

STATISTICAL ANALYSIS OF ECOLOGICAL DATA

I. Objectives:

1. Discuss why scientists employ statistics to understand ecological problems.
2. Calculate the descriptive statistics including the mean, variance, and standard deviation of describe a population.
3. Describe the concept of normal distribution.
4. Understand and apply the t-test to differences in populations.
5. Understand and apply the chi-square test to frequency or count data.

II. Introduction:

Ecologists are often concerned with numbers of organisms (density) and their patterns of distribution in nature. This makes ecology a quantitative science. However, ecologists cannot count and determine the location of every organism in a given area. Rather, ecologists must collect and analyze data from samples taken within the population. The quantitative data collected by ecologists can be (actually, must be) analyzed using statistics.

The term statistics is commonly used to describe two distinct concepts. Statistics is the science that deals with summarizing and analyzing data. Within the science of statistics, estimates of true values of parameters like the true mean or true standard deviation are also known as statistics. In this sense, a statistic like the sample mean is an estimate of a value that actually exists in nature.

A. Terminology:

Before we begin our work, some definitions are needed:

x_i = single observation or measurement
 i = indicates the specific observation i in a set

n = sample size (number of observations)

$\sum_{i=1}^n x_i$ = sum of $i = 1, 2, \dots, n$ observations

\bar{X} = mean; "X-bar"

s^2 = variance

s = standard deviation

SS = sum of squares

df = degrees of freedom, often = $n-1$

B. Descriptive Statistics:

Descriptive statistics summarize some aspect of the population. The most commonly used are the mean, median, mode, variance and standard deviation. For ecological studies, the mean, variance and standard deviation are most often used.

B1. Mean:

The mean is a measure of the central tendency for a population. This is also known as the average.

$$\text{Mean} = \bar{X} = \sum_{i=1}^n x_i$$

Example 1: In population #1, the following numbers of trees are counted in 5 quadrats: 1, 6, 11, 16, 21

$$\text{Mean} = 55/5 = 11$$

Example 2: In population #2, the following numbers of trees are counted in 5 quadrats: 10, 11, 11, 11, 12

$$\text{Mean} = 55/5 = 11$$

B2. Variance:

The two populations shown above have the same mean though the variation in the counts is quite different. Population #2 has a narrow range of abundances per quadrat (10-12) while population #1 has a relatively wide range of numbers per quadrat (1-21). One way to measure this range of possible results is to calculate the variance.

$$S^2 = \text{SS}/\text{df}$$

SS can be calculated as $\sum_{i=1}^n (x_i - \bar{X})^2$, but this is a cumbersome equation to use when there are large numbers of data. A simpler way of calculating SS on most calculators is:

$$\text{SS} = \sum_{i=1}^n x_i^2 - \left[\left(\sum_{i=1}^n x_i \right)^2 / n \right]$$

Example 1: For population #1:

$$\text{SS} = 855 - [3025/5] = 855 - 605 = 250$$

$$S^2 = 250/(5-1) = 62.5$$

Example 2: For population #2:

$$\text{SS} = 607 - 605 = 2$$

$$S^2 = 2/(5-1) = 0.5$$

B3. Standard Deviation:

The standard deviation S is calculated as the square root of the variance. For population #1, the standard deviation is 7.9, while the standard deviation of population #2 is 0.7. The standard deviation is important because it provides an easily visualized measure of the variation from the mean for normally distributed data.

What does normally distributed mean? A normal distribution is a typical bell curve, with the peak of the curve corresponding to the mean. However, a bell curve can be narrow and tall or broad and short, depending on whether the data has a low or high variance. The standard deviation provides an easily understandable estimate of this variability. For normally distributed data, 95% of all possible observations (such as counts in quadrats) will lie within 2 standard deviations of the mean. These values are known as the 95% confidence limits. For example, population #2 has a standard deviation of 0.7. Two standard deviations would thus be 1.4. Therefore, the 95% confidence limits for this

population are 11 (the mean) ± 1.4 (or 9.6-12.4). This means that if we took further quadrat samples from this population, on average 95% of these additional quadrats would have densities between 9.6-12.4 trees per quadrat.

