ENV 226 Lab Online R Manual

Sara Souther

2025-08-22

Contents

Chapter 1

Introduction

This course surveys the central concepts in ecology: evolution, population dynamics, community interactions, biogeochemical cycling, and limiting factors, as well as how those factors are measured, quantified, and interact with drivers of global environmental change. This course is required for the B.S. in Environmental Sciences degree program, and also the B.S. in Environmental and Sustainability Studies degree. This course acquaints students with foundational concepts and theories in ecology and provides a broad basis for more advanced courses in subdisciplines and applications of ecology.

In lab, we will conduct experiments and practice skills used in the ecological sciences. In addition to field and experimental techniques, these skills include data collection and analysis. Learning how to manage, manipulate and analyze data in R will serve your undergraduate career and beyond! In ENV 226 lab, we will ease you into using R for all your data needs!

How to use this resource

Each chapter in this online book corresponds to lab that you will complete. For most labs, you will download R code that you will use to analyze your data, and that can be also be used for other analyses and projects that you have in college and in the workplace.

Chapter 2

Welcome to R

R is an open source statistical software package commonly used by researchers and other folks, who crave a free way to manipulate, analyze, and visualize data. R uses its own programming language, which is similar to S+ (the paid precursor to R). R employs an object-oriented programming (OOP) paradigm to manage and manipulate data. Object-oriented programming (OOP) is a paradigm where data and functions are grouped into 'objects' that can be reused, helping organize and simplify code.

Today, R is essential for ecological work that involves data analysis, whether you are participating in a graduate program and analyzing data for your thesis or working at a non-profit analyzing the efficacy of restoration treatments. We are using R in this lab to introduce you to the basics of data analysis. Managing, manipulating and analyzing data are important skills to include on your resume and will hopefully help you in other classes at NAU. Let's walk through the basics of installing and using R!

2.1 R and R studio installation

You will first want to download R statistical software and R studio, which is a powerful program that interfaces with R to make your coding experience more organized and enjoyable. Notice that you need to select a version of R depending on your operating system.

2.2 Download the R statistical software from the official R Project website.

Open your web browser and go to the official R Project website at https://www.r-project.org/.

Choose a CRAN Mirror: On the R Project website's main page, you'll see a section that says "Download and Install R." Click on the link that says "CRAN (Comprehensive R Archive Network)." This will take you to the CRAN website.

1. Select Your Mirror: On the CRAN website, you'll find a list of mirrors (servers) from which you can download R. Choose a mirror that is geographically close to your location, as this will generally provide faster download speeds. Click on the mirror's link.
2. Download R for Your Operating System: On the mirror's page, you'll see options to download R for various operating systems (e.g., Windows, macOS, Linux). Click on the appropriate link for your operating system.
3. Choose the Latest Version: You'll typically see multiple versions of R available for download. It's recommended to choose the latest stable version unless you have a specific reason to use an older version.
4. Download and Install: After clicking on the download link, the installation file for R will begin downloading. Once the download is complete, run the installer and follow the installation instructions for your operating system.
5. Start Using R: After the installation is complete, you can launch R from your computer. First, let's install RStudio, a popular integrated development environment (IDE) for R, to enhance your R programming experience.

2.3 Download R studio

Now download R studio!

1. Visit the RStudio Website: Open your web browser and go to the official RStudio website at https://www.rstudio.com/.
2. Download RStudio: On the RStudio website's main page, click on the "Products" menu at the top, and then select "RStudio" from the dropdown menu.
3. Choose the RStudio Edition: RStudio offers different editions, including RStudio Desktop (for use on your local machine), RStudio Server (for remote access), and RStudio Workbench (formerly known as RStudio Server

Pro, designed for collaboration and sharing in enterprise environments). You will want to choose the free version, RStudio Desktop.

4. Download the Installer: After selecting the edition, you'll be directed to a page with download options. Click on the download link for your operating system (e.g., Windows, macOS, Linux).

5. Download and Install: The installation file for RStudio will begin downloading. Once the download is complete, run the installer and follow the installation instructions for your operating system.

Start Using RStudio: After the installation is complete, you can launch RStudio from your computer.

2.4 Optional: Set up R studio

Alright, now that you've downloaded R and R studio, open R studio. You can customize the panes that you are visualizing in R.

In RStudio, the four panels or panes are commonly referred to as:

Source Pane: This is where you can write, edit, and save your R scripts and code files. It is typically used for script development and editing. You can open and create new R script files in this pane.

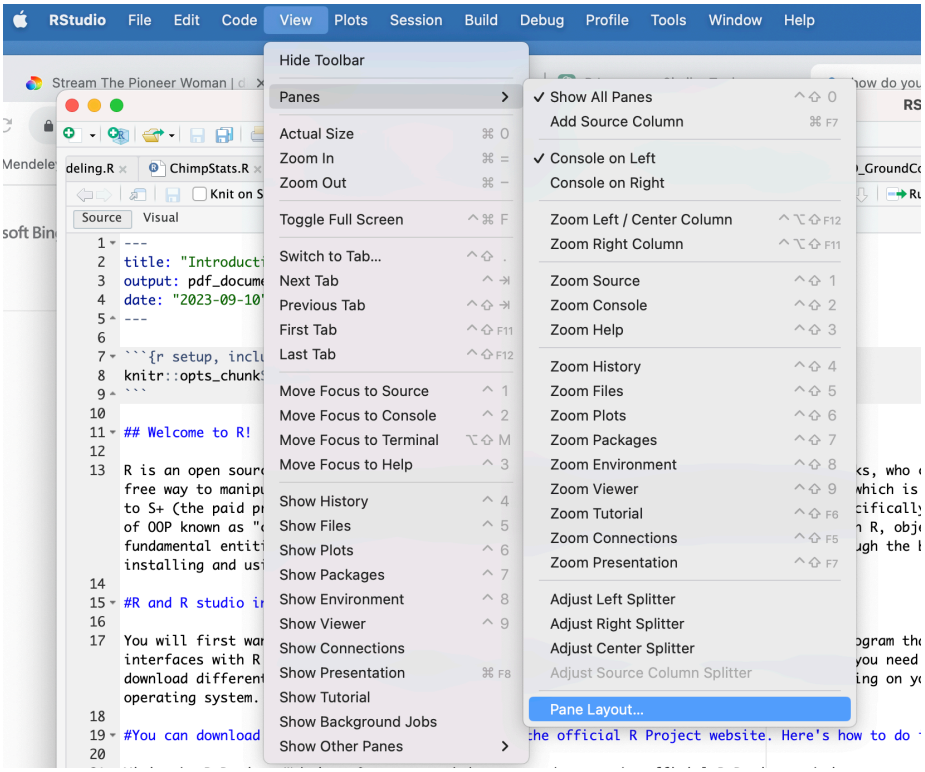
Console Pane: The console is where you interact with R directly. You can execute R commands and see their output here. It's an interactive environment where you can test and run R code line by line or in batches.

