Analyzing Variation

By Eric Pauley and Chris Hill

A major task faced by ecologists is to detect differences among groups of organisms (and ultimately explain those differences). Most of the time, we are more concerned with the broad differences among groups, rather than the differences among individuals within a group. While each organism is unique, we usually want to distinguish between variations that are "**random**" (i.e., lacking in consistent explanation) and variations that are "**significant**" (i.e., greater than random). Most statistical techniques are designed to make this distinction, and thus ecologists have to be part-time statisticians.

Sampling

Since ecologists can't usually measure all individuals in a group of interest (for example, all the leaves in a forest), they measure a **sample,** a subset of the whole. It is important to pick the sample in such a way that the characteristics of the sample accurately reflect the characteristics of the whole, usually called the "**population**." It turns out that the most reliable way to pick an unbiased sample is to pick a **random** sample.

The "Null" Hypothesis

Most statistics are based on a simple premise: *It is easier to disprove a statement than to prove it*. To disprove a statement, all you need is some convincing evidence to the contrary. Moreover, ecologists tend to be conservative, in that we don't want to say there is "significant" variation unless faced with clear evidence of it. Thus, before conducting any kind of test, we usually devise a formal statement called a **"null" hypothesis**, which we assume to be true unless we find compelling evidence against it. The null hypothesis is usually what is being tested. If the null hypothesis is rejected (i.e., disproved), then you can conclude that your results are statistically significant.

You will collect needles from different species of pine trees on campus to determine whether needle length varies significantly among them. The species are:

1. Longleaf pine (*Pinus palustris*)
2. Loblolly pine (*Pinus taeda*)
3. Pond pine (*Pinus serotina*)

In this lab, we will examine whether needles vary significantly in length among these species. On the last page, you'll need to state the null hypotheses for today's lab.

PART I - Comparing Two Groups

The Data

Work in groups of 2. Using the rulers, measure the length of pine needles from (1) longleaf pine, (2) loblolly pine, and (3) pond pine. Note that the needles occur in clusters called "**fascicles**." Measure the longest needle in each fascicle from the base of the fascicle to the tip. If all needles in a fascicle are damaged, discard that fascicle. Measure each needle to the nearest 0.1 centimeter (cm), recording your data on the sheet provided. Later, you will enter these data in MS Excel for the analysis of variance statistical test.

The "t" Test

One technique used to determine if a significant difference exists between two groups is the "t" test. It compares their **means** (average values), with reference to how variable each sample is. Of course, two means will rarely be identical. Whether the difference between means is "significant" depends on three factors:

- **How big is the difference between means?**
- **How much variation around each mean is there?** High variation can mask a small difference.
- **How many samples is each mean based on?** The fewer needles in each sample, the less precise your estimates of the mean and variation will be. Thus, it will be difficult to detect a real difference, if one exists. Larger samples enable you to be more certain about the mean and the degree of variation within each sample.

To do a t-test, you'll need answers to all three questions.

Now it is time for some calculations. Calculate and enter the following values on the Data Sheet:

1. For *each* of the "Longleaf pine" and "Loblolly pine" samples, calculate the **sum** of all raw values (ΣX_i).
2. Calculate the **mean** (Y) for *each* sample: the sum (ΣX_i) ÷ the number of needles measured (n)
3. Go back, square each value for a needle, and calculate the **sum** for these squares for *each* sample: ΣX_i^2
4. Calculate the "**sum of the squares of the deviations**" (SS) for *each* sample:

Answer sheet columns→ *Right* *Left*

$$SS = \Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}$$

5. For *each* longleaf and loblolly pine sample, calculate the **variance**: $s^2 = SS \div (n - 1)$

Calculating "t"

Calculate the value of "t" accordingly:

$$t = \frac{|Y_1 - Y_2|}{\sqrt{\left(\frac{{s_1}^2}{n_1}\right) + \left(\frac{{s_2}^2}{n_2}\right)}}$$

Deciding If "t" Represents a Significant Difference Between Groups

You are now ready to compare your *calculated* value of "t" with that of a *theoretical* value. This theoretical value is the maximum value expected for "t" if two means are the same. If your calculated value exceeds the theoretical value, then you have evidence to "reject" the null hypothesis that there is no difference.

Look at the table of "Critical Values of Student's t-Distribution". Each row consists of values for ν (Greek: nu), the "**degrees of freedom**." The meaning of this number is a little obscure, but in essence *each mean* has a "degree of freedom" of 1 less than the total number of values included in it:

$$\nu = n - 1$$

However, we are comparing two means in the "t"-test, so the degrees of freedom in the whole test is thus:

$$\nu = n_1 + n_2 - 2$$

Degrees of freedom for the t-test

Now look at the column headers for the Critical Values table. These numbers are values of α (Greek: alpha). They represent the probability of making a "**Type I error**," or rejecting the null hypothesis when in fact it is true. You wouldn't want to say there's a difference between two groups when in fact there is no difference, would you?

