Chapter 4

Introduction to Scientific Thinking and Statistics in Ecology

Adapted from:

Taylor, J. 2008. Appendix A: A Primary of Biometry. Ecology Laboratory Manual. University of New Hampshire.

4.1 Introduction to Statistics

To do science is to describe and explain nature through observation. In ecology, most of these observations are quantitative. They are either measurements (e.g. temperature, pH, mass) or counts (e.g. number of individuals in a population). Unfortunately, quantitative phenomena are highly variable, as two observations are rarely identical. We even use “variable” as a synonym for the ecological attirbutes we measure. Therefore, in order to describe nature, we must find ways to describe and expalin the variation we observe. Similarly, we only sample nature so we need to use statistics to test hypotheses if we want our inferences to be relevant to populations rather than just to the specific samples we took.

4.2 Describing Central Tendancy and Variation

The first step in describing observations or samples is to summarize them by reducing many individual observations to one or two statistics that describe the original observations without listing all of them. The simplest **descriptive statistics** is the **mean** (sometimes call the average). The mean is defined as

$$\bar{X} = \frac{\Sigma X_i}{n}$$

where ΣX_i is the sum of all observations and n is the total number of observations.

The mean can reduce many observations into a single number for descriptive or comparative purposes. However, the mean does not describe the amount of **variation** among observations. It is possible for two different groups of observations to have the same mean, but different amounts of variation. Consider Figure 4.1, which shows two normal curves with the same mean(5), but with different variances.

Note that in the sample with less variation (the taller, narrower one), most of the observationsare close to the mean, while in the sample with more variation, there are more observations at greater distances from the mean. This provideds a clue about ways to measure variation: Find a statistics that summarized distances