Environment Pane: The environment pane displays information about the objects, data frames, variables, and functions currently loaded in your R session. You can also use this pane to view data frames in a spreadsheet-like format and manage your workspace.

Files/Plots/Packages/Help Pane: This pane has multiple tabs and serves various purposes: *Files*: It shows the file system of your project, allowing you to navigate and manage files and directories. *Plots*: When you create plots in R, they will appear in this tab. You can interact with and export the plots from here. *Packages*: This tab displays information about installed packages, and you can use it to install, update, or load packages. *Help*: When you need documentation or help for R functions or packages, you can use the Help tab to search for and view documentation.

Typically, I select a structure in which I have my **Source pane** in the upper left, my **Console Pane** in the lower left position, my **Environment Pane** in the upper right corner, and the **Files/Plots/Packages/Help Pane** in the lower right position. You can select any position that you'd like, but if we create the same work environment, it will be easy for me to direct you when we

are trouble-shooting code. To adjust the panels positions, use the pane layout function. Here's what that looks at for a Mac, but typically this arrangement is the default positioning for panels in R studio, so you likely won't have to adjust positioning!



2.5 Downloads and video links

Watch the tutorial for this lab project: YouTube Video

Here is a nice example of a R script: Download the R file Download and use the R script to practice with R.

Also, download this .csv file Download the data file We will use it to practice importing a file.

2.6 Using R

1. Create a folder on your desktop called: EcologyLab

2. Download the R file and .csv file above
3. Once you've downloaded the R script, open it in R studio:
    - File -> Open file
4. You will see several things:
    - Text with a hastag in front of it (#This code …). R won't run code with a hastag, so we can use this to explain what chunks of code do
    - Code to load and install packages.
    - Code to set your working directory
    - Code in import data
    - Code to export data

We will go through each of these codes during this laboratory exercise.

2.6.1 Packages and libraries

Base R, the fundamental, built-in set of functions, data structures, and libraries that come with the R programming language without the need for additional packages or extensions, including basic math functions, statistical analyses and visualization tools. However, one amazingly cool think about R is that folks are out there creating 'libraries', or a collection of R functions, data sets, and documentation bundled together into a single package, to do specialized analyses. For most tasks in R, you will need to install and load libraries.

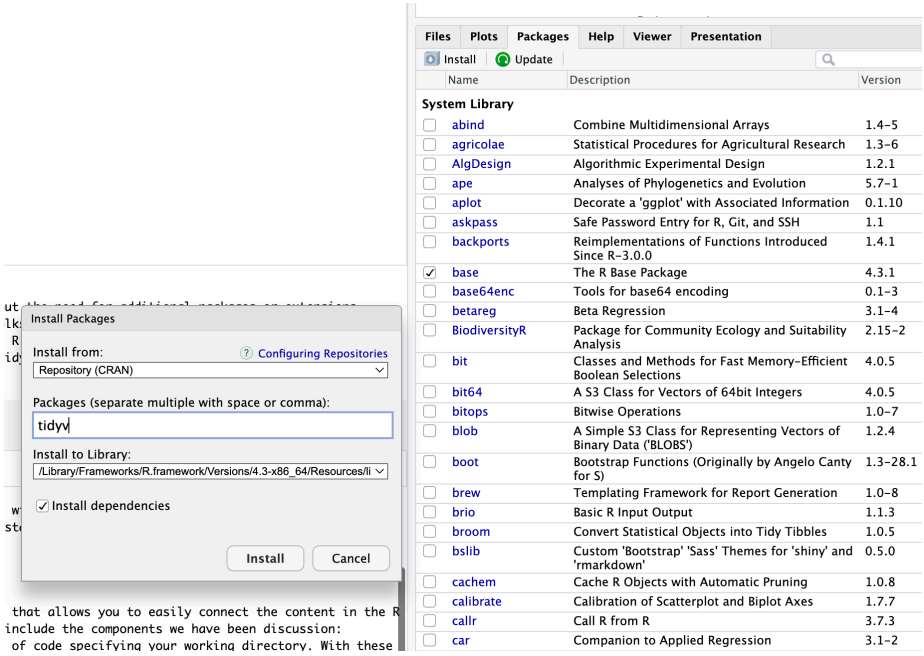
2.6.1.1 Install a library

There are two methods to install libraries:

1. Manually: Go to the Rstudio pane 'Packages' -> Install -> Search for your package and install
2. Run code: install.packages("tidyverse")
    - To run code, place your cursor on a line of code (any where on install.packages("tidyverse")) and press the **Run** button at the top of this pane.

Please install a library (or package) for data manipulation, called 'tidyverse' (actually several packages - hence why the name references a universe) using either method. When prompted, be sure to **install dependencies** - this will make sure that you have any pieces of code that the library that you are installing needs to operate.

The weird thing about including install packages code is that you don't want to re-install packages every time that you use R (in fact, it caused my R markdown code to freak out, which is why I've included the install function in the text). You will need to load packages, you generally won't need to install packages after you have done it once. You can either then install the packages and delete/hash out the install code OR you can install through the R studio interface by going to packages, selecting install and searching for and installing the packages that you are interested in.



Excellent! You have installed a library! **Now, we need to load it.** To load a library, run this code:

```
library(tidyverse)
```

Alternatively, to load packages, you can select the package name in the 'Packages' pane. You WILL need to load R packages every time that you use R. Any functions associated with your package won't work, unless the package is loaded, so I suggest keeping the load library code in your R script, rather than loading manually.

2.6.2 Working directories

When you are coding in R, you will want to save your R or R markdown scripts and any other data files (e.g., .csv or spatial files) that you are analyzing in a

common file. You can use R studio interface to import files (Go to files tab); however, with good housekeeping, you will be able to seamlessly rerun your analyses at in point in time, allowing you to pick back up on projects that may have been dormant!

The first step in good R housekeeping is to set a **working directory**. A working directory tells R where to look for and save files. Now, let's set your working directory to the folder containing the R script and .csv file called 'EcologyLab'. As always, there are two options:

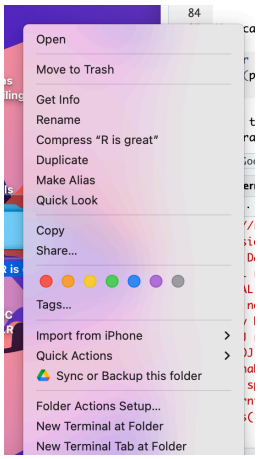
1. Manual working directory selection: You can change your working directory using the R studio interface by selecting a working directory at the top of the console panel in the "Files" tab. If you want to change your working directory to a different location, click on the "..." (ellipsis) button in RStudio's "Files" tab.
2. Set the working directory using code.

You will be *far* better served by including code in your R script that directs R to your working directory. I prefer to set it within the code in order to allow you to instantaneously be able to pick up work where you left off rather than searching through files and trying to remember how you set up the code. You can view your working directory by running a simple bit of code (run code below).