C. Comparative Statistics

C1. What are comparative statistics?

Statistics can also be used to determine whether populations (or measurements of population characteristics) are similar or different. For example:

Is the density of pine trees in two areas similar or different?

Is the number of crabs in the Cape Fear estuary more now than a decade ago?

This use of statistics is called significance testing. Using the scientific method, even using statistics, the scientist cannot prove anything. Statistics can only demonstrate that an event is very unlikely, but nothing is ever proved in the process. Typically, the investigator establishes a hypothesis generally suggesting a difference or pattern of some kind, and then tries to determine if that hypothesis is likely by showing that the alternative possibility, that there is no difference or pattern (called the null hypothesis) is not likely. For example, one may have a hypothesis that densities of pine trees are different between 2 forests. However, because statistics do not prove differences, the investigator actually seeks to show that the null hypothesis of no difference between the forests is unlikely to be true (confusing isn't it!).

C2. Example using the t-test.

Let's run through an example:

Step 1: A researcher develops the following hypothesis:

H_a = There is a difference in the density of pine trees between a recently burned forest and a forest that has not been burned for 25 years.

Step 2: Form a null hypothesis:

H_o (the null hypothesis) = There is no difference in pine tree densities between the two forests.

Step 3: Data collection:

For our scenario, the researcher uses quadrat sampling to obtain the following counts of pine trees per 100 m² quadrats:

Unburned forest (4 quadrats): 5, 2, 3, 8 Burned forest (5 quadrats): 15, 25, 20, 11, 15

Step 4: Data Analysis:

Now the densities (no. per quadrat) of pines can be compared statistically to determine if there is a difference. Since this data represents replicate measures from 2 groups, an appropriate test for comparing the groups is the t-test.

Calculation by Hand:

The t-test has the following formula: $t = \frac{|\bar{X}_1 - \bar{X}_2|}{S_{x1-x2}}$

Where

$$S_{x1-x2} = \sqrt{(S_p^2/n_1) - (S_p^2/n_2)} \quad \text{and} \quad S_p^2 = (SS_1 + SS_2)/(df_1 + df_2)$$

This calculation assumes the variances are pooled.

The following numbers are calculated to determine the t-statistic for the two populations (b=burned, u=unburned)

$\text{mean}_u = 18/4 = 4.5$	$\text{mean}_b = 86/5 = 17.2$
$n_u = 4$	$n_b = 5$
$SS_u = 102 - (324/4) = 21$	$SS_b = 1596 - (7396/5) = 116.8$

$$S_p^2 = (21 + 116.8)/(3 + 4) = 19.7$$

$$S_{x1-x2} = \text{sqrt}[(19.7/4) + (19.7/5)] = \text{sqrt}(8.97) = 2.99$$

$$t = \frac{|4.5 - 17.2|}{2.99} = 4.26$$

The degrees of freedom (df) for this test are $(n_u - 1) + (n_b - 1) = (4 - 1) + (5 - 1) = 7$
This t-value can be looked up in a t-table. If the calculated value is greater than the value under row df for 0.05 probability level, then you reject the null hypothesis and conclude there is a significant difference between the burned and unburned forests. In this case, the table value for 7 df and 0.05 significance level is 1.895. Therefore, we can conclude that pine tree density is greater in the burned forest.

For clarity, ecologists tend to reserve using the word *significant* for when they are referring to statistical significance. Please follow this convention in this lab and when writing your lab reports.