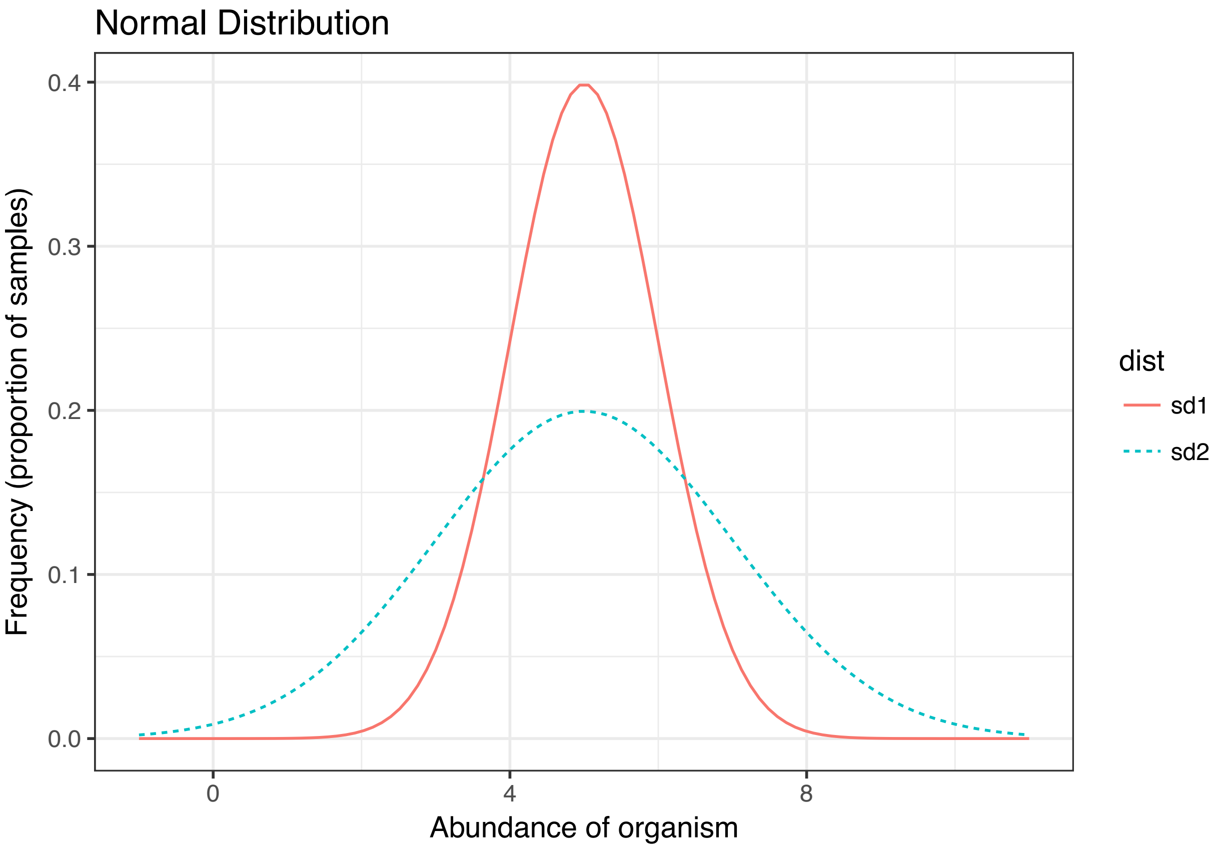


Figure 4.1: Two curves with equal means (5), but unequal variances. The higher, narrower curve has a variance of 1, while the broader, flatter curve has a variance of 2

from the mean to all the observations. Distance from the mean is simply $x_i - \bar{X}$, where x_i represents any one observation and $barX$ represents the mean of all the observations. Distance from the mean tells us how deviant a single observation is. To summarize all thedistances into an estimate of variability, we calculate the **variance**, which is the average squared distance from the mean. To calculate variance, subtract the mean from each observation:

$$x_i - \bar{X}$$

Square each deviation (This removes negative distances):

$$(x_i - \bar{X})^2$$

Sum the squared deviations:

$$\sum (x_i - \bar{X})^2$$

Find the average of the squared deviations. When taking an average we typically divide by n, the number of observations. In this case, however, we calculate the average by dividing by $n-1$ instead of n. This corrects for a bias caused by sampling only a few of all possible observations. If we observed all members, then it is apprpriate to divide by n.

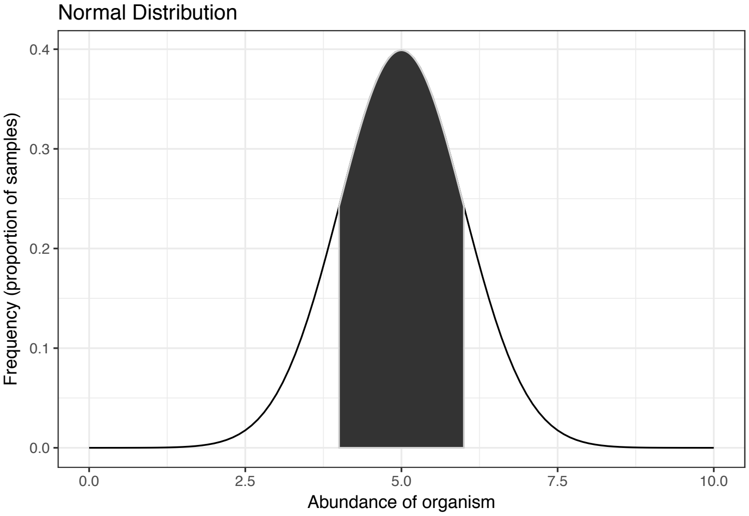
$$\frac{\sum (x_i - \bar{X})^2}{n-1}$$

The average of all the squared deviations is also known as the variance, and it is traditionally reresented by σ^2. The variance is a very important descriptor of ecological variation: The greater the variation among the observations, the greater the variance. Knowing the amount of variation in ecological attributes can tell us much about ecological patterns and processes.

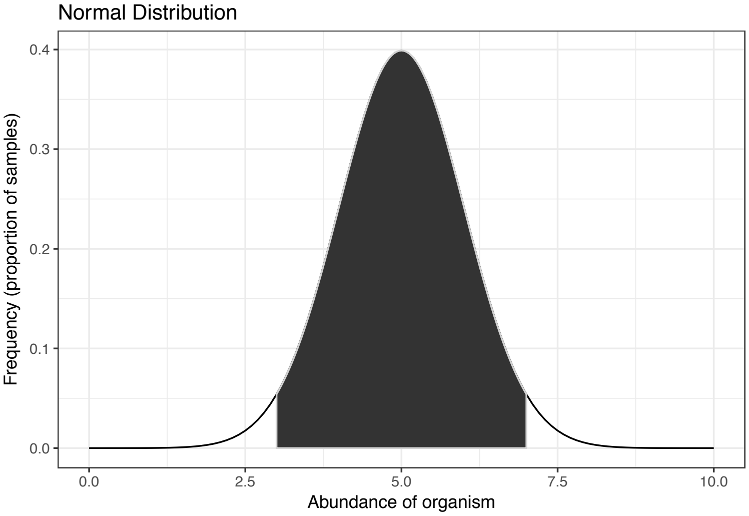
Another descriptor of variation is the **standard deviation**, abbreviated as σ or SD, which is the positive square-root of the variance:

$$\sigma = \sqrt{\sigma^2}$$

The standard deviation may be thought of as the average absolute deviation from the mean. It has the same units as the original observations. If our data are normally distributed, such as in Figure 4.1, then about 68% of the observations will fall within 1 standard deviation on either side of the mean (Figure 4.2), and about 95% of the observations will fall within 2 standard deviations of the mean (Figure 4.2).



\begin{figure} \caption{Normal curve with colored area indicating occurance of +/-1 standard deviation. Approximately 68% of all observations fall in this area.} \end{figure}



\begin{figure} \caption{Normal curve with colored area indicating occurance of +/-2 standard deviation. Approximately 95% of all observations fall in this area.} \end{figure}

4.3 Explaining Variation: Measuring association with the Coefficient of Determination

To explain things, we try to find other things that are associated with them that change as they change and remain constant when they stay constant. The search for pattern in nature is thus often a search for associations. Although cause and effect can rarely be inferred from such associations, the do have predictive use. If you variables are associated with each other, then knowledge of one tells us something about the other.

A good descriptor of the association between two variables is the **Coefficient of Determination**, R^2. It is the ratio of the variance shared with another variable to the total variance in the two variables.

$$R^2 = \frac{Variance\ in\ Common}{Total\ Variance}$$

Thus, R^2 is the proportion of the variance in one variable explained by, or associated with, another variable. As a proportion, it is a unitless number ranging from 0 to 1. High values, those approaching 1, indicate very close association, because one variable explains a large proprotion of the variance in another variable. Low values, those approaching 0, indicate little or no association. The equation for calculating the Coefficient of Determination is

$$R^2 = \frac{\left(\frac{\sum(x_i-\bar{X})(y_i-\bar{Y})}{n-1}\right)^2}{\sigma_x^2\sigma_y^2}$$

where X and Y are the data for the two variables and σ_x^2 and σ_y^2 are the variances of the two variables. The part within the parentheses is the **covariance**.

R^2 can also be used to determine which of several hyptheses best explains an association. The hypothesis that generates the largest R^2 between two variables is the best explanation of their association. There are other, better methods ecologists generally use for model comparison, but they are much more complicated and R^2 will be useful for our purposes.

The **Correlation Coefficient**, r, is another measure of association. It varies from 1 (perfect association where the two variables increase or decrease together), to 0 (no association), to -1 (perfect inverse association, where one variable increasesas the other decreases). It is caluclated as the square root of the Coefficient of Determination:

$$r = \sqrt{R^2}$$

Computers easily calculate r and R^2.

4.4 Explaining Variation by Association: Regression

R^2 gives us a measure of the relationship between two variables, tellingus how much variation in one variable is explained by another variable. Sometimes, however, we would like to visualize the relationship. One way to do this is to fit a trend line that shows the change in one variable with change in another. The simplest trend line is a straight line drawn through the data, such that it minimizes the distance from each observation to the line. This is called **linear regression**.

Figure 4.2 shows a linear (straight-line) regression fitted to measurements on some iris flowers. It illustrates the trend of increasing flower petal length with increasing petal with. The equation for a straight line is

$$Y = mX + b$$

and is often written as $Y = b + mX$ since the order of addition doesn't matter and b represents the intercept or overall mean before accounting for the effect of the the independent variable X. In this case, m is the slope of the line (change in Y for one unit change in X), which represents the effect of X on the dependent or response variable Y. The distance between each observed point and the best fitted line are referred to as the **residuals** and in a linear regression it must be assumed that they are normally distributed with a mean of zero (equal amount over and under the line).

It is also possible to fit other lines to data or manipulate equations so they fit the linear regression equation. Figure 4.3 shows the same data with a log-log linear regression.