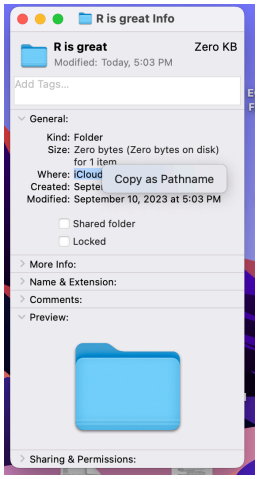
The code to set your working directory is in the R script that you loaded: setwd("your_path_to_your_working_folder"). The only challenging part is to identify your working directory. However, once you have found it, and if you put all your files into one folder, you can copy and paste this script into every code you use throughout the semester to set your working directory.

2.6.2.1 Setting your working directory on an Apple device

To find the file path to the Class1_IntroToR on a mac, double click on the file and should see several option, including 'Get info' (check out picture below).



Then, select ‘Get info’. Then, highlight the information after ‘where’ and copy it as a path name (see picture). Now you can paste the path to your working folder in your R code to set the working directory.



2.6.2.2 Setting your working directory on a PC

For PCs, start by opening your File Explorer:

Press the Windows key + E on your keyboard. Alternatively, you can click the “File Explorer” or “This PC” icon on your taskbar or Start menu. Navigate to the Folder: Use the File Explorer to navigate to the folder where the file is located. You can click on folders to open them and view their contents. Then, proceed to:

1. Find the File: Locate the file you are interested in within the folder.

2. View the File Path: Once you've found the file, you can see its full file path in the address bar at the top of the File Explorer window. The file path will be displayed as a sequence of folder and file names separated by backslashes. You can click in the address bar and copy the file path to the clipboard by pressing Ctrl + C after selecting it.

2.6.2.3 Adjust your code

Once you have copied your file path, paste that path name into the following code and set your working directory: setwd("your_path_to_your_working_folder"):

- Example for my mac: setwd("/Users/sks379/Desktop/EcologyLab/")

Now R studio is directed to upload and save work to this folder.

2.6.3 Trouble-shooting issues

When coding, every symbol is important - meaning that it is extremely easy to make mistakes and have your code not run. One easy way of troubleshooting your code is to use AI. For this class, I suggest using the free version of ChatGPT.

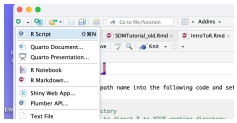
1. Access ChatGPT on a Web Browser
    - No Installation Required! The free version of ChatGPT does not require downloading or installation:
        1. Open a web browser (e.g., Chrome, Firefox, Safari).
        2. Visit ChatGPT's website.
        3. Log in using an existing account, or sign up for free using an email address, Google account, or Microsoft account.
2. Use ChatGPT on Mobile
    - Mobile-Friendly Website:
        - Simply visit chat.openai.com on your mobile browser and log in.
    - Mobile Apps (Optional):
        - OpenAI offers official apps for ChatGPT on iOS and Android devices. Here's how to download them:
        - iPhone/iPad:
            1. Open the App Store.
            2. Search for "ChatGPT" by OpenAI.
            3. Tap Download and install the app.
        - Android Devices:

1. Open the Google Play Store.
2. Search for "ChatGPT" by OpenAI.
3. Tap Install to download the app.

If you are having a problem setting your working directory, copy and paste your code into chat and explain the error. **Were you able to troubleshoot your problem?**

2.6.4 Annotating your code

Notice anything about the code in the section you just ran? You can use the hast tag symbol to tell R not to run a section of code. Whenever I am generating code, I try to add lots of notes to myself, so that that future me knows what code I created and why. Annotating your code is just good practice for coding! Alternatively, you can create R markdown files (what this tutorial has been created in), but R markdown, while generating pretty PDFs and websites, adds an extra layer of complexity that you generally don't want or need while coding, so I typically recommend creating an R script and annotating your that file!

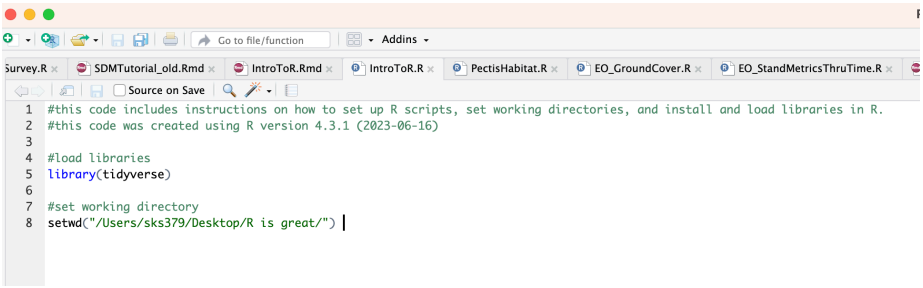


One thing that you might want to include in your code description is the version of R that you are using (you may need to load older versions of R if your scripts stop working due to updates to the program). To check the version of R that you are using, paste this in the command line: R.version.string

2.6.5 Organizing your code

A well-written R script will include the components we have been discussion: Annotated notes on what the script is doing and potentially even the version of R, loading commands for your libraries, and a line of code specifying your working directory. With these elements in place, you are ready to code your heart out!

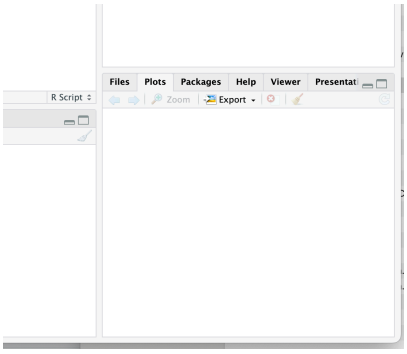
Here is a glimpse at what your R scripts should look like:



2.6.6 Importing and exporting csv files

You may want to import data, and export a .csv file after analyses to include as a table in your results. The code to do this is simple:

Finally, one easy way to export plots is to use the 'Plots' tab in your R studio window. Simply click on the 'Plots' tab and select export - you can choose different file types and sizes when you export.



Deliverables

Download both files, modify your working directory and import the file. Show your TA that you successfully altered your working directory and imported a file to receive credit!

And with that, welcome to the wonderful world of coding in R!

Chapter 3

Descriptive statistics

3.1 A short statistical review

3.1.1 What can statistics tell us?

Welcome to our statistical exploration of the natural world!

Almost all statistical analysis boils down to answering 1 of 2 questions:

- Do these groups differ?
- Is there a relationship between these variables?