C3. Calculation using JMP IN:

In this course, we can also calculate a t-test using a standard, commercial statistical package, JMP IN. To do this, do the following steps:

1. Double click on the JMP icon
2. Choose "New Data Table" from the start menu that appears. On the new data table created, there is only 1 column (labeled column 1), you will need to create a second by double clicking in the right side space (to the right of "column 1").
3. Click on the column 1 square and then click on the name. Type in "forest type". Do the same for column 2, typing in "density". By double clicking on the column name you can specify more data attributes including data type and model type (for the data). For forest type, change the data type to "Character" and set the model type to

“Nominal” because the data you are entering is categorical. You can leave density as the default data type “Numeric” and model type “Continuous”.

4. Enter the data in the following format (u=unburned forest, b=burned forest):

	Forest type	Density
1	U	5
2	U	2
3	U	3
4	U	8
5	B	15
6	B	25
7	B	20
8	B	11
9	B	15

Note: if the program still doesn’t allow you to enter anything in the column for forest type, you need to highlight the top of the column, then select Cols – Column Info from the menu bar. A box will appear and you need to change Data Type to Character.

1. From the menu bar, choose **Analyze – Fit Y by X**
2. Choose forest type as **X** and density as **Y**
3. Do the group means/one-way ANOVA comparison (which will be the default comparison for you data).
4. You will get a graph of the data. Choose **means/ANOVA/pooled t-test**.
5. The results of the t-test will be displayed along with the results of several other tests. Please note that the calculated t-value is the same as we calculated by hand, and a p-value of less than 0.05 is shown, indicating a significant difference.

C4. Chi Square (χ^2) test

Another test that is useful for comparing totals, counts or frequencies is the χ^2 test. Using the χ^2 test, scientists can determine if observed values are the same as values expected for a given situation.

For example, imagine that you have surveyed the abundance of crabs under “large” rocks and “small” rocks in a swift current to determine if there is a difference in the number of crabs under each rock type.

The total number of crabs under 20 rocks was:

	Large rocks	Small rocks
Observed	200	10
Expected	105	105

The expected number is established by determining the number of crabs expected if the null hypothesis were true. In this case there is a hypothesis (H_a) of a difference between

the rocks and a null hypothesis (H_0) of no difference. So, if there are a total of 210 crabs collected, with no difference in the number found under each rock type, there must be an expected number of 105 for both large and small rocks ($105+105 = 210$).

The χ^2 statistic is then calculated by:

$$\chi^2 = \sum \left[\frac{(\text{observed} - \text{expected})^2}{\text{expected}} \right]$$

In this example, $\chi^2 = (200-105)^2/105 + (10-105)^2/105 = 171.9$

For this case, the degrees of freedom (df) for the test is determined by the number of groups minus 1 ($2-1=1$). For 1 degree of freedom at a 0.05 significance level, the critical table value is 3.84. Since your calculated value is greater than the table value, the null hypothesis is rejected and you conclude there is a difference in the number of crabs under large rocks versus small rocks.

NOTE: In a Chi Square test, if $df=1$, you must also use the **Yate's correction**. This correction helps avoid a Type I error (rejecting the null when it's true) and helps make the test more rigorous with small samples. To complete the Yate's correction, use the formula:

$$\chi^2 = \sum \left[\frac{(|\text{observed} - \text{expected}| - 0.5)^2}{\text{expected}} \right]$$

The vertical line brackets around observed - expected refer to the absolute value (ignore any negative values) in the calculations.

III. Questions to Consider

Please turn in your answers to the following questions at the end of the laboratory:

- (1) For your forest data, what are your null and alternative hypotheses, and how will you decide to reject/accept them? What decision-making went into your experimental design to allow you to be confident in the way you chose to test these hypotheses?
- (2) Write a brief summary you could use to explain to someone with no connection to this lab why and how we were exploring scientific and statistical differences in populations in these forests labs and what your main conclusions are based on your results.
- (3) If your job was to report to a land management advising council about how best to manage this forest, what would you recommend to them based on your results? Why?

IV. Acknowledgments

The faculty and graduate teaching assistants at UNCW compiled this laboratory.