This assumes an equation following

$$Y = bX^m$$

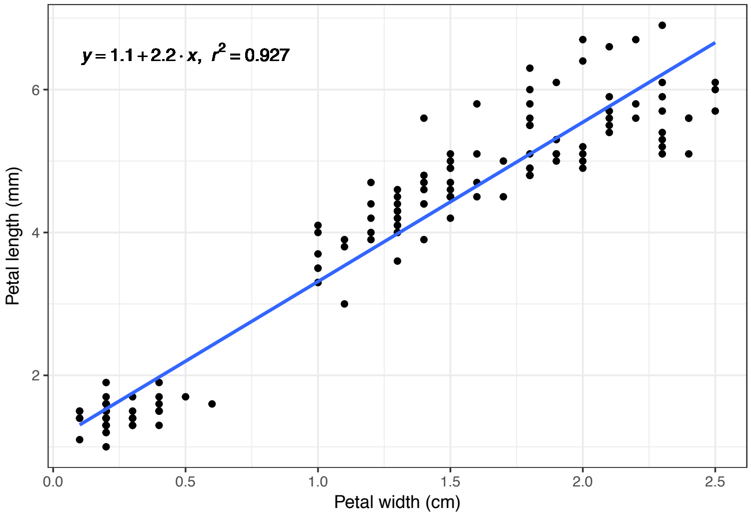


Figure 4.2: Example of a linear regression plot.

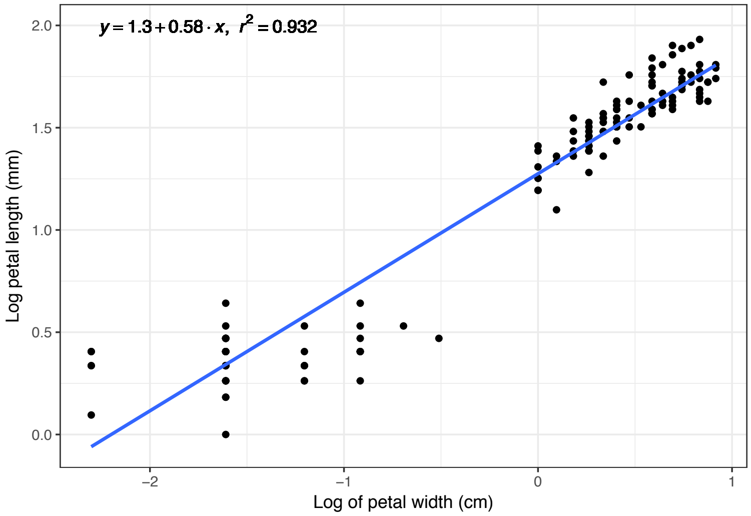


Figure 4.3: Example of a linear regression plot.

and would require a nonlinear regression without transformation of the data. By taking the logarithm of both sides of the equation is is transformed into

$$log(Y) = m * log(X) + log(b)$$

Which of the two lines fits the data better? To answer this kind of question, one is interested in the R^2 value. However, they both assume that the data meet the assumptions for the model which must be checked after the analysis.

4.5 Explaining Variation: Identifying random and nonrandom variation with probabilities

Odd things do happen, even if the probability of their occurrence is extremely small. For example, the probability that two different states would pick the same fout-digit lottery number on the sam day seems remote (it's typically on the order of 1 in one hundred million on any given day), but it has happened. Likewise, it is possible that a seemingly meaningful R^2 or other statistic of association was derived by chance alone rather than from variables that were actually related. To guard against this possibility we can determine the probability that the observed result did occur by chance. If this probability is low, then we can be reasonably (but never absolutely) sure that there is a real association in our observations. A "low" probability is often, but not always, defined as 0.05, or one chance in 20 that the result occurred by chance.

The probability of getting a certain result by chance is called its **statistical significance**. Significance is calculated from the known behavior of random variables, and the particular method depends on the statistic of interest. Computer programs are adept at calculating significance in an apprpriate manner. Significance is often reported as "p =", were p stands for probability. The probability values are often termed "p-values".

Thus, when inferring an association in your data, always look at the p-value as well as the measure of association or effect size of a variable. Only if the p-values is low (e.g., less than 0.05) can one have confidence that one has discovered a real relationship.

Correct interpretation of p-values: http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-correctly-interpret-p-val

4.6 Explaining Variation: Associating differences with group membership

Sometimes one wishes to determine if certain attributes are associate with certain groups or determine whether some characteristic differs among groups. For example, we might hypothesize that wormwood plants growing in the open are taller than those growing in the shade. In this example, we want to determine if "taller" is associated with "open" and if "shorter" is associated with "shade". In order for this to be true, the two populations must be truly different from each other. Suppose we collect data on sun grown plants and shade grown plants and observe the results in Figure 4.4.