These seem like relatively simple questions to answer, perhaps just by looking at our data, so **Why do we need statistics?**

The short answer is: **error** and **sampling**! Whenever we collect data, we introduce **error**; our instruments are imprecise and do not capture an exact measure of whatever you are measuring (e.g., height, weight), and humans make mistakes during measurement collection. Secondly, we are **always** measuring a sub-sample of the true population (*true population* meaning all representatives of whatever you are trying to measure; this can be grass, marbles, or the tibia of humans). Not only is it intractable in most cases to measure all individuals of whatever you are interested in, even when it is possible to attempt to measure **all** individuals (like in the case of rare plant work), statistics acknowledges that it is **still** unlikely that we are able to do so, since individuals may be dormant or challenging to locate. If we could measure all individuals of our focal population with perfect accurately, we could calculate population **parameters**, or quantities describing populations like averages and variation, rather than estimating these metrics, and just compare them. In this way, statistics is inherently practical, and asks: What can we say about whatever we are looking at, given our numerous flaws?

3.1.2 Sampling populations

After a few classes, we will explore sampling methodology in greater depth in order to design appropriate experiments that test a statistical **hypothesis**. Let's quickly talk about sampling now so that we have a shared understanding and vocabulary to build on - after all, statistics really centers around estimating characteristics of a true population from a sample. The **really, truly** amazing thing is that by properly applying statistics, we can learn practically anything about almost any population using samples!

In statistics, a **population** refers to the all units of the thing that you are interested (i.e., all suriname frogs, all grains of sand, all aspen leaves from a genotype found in southern Arizona). **Note**: Population in statistics differs from the term population in population ecology, where a population refers to a group of individuals in a particular area that interbreed.

A **sample** is a subset of the population that we measure to infer something about the population. **Statistical analysis is only one part of presenting your research results.** Generally, a results section in a manuscript includes: **statistical results**, **data description** (e.g., describing means, ranges, maxima, minima of groups of interest), and **data visualization** (i.e., creating beautiful figures).

3.2 Objectives

We will build to towards always providing a complete results section with these three components. In this lab, however, we will focus exclusively on data description and a teeny bit of visualization.

3.3 Downloads for this lab

Last week, you should have created a folder to keep files for lab. Download these materials and place in your folder:

- For your reference, here is a guide that shows the basics formulas for calculations in R: Supplementary Material
- Download the example dataset

The code can be used to import the data, but in case you wanted to access the data yourself - provided above!

- Build on this code, to generate your description of the petal lengths of your focal species: R Script for Chapter 1

- Watch this video if you need additional help

Remember to alter the working directory, so that you are importing and exporting files from the folder for this lab.

3.4 Data types

Before we start learning to present research results (analysis, description, visualization), let's talk about data! Data comes in several varieties, and the variety dictates which statistical analysis we choose!

Categorical variables are non-numeric variables. **Examples**: Pet type (dog, cat, fish, bird), Size (small, medium, large), Car type (sedan, SUV), Present/Absent

Numerical variables are variables that are numbers, and occur in two forms: *Discrete* = Counts of things (no decimal points/fractions) Data are discrete when it does not make sense to have a partial number of the variable. For instance, if counting the number of insects in a pond, it does not make sense to count a half a species. **Examples**: Number of people in a building, number of trees in a plot, number of bugs in a pond

Continuous are numerical data that can occur at any value. These are variables that can occur in any quantity. If you can have a fraction of this variable, it is continuous. **Examples** = Height, Weight, Length

Ordinal variables (sometimes referred to as ranked) can be categorical or numerical, but the order matters. **Examples** = Grades (A, B, C, D, E), Likert scale variables (Strongly disagree, Agree, Strongly Agree), Class rank (1, 2, 3, 4, 5)

3.5 Descriptive statistics

Descriptive statistics quantify characteristics of a population. Critically, they are *not* statistical tests, which tell us whether groups differ or if variables are related, BUT they are an important part of scientific work and data analysis. First, we will start by describing continuous data.

Let's use a simplified version of a dataset that I'm working with right now to look at the performance of several species of pollinator-friendly native species in agricultural gardens. Eventually, we'd like to develop seed to provide to restorationists for restoration of arid and semiarid grasslands. To do this, we need to understand how reliable these species are at establishing, producing seed, and attracting pollinators. Initially, we are conducting experiments with multiple populations of each species to determine how consistently plants grow,

reproduce, and perform. Here, We will take a look at the initial heights of 1 population of one species, *Asclepias subverticulata*.

Most of the time when writing up results, you present a mean (sum of numbers divided by the number of observations), and an estimate of variation (a measure of how different the observations are). Here, we calculated three estimates variation, variance, standard deviation, and standard error.

For the following sections of code, practice code is in your R file

```
#create vector of heights (cm) of one population of A. subverticulata
sedonapopulation <- c(3, 3, 3, 3, 7, 8, 9)
#take the mean
mean(sedonapopulation)
```

```
## [1] 5.142857
```

```
#calculate variance
var(sedonapopulation)
```

```
## [1] 7.47619
```

```
#calculate standard deviation
sd(sedonapopulation)
```

```
## [1] 2.734262
```

```
#calculate standard error
#base r doesn't have this function
#so we have to write our own
std_error <- function(x) sd(x)/sqrt(length(x))
std_error(sedonapopulation)
```

```
## [1] 1.033454
```

Since you will occasionally need to include equations in your write-ups, let's get use to mathematical syntax, with these simple examples.

The formula for the sample mean is: $\mu = \frac{\Sigma x_i}{n}$; where μ indicates the sample mean (sample = group of numbers we are looking at); Σ means to add what ever follows; x_i is the value of one observation; (subscript i is often used to indicate that the action should be repeated for all values); n is the number of observations

Why didn't we just use $\bar{x}$ to indicate the mean? Because statisticians typically use $\bar{x}$ to indicate the true mean of the population, and μ to indicate the sample mean!

Just to show you, what the mean() function is doing, let's run:

```
sum = 3+3+3+3+7+8+9 #add all the numbers in the sample
n = length(sedonapopulation) #or you can just calculate the number of height measurements
mean = sum/n; mean #divide sum by number
```

```
## [1] 5.142857
```

This formula is simple, but sometimes with more complex formulas, I will solve the equations by hand, to make sure that I understand what is happening!