The **α-level** is merely a formal statement of how willing you are to incorrectly reject a true null hypothesis. Most people settle on the **0.05** level (although there are often good reasons to go lower or higher).

On the last page, record your results, along with your conclusion about whether to reject the null hypothesis.

PART II - Comparing Three (or More) Groups

A logical extension of the t-test involves comparing 3 or more groups. The simplest such case is called "**analysis of variance**," or "**ANOVA**." The t-test, ANOVA, and many other techniques are based on distinguishing among-group variation from within-group variation. However, the calculations quickly get very complex. Enter the computer.

Analyzing the Data with Microsoft Excel

Excel can do ANOVA of several types. Our example is called a "**single-factor**" ANOVA. The "factor" here is the species. We are testing the effect of species on needle length, and the completed ANOVA allows us to decide if the differences among needles from different species are statistically significant (and thus of possible ecological importance).

1. **Enter the data.** Open Microsoft Excel. To do an ANOVA, Excel expects the data for each group to be arranged in side-by-side columns, so enter your needle length data for each species like this:

Longleaf pine	Loblolly pine	Pond pine
value	value	value
value	value	value
value	value	value
. . .	. . .	. . .

2. **Do your ANOVA.** Follow these directions:

   a. Go to the "**Data**" menu and select "**Data Analysis**". In the dialog box that appears, select "**Anova: Single Factor**" and click OK. In the "Anova: Single Factor" dialog box, our **Input Range** will be all the cells that contain data, *including the column headers* (Longleaf pine, Loblolly pine, etc.). Click in this box, and then without closing this window, use the mouse to select all cells containing data, including the column headers. Excel will put the correct cell references in the Input Range space.

   b. Our data are "**Grouped By**" columns.

   c. Check the "**Labels in First Row**" box so that Excel knows to use the column headers as labels. If the value of "**Alpha**" is *not* 0.05, type that in.

   d. Under "**Output Options**," click "**Output Range**" then click in the box next to it so that your cursor is there. Now click somewhere to the right of the data on the spreadsheet. This is where Excel is going to put a whole bunch of stuff, so you need plenty of room.

   e. Click OK. The computer will calculate the ANOVA, and two tables should appear. The first table gives you a summary of the sums, means, and variances for each of your groups (species). The second table is a standard ANOVA table, which looks like this:

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	xxx	xxx	xxx	xxx	xxx	xxx
Within Groups	xxx	xxx	xxx			
Total	xxx	xxx				

A little explanation is in order:

"***Groups***" are different species, in this case.
"***SS***" is the sum of the squares of each value for each source of variation.
"***df***" is the degrees of freedom for each source of variation (Between = n - 1, Within = n - 3).
"***MS***" is the mean square for each source of variation, computed as: SS ÷ df
"***F***" is the calculated F-statistic, computed as: $MS_{Between} \div MS_{Within}$
"***P-value***" is the probability of obtaining a calculated statistic bigger than the "*F crit*" by chance alone.
"***F crit***" is the theoretical "critical" value of F.

Your conclusion to determine statistical significance is based on whether (the calculated) *F* is larger than *F crit*. It is also statistically significant if the *P-value* is smaller than your pre-chosen level of α (again, usually 0.05).

Complete both tables and answer questions on the last page.

Data Sheet for the "t" Test Lengths (cm) of needles from different species of pine. *Write down the length of pond pine needles in the table at the **bottom** of the page*.

	Longleaf pine			**Loblolly pine**	
	Raw	*Squared*		*Raw*	*Squared*
Sums	$\Sigma X_i =$	$\Sigma X_i^2 =$	Sums	$\Sigma X_i =$	$\Sigma X_i^2 =$
n_1			n_2		
mean (Y_1)			mean (Y_2)		
SS_1			SS_2		
variance (s^2_1)			variance (s^2_2)		

Calculated "t"	
α-level	
df	
Critical "t"	

From critical value table→

Enter the pond pine needle lengths (cm) in the table below. You will ***not*** be statistically analyzing these data by hand.

BIOL 370L, Analyzing Variation **Names:** _____

1. Null hypothesis for **t-test**:

2. Record the results of your t-test here:

	Longleaf pine	**Loblolly pine**
n		
mean (Y)		
variance (s^2)		

Calculated "t"	
α-level	
Degrees of freedom	
Critical "t" (from critical values table)	
Should you reject the null hypothesis? Why?	

---

3. Null hypothesis for **ANOVA**:

4. Record the results of your ANOVA here by copying the Excel output.

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Longleaf pine				
Loblolly pine				
Pond pine				

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups						
Within Groups						
Total						

5. Should you reject the null hypothesis? Why?