We observe a difference btween the two samples, because they do not totally overlap and the means are slightly different (the mean for the shade plants is 5 and the mean for the sun plants is 8). They both have a standard deviation of 2. Can we conclude that our hypothesis ("taller" is associated with "sun" and "shorter" is associated with "shade") is true? The differences between the two groups could be due to chance or sampling erros. We didn't measure every wormwood in the whole world, and we may have accidentally take more taller plants from the "sun" group, when in fact the "sun" may have no affect on wormwood size.

Because out results do suggest a difference in plant height between the two groups, we ask ourselves, *"What is the probability of obtaining our results when in fact the observed difference between groups is jus due to sampling error?"*. If the probability of getting our results when there is no true difference is low, then the

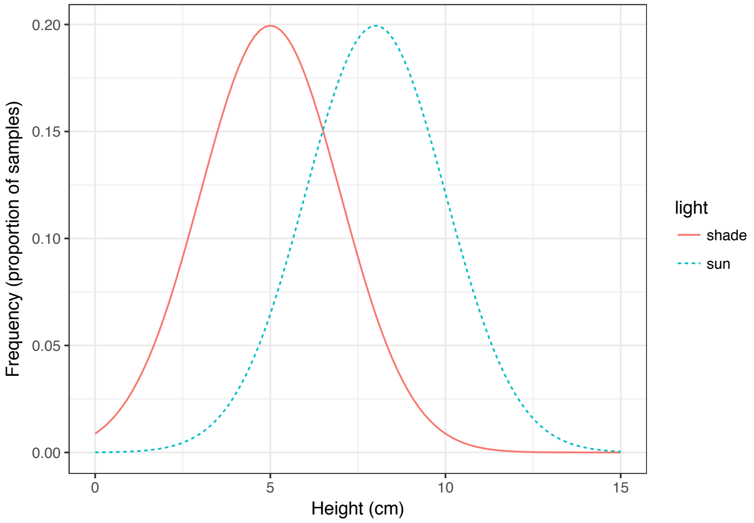


Figure 4.4: Two wormwood populations grown in sun and shade.

difference is probably real. Low probabilities of differences due to chance alone usually indicate meaningful (significant) differences between groups.

There are many ways of determining the probability of obtaining our result when there is no true difference. The two simpliest and most common ways are **t-tests** and the **Analysis of Variance (ANOVA)**. There are many variations of each and choosing the appropriate form requires skill and experience. In pratice it's best to consult with a statistician when designing and analyzing your study. *For this course, you will gain experience determining the appropriate analysis, analyzing data, and drawing inference based on the analysis. We will stick to basic forms of analysis that you can build from in future courses, projects, and jobs.*

An ANOVA is used for testing for a difference among multiple groups. A t-test is essentially a special case of an ANOVA when there are only two groups. Since ANOVA can still be used in cases of two groups, we will focus on ANOVA for this class.

4.7 Activity

Determine what type of test you would use to test each of your hypotheses developed previously? What data would you need to test the hypotheses?

What is one additional question that you had that would require a regression to evaluate (not one used for your hypotheses)?

What are general types of hypotheses that you can test with regression analysis?

What is the hypothesis you would test with an ANOVA?

When do you need to use an ANOVA instead of a t-test to answer the same type of question?

4.8 Introduction to Scientific Programming

Microsoft Excel is capable of doing some basic statistics including t-tests, ANOVA, and regression. However, the user has little control over the analysis, the software is proprietary, and you cannot build on these basics to analyze more complicated data or ask more interesting questions. MS Excel also varies considerably among computer operating systems and versions. Therefore, we are going to be using software designed for statistical analysis in this course, although you may use any appropriate means on your own homework (e.g. hand calculations, MS Excel, Stata, JMP, SAS, SPSS, Minitab, R, etc.). The demonstrations and instruction in this course will be using the **R Statistical Programming Language** and associated integrated development environment (IDE) **RStudio**. Both of these programs are free and open source and available on all major computer operating systems. They also generally look and act the same across operating systems and versions.

R has become the *lingua franca* of statistics and in the field of ecology. Besides the utility in this course and other science courses, having some familiarity with R and programming languages in general can open many graduate school and profession opportunities.