The formula for variance is: $S^2 = \frac{\Sigma(x_i - \mu)^2}{n-1}$ where S^2 is the sample variance; μ is the sample mean (remember from above); x_i is the value of one observation; n is the number of observations

In other words:

```
#We determine how much each observation varies from the mean.
diffobs1 = mean - 3
diffobs2 = mean - 3
diffobs3 = mean - 3
diffobs4 = mean - 3
diffobs5 = mean - 7
diffobs6 = mean - 8
diffobs7 = mean - 9

#Then we square each of these.
diffobj1_sq = diffobs1^2
diffobj2_sq = diffobs2^2
diffobj3_sq = diffobs3^2
diffobj4_sq = diffobs4^2
diffobj5_sq = diffobs5^2
diffobj6_sq = diffobs6^2
diffobj7_sq = diffobs7^2
```

Why do we square the differences rather than just adding them up? Because differences will be positive and negative. If we added them without squaring, sample differences would negate each other. We want an estimate of the absolute differences of samples from the mean.

```
#Then we add the differences up.
sumofsquares = sum(diffobj1_sq, diffobj2_sq, diffobj3_sq, diffobj4_sq, diffobj5_sq, di
#Divide the sum of squares by n - 1.
variance = sumofsquares/(n-1); variance
```

```
## [1] 7.47619
```

Why n - 1 instead of n? One reason is that, theoretically, because we are taking the mean of a sample, rather than all individuals, we underestimate the variance, so taking n-1 corrects that bias. Consider it a penalty for measuring a sample, not the entire population! Another practical reason is that dividing by n-1 makes the variance of a single sample undefined (unsolvable) rather than zero (solvable)

For standard deviation, we just take the square root of the variance, to remove the effect of squaring the differences when calculating the variance, and thus contextualizing our estimate of variation with regard to the mean. For example, the variance for the Sedona population is 7.48, larger than the sample mean of 5.12; while the standard deviation is 2.73, indicating that you would expect most observations to be 5.12 +/- 2.73 (we'll get to quantiles in a minute).

The formula for standard deviation is: $\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{n-1}}$ where σ is the sample variance; μ is the sample mean; x_i is the value of one observation; n is the number of observations.

Finally, standard error and confidence intervals (we'll get to confidence intervals later) are the most common metrics of variance presented in journals.

The formula for standard error is: $SE = \frac{\sigma}{\sqrt{n}}$ where SE is standard error of the sample; σ is the standard deviation; and n is the number of samples.

Why do we divide the standard deviation by the square root of the sample size to get standard error? While standard deviation measures the variation of the sample, standard error is meant to estimate the variation of the entire population of samples, if we could measure all individuals accurately. By dividing by the $\sqrt{n}$, the larger the sample size, the lower the error, because you have a more complete estimate of the true mean. In other words, standard deviation is just a measure of the variation of our sample, while standard error also incorporates information about our sampling process (how many individuals we have sampled). *Want to delve deep into standard error and deviation (me neither - ha)?: Google central limit theorem + standard error / standard deviation.*

Means and variance measures are the most common way to describe quantitative data. However, several other metrics are useful for understanding the nature of your data and making decisions about analyses. A comprehensive understanding of your dataset includes describing these four features: *Location (Mean, Median)* Spread (Variability) *Shape (Normal, skewed)* Outliers

We've talked about means. The median is just the central number in the dataset, and helps you identify skewness.

```
#an example of an unskewed population
sedona_unskewed <- c(1, 2, 3, 4, 5, 6, 7)
mean(sedona_unskewed)
```

```
## [1] 4
```

```
median(sedona_unskewed)
```

```
## [1] 4
```

```
#previous sedona population; skewed
sedonapopulation <- c(3, 3, 3, 3, 7, 8, 9)
mean(sedonapopulation)
```

```
## [1] 5.142857
```

```
median(sedonapopulation)
```

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

In an unskewed population, the mean will equal the median. Skew may not seem important, but it has statistical ramifications, AND it tells us something meaningful about the data. For instance, what if I said that mean price of a home in Flagstaff is 350K, but the median price of a home is 300K? We would know the that average house prices are driven up by a smaller number of expensive homes.

We can quantify skew by comparing means and medians (mean > median = right-skewed; median > mean = left-skewed), but it is helpful to visualize the shape of data with a **histogram**. A **histogram** is a graph of the frequency of different measurements.

Let's add a few more observations to our Sedona populations (skewed and unskewed) and check out the look of the data!

```
sedona_unskewed <- c(7, 2, 2, 3, 3, 3, 3, 6, 6, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 0.5)
mean(sedona_unskewed)
```

```
## [1] 3.975
```

```
median(sedona_unskewed)
```

```
## [1] 4
```

```
#I'm renaming sedonapopulation, sedona_skewed for this example
sedona_skewed <- c(3, 3, 3, 3, 7, 3, 4, 5, 6, 3, 3, 3, 4, 4, 6, 7, 8, 9, 3, 4, 5, 2)
mean(sedona_skewed)
```

```
## [1] 4.454545
```

```
median(sedona_skewed)
```

```
## [1] 4
```

```
## pdf
##   2
```

In this relatively unskewed example, the tails are approximately even. This shape is also referred to as a normal or Gaussian distribution.

```
## pdf
##   2
```

Here, we superimposed the bellshaped Normal or Gaussian distribution.

```
## pdf
##   2
```

In this example of skewed data, the tail tapers to the right, indicated that the data is skewed to the right.

In order to explain outliers, we need to look at quantiles! Quantiles are proportions of your data, in other words a way to break your data into chunks to understand spread. You can break your data into as many quantiles as you would like, but it is most common to break your data into 4 parts, also called quartiles. (If you break data into 5 parts, the components are called quintiles, 10 parts = deciles, 100 parts = percentiles).

When you break data into quartiles, roughly 25 percent of the data occurs within each data chunk. The first chunk of the dataset contains 25% of the data (25th percentile; 25% of the data fall at or below this cut-off) is called the first quartile, the 50th percentile is called the sample median or the second quartile, the 75th percentile is called the third quartile.

Box and whisker plots are commonly used to quickly examine quartiles. Let's check out our plant height data again, using a box and whisker plot.

