not tell us where the difference is. There are many common ways to test this. We will use the Student’s t-Test.

## Step 2 – Comparing among VIs using t-Tests

The t-Test is a common method used to determine if there is a significant difference between populations. However, unlike the ANOVA the t-test can only determine the difference between two populations. The t-Test is also a type of hypothesis test. Once again our null hypothesis is that the means of the two populations are equal.

H0: μ1 - μ2 = 0

H1: μ1 - μ2 ≠ 0

The t-Test results will tell us whether we should accept (no statistical difference between means) or reject (the means are significantly different) the null hypothesis.

Go to Data > Data Analysis > t-Test: Paired Two Sample for Means

Select the dataset of NDVI for variable 1 range and the dataset of SAVI for variable 2 range. Select an output option and click OK.

To interpret this we must look at the t Stat and t Critical values. Since the VIs we are comparing can be either larger or smaller than each other we are conducting a two-tailed test. Therefore, we are only interested in the two-tail part of the results. If **t Stat** < **-t Critical** **OR** if **t Stat** > **t Critical** then we must reject the null hypothesis. In other words, if the absolute value of t stat is greater than T critical **(|t Stat|** > **T Critical**), then the null hypothesis is rejected.

In addition to comparing t stat vs t critical, we can also look at the **p-value**. Once again, a p-value of less than alpha (default 0.05) indicates a rejection of the null hypothesis while a p-value of greater than alpha indicates an acceptance of the null hypothesis.

Repeat the t-Test to test for differences between all the VIs.

## Step 3 – Linear Regression

**Now that we have determined whether or not the dependent variables (VIs) are different we will now look at the relationship between the independent (distance) and dependent variable (NDVI).

Go to Data > Data Analysis > Regression

Select the dependent variable (VI) as your Y range and independent variable (distance) as your X range. Select an output option and click OK.

For the purpose of this lab we will focus on interpreting the top part of the results from the regression analysis. **Multiple R** is the *correlation coefficient* which tells you how strong a relationship is, with 1 being a perfectly positive relationship and 0 being no relationship at all. **R squared (R2)** is the *coefficient of determination* which tells you how many points fall on the regression line. In simpler terms, multiple R is the correlation between actual and predicted values for the dependent variable while R squared is the regression model’s accuracy in explaining the dependent variable. For more information click this [link](http://blog.uwgb.edu/bansalg/statistics-data-analytics/linear-regression/what-is-the-difference-between-coefficient-of-determination-and-coefficient-of-correlation/).

The p-value for the regression can be found under **Significance F** in the ANOVA section of the result output. The p-value tests the null-hypothesis that there is no effect (R2=0) between your variables. A low p-value (< 0.05) means that the null hypothesis should be rejected.

Repeat the regression analysis for all three VIs.