4.9 Assign and Recall an Object

The most basic action in R is to assign a value to an object so you can use the object or recall it later.

```
a <- 5
b <- c(1, 2, 2, 3)

print(a)
```

```
## [1] 5
```

```
b # you generally don't need the "print" command
```

```
## [1] 1 2 2 3
```

```
a * b
```

```
## [1]  5 10 10 15
```

4.10 Calculate the Mean, SD, and Variance

R also has a massive number of built-in functions, especially associated with data manipulation, printing, and statistics. Therefore, it's easy to calculate basic summary statistics such as the mean, median, mode, SD, and Variance.

```
mean(b)
```

```
## [1] 2
```

```
median(b)
```

```
## [1] 2
```

```
sd(b)
```

```
## [1] 0.8164966
```

```
var(b)
```

```
## [1] 0.6666667
```

```
sd(b)^2
```

```
## [1] 0.6666667
```

```
sqrt(var(b))
```

```
## [1] 0.8164966
```

4.11 Import Data

When interested in real data, we have to import the data into R. You can code it which as the advantage of being reproducible (like other methods in science).

```
salamander_data <- read.csv(file = "Data/salamanders.csv", header = TRUE, stringsAsFactors = FALSE)
```

RStudio also has the option under `File -> Import Dataset` to do this through the GUI.

4.12 Calculate Summary Statistics on the Data

One you have imported a dataset and assigned it to an object in R, you are ready to work with it. R has a nice `summary` function for datasets. You can also view the data through the options in RStudio or through R functions.

```
summary(salamander_data)
```

```
##      Plot              Species              Date               Count       
##  Length:72          Length:72          Length:72          Min.   :0.0000  
##  Class :character   Class :character   Class :character   1st Qu.:0.0000  
##  Mode  :character   Mode  :character   Mode  :character   Median :0.0000  
##                                                           Mean   :0.4444  
##                                                           3rd Qu.:0.0000  
##                                                           Max.   :5.0000  
##  Cover_Objects  
##  Min.   : 0.00  
##  1st Qu.:10.00  
##  Median :15.50  
##  Mean   :15.22  
##  3rd Qu.:17.00  
##  Max.   :35.00
```

```
head(salamander_data)
```

```
##   Plot                           Species    Date Count Cover_Objects
## 1 1--1 Allegheny Mountain Dusky Salamander 10/4/16     0            19
## 2 1--2 Allegheny Mountain Dusky Salamander 10/4/16     1            15
## 3 1--3 Allegheny Mountain Dusky Salamander 10/4/16     0            16
## 4 1--4 Allegheny Mountain Dusky Salamander 10/4/16     0            13
## 5 1--5 Allegheny Mountain Dusky Salamander 10/4/16     0            15
## 6 1--6 Allegheny Mountain Dusky Salamander 10/4/16     0            17
```

```
str(salamander_data)
```

```
## 'data.frame':    72 obs. of  5 variables:
##  $ Plot         : chr  "1--1" "1--2" "1--3" "1--4" ...
## $ Species    : chr "Allegheny Mountain Dusky Salamander" "Allegheny Mountain Dusky Salamander" "Alleghe
##  $ Date         : chr  "10/4/16" "10/4/16" "10/4/16" "10/4/16" ...
##  $ Count        : int  0 1 0 0 0 0 3 1 1 0 ...
##  $ Cover_Objects: int  19 15 16 13 15 17 10 8 16 6 ...
```

4.13 Subset and Recalculate Summary Statistics

You can also subset the data and perform actions on the new object.

```
pcinereus <- salamander_data[which(salamander_data$Species == "Red-backed Salamander"), ]

summary(pcinereus)
```

```
##      Plot             Species              Date               Count       
##  Length:18          Length:18          Length:18          Min.   :0.0000  
##  Class :character   Class :character   Class :character   1st Qu.:0.0000  
##  Mode  :character   Mode  :character   Mode  :character   Median :0.0000  
##                                                           Mean   :0.9444  
##                                                           3rd Qu.:2.0000  
##                                                           Max.   :5.0000  
##  Cover_Objects  
##  Min.   : 0.00  
##  1st Qu.:10.75  
##  Median :15.50  
##  Mean   :15.22  
##  3rd Qu.:17.00  
##  Max.   :35.00
```

```
mean(pcinereus$Count)
```

```
## [1] 0.9444444
```

```
sd(pcinereus$Count)
```

```
## [1] 1.349171
```

4.14 Save Script

Once you have finished working on something, it's important to save the script so you can rerun the analysis later, share the code, or reuse code in other places. You do this through RStudio and the file ending is `.R`. You can also output objects as CSV files or as RData files to work with later or work with outside of R.