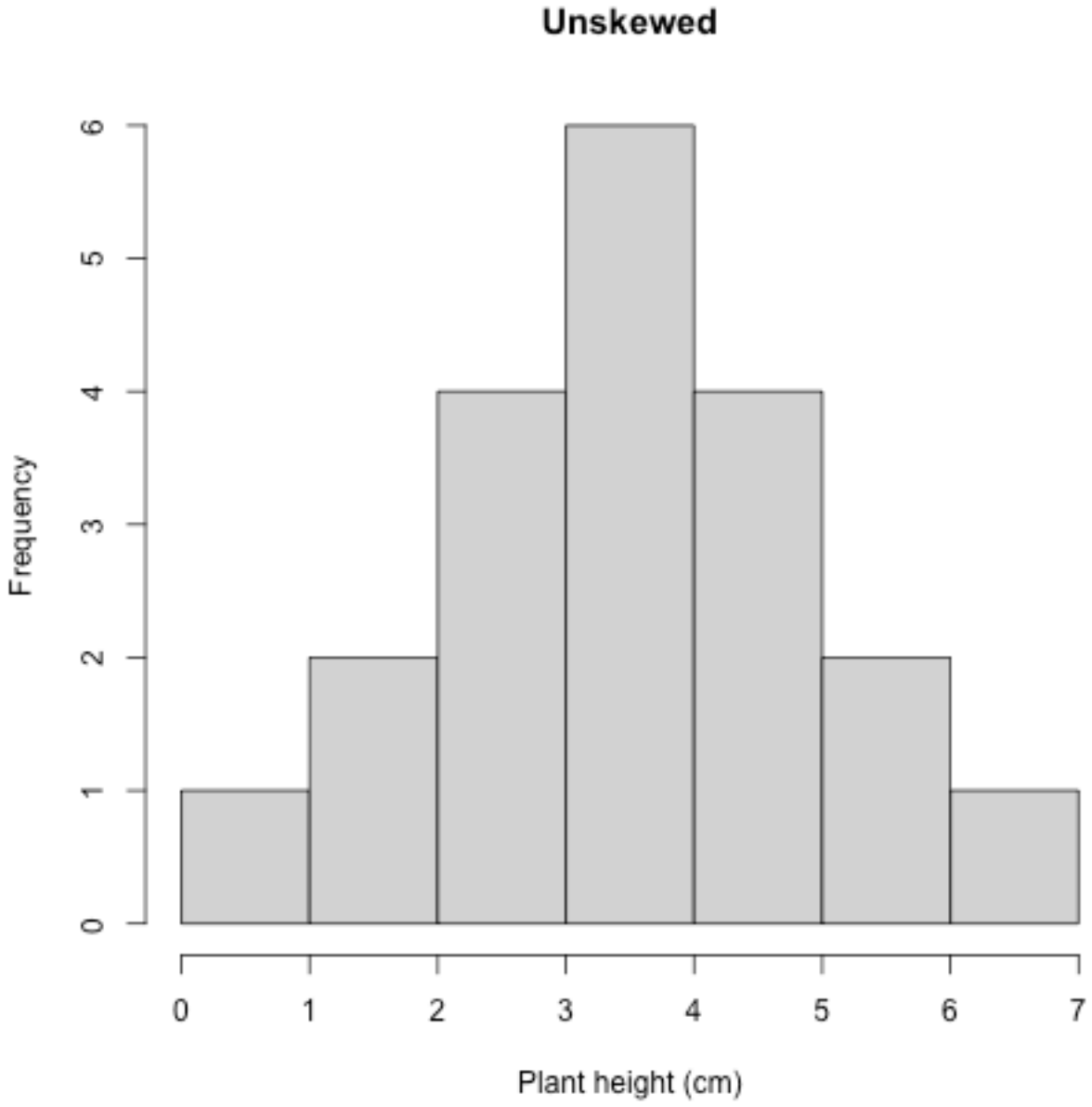


Figure 3.1: Histogram unskewed plant height

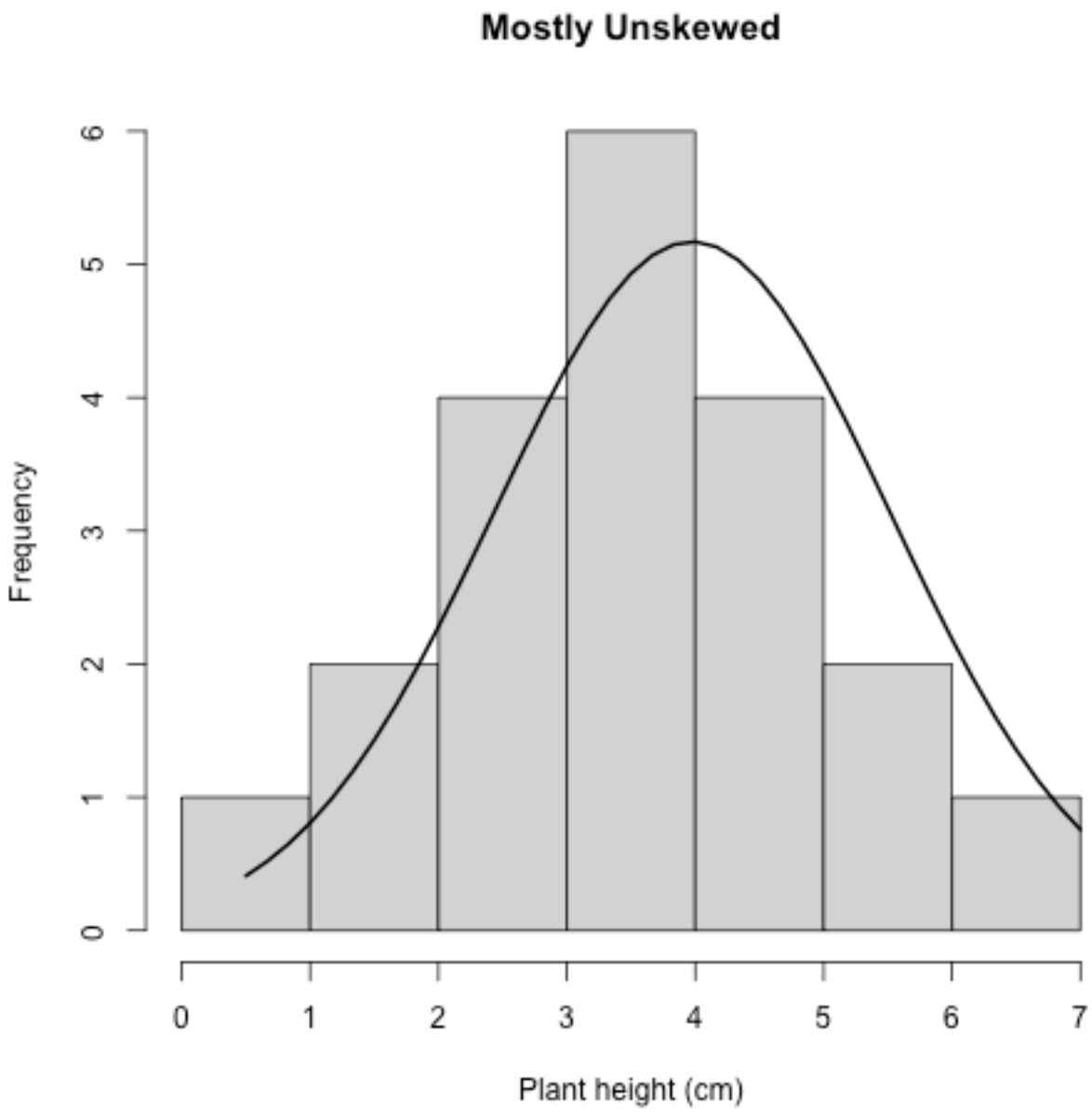


Figure 3.2: Histogram unskewed plant height

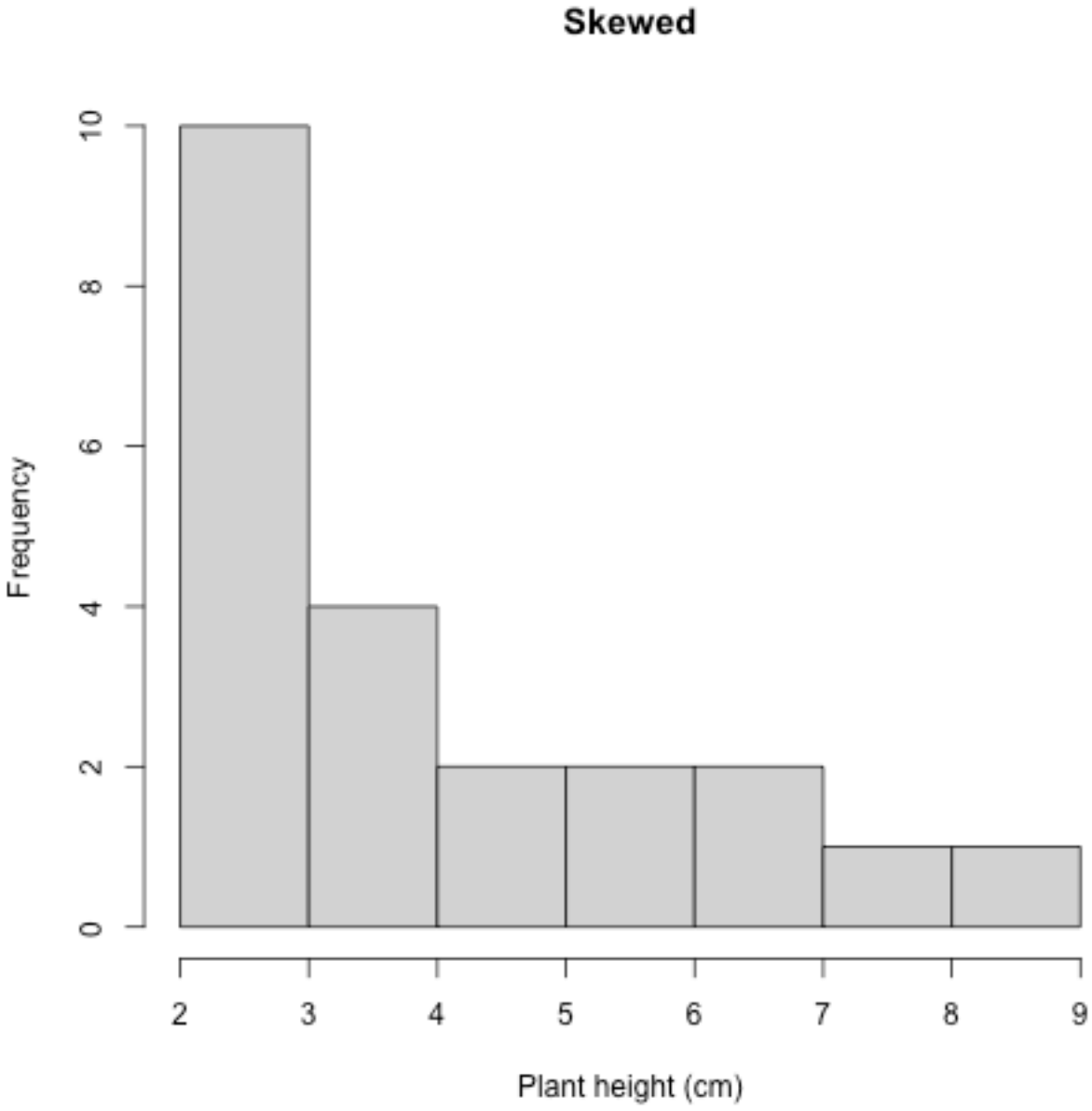


Figure 3.3: Histogram skewed plant height

```
## pdf
##   2
```

In the plot shown here, the box encapsulates the Interquartile Range (IQR); the center of the data ranging from the 25th percentile to the 75th. The black line in the middle of the box is the median (also called the 50th percentile, because it bisects the dataset; half of the data occur above the median and half below). The lines emerging from the box (whiskers) indicate the extent of the first and third quartiles, and usually corresponding with the minimum and maximum values of the dataset, unless there are **outliers**. An outlier is a datapoint that occurs outside of the 1st or 3rd quantile. Let's add one to our Sedona dataset, and see how it is represented on the box and whisker plot.

```
## pdf
##   2
```

The outlier appears as a dot on the box and whisker plot, and is the maximum value of the dataset.

One other thing to note: Standard deviation also breaks data into meaningful segments, but is only used when data conform to a normal distribution; the mean +/- 1 SD accounts for 68% of the data, +/-2 SDs contains 95% of data, and +/- 3SD includes 99% of data. That said, I've never presented standard deviation in a manuscript; it is much more common to include standard error or confidence intervals (discussed later).

3.6 Figures

Figures are a common way to show key elements of your data or statistical analysis. In this tutorial, you've seen both histograms and box plots (aka box and whisker plots). When presenting results in a scientific manuscript or report, figures will be accompanied by figure legends. A good figure legend provides the following information:

1. A description of what the figure is showing you.
2. For complex figures with multiple panels, an orientation to the structure of the figure
3. Explanation of any symbols, colors, and/or lines
4. Description of how variance is quantified
5. Definitions of axes or units, if unclear
6. Acknowledgment of data source, if data source requires attribution
7. Implications of the figure (optional)

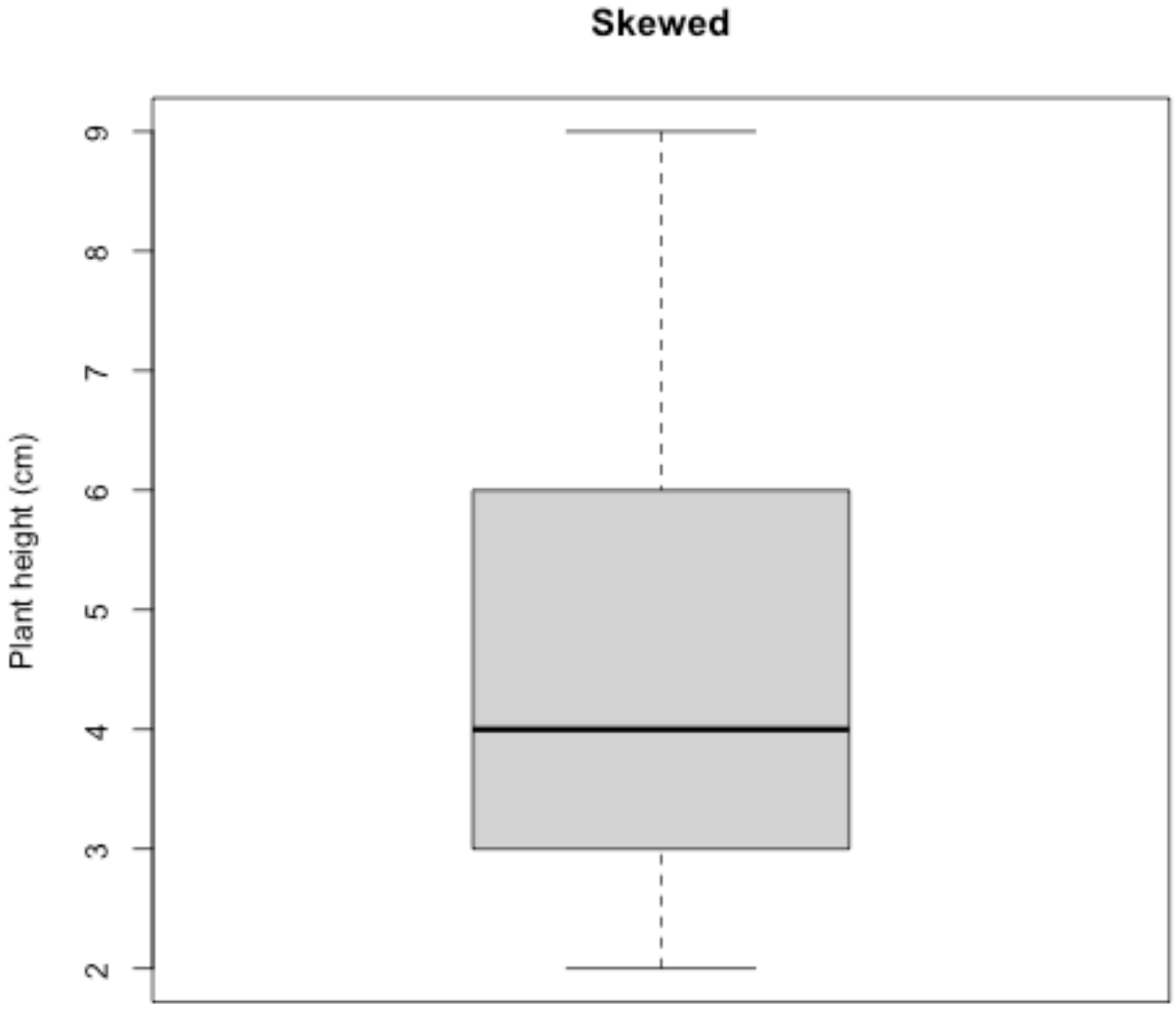


Figure 3.4: Histogram skewed plant height

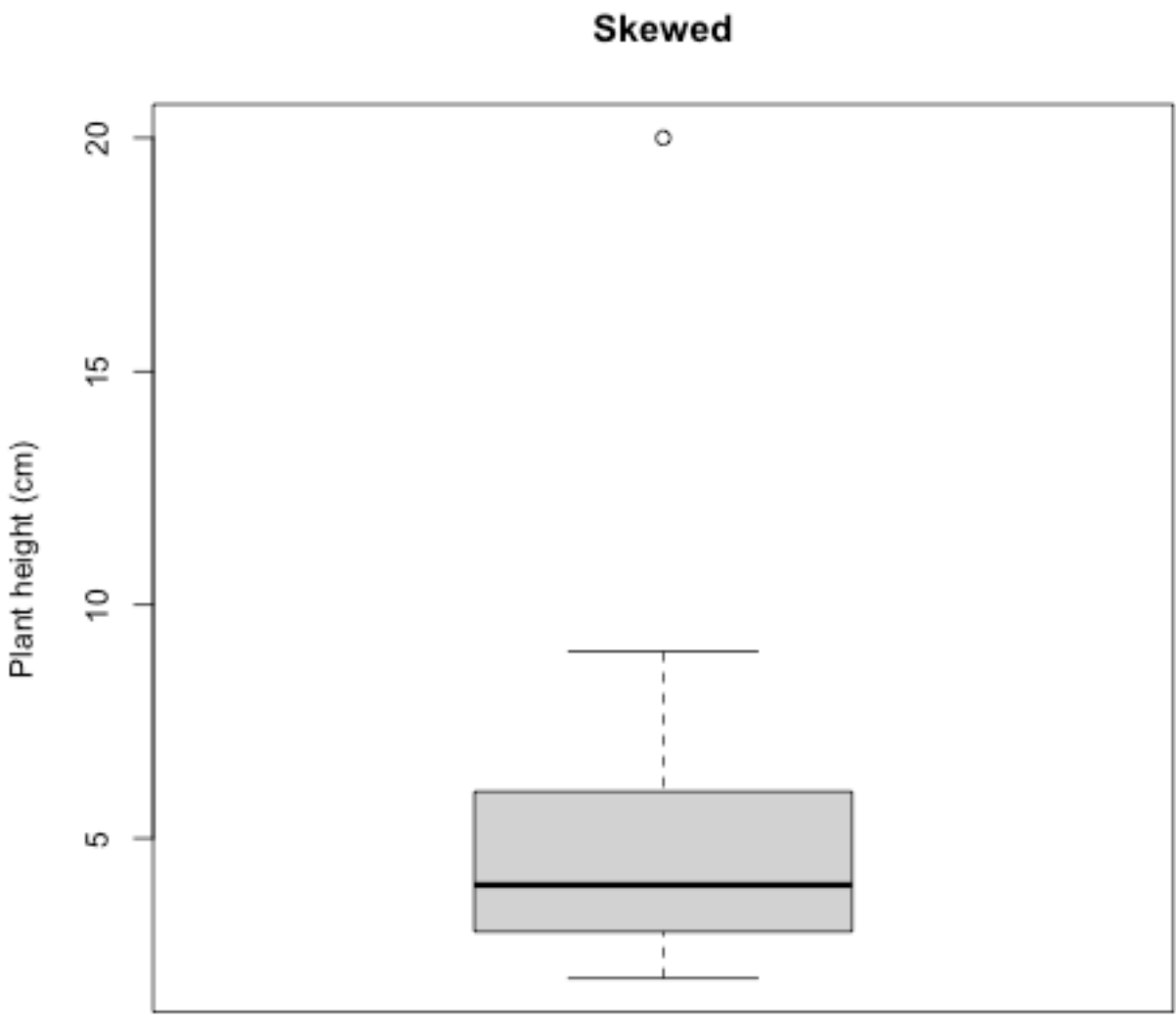


Figure 3.5: Histogram skewed plant height

Here is an example figure legend of a hypothetical figure showing species richness as a function of temperature across ecosystems:

Figure 1. Species richness increases with temperature across ecosystems. Panel A shows the mean species richness across three ecosystems: grassland (green), forest (blue), and desert (orange) from 2015 to 2020. Panel B depicts the temperature variation in these ecosystems during the same period. Circles represent data from 2015, while triangles represent data from 2020. The x-axis shows mean annual temperature (°C), and the y-axis shows species richness (number of species per 1000 m^2). Error bars indicate ± 1 standard deviation. The dashed line represents the best-fit linear regression ($R^2 = 0.85$, $p < 0.05$), showing a positive correlation between temperature and species richness.

3.7 Summary

We've played around a lot with data, but what do you actually need to take away from this? Here are the key takeaways:

1. Understand data types (Categorical, Numerical discrete, Numerical continuous, Ordinal)
2. Know ways to describe numerical continuous data (Location, Spread, Shape, Outliers)..
3. Be able to calculate mean, median, and standard error.
4. Familiarize yourself with mathematical annotation.
5. Familiarize yourself with R code.
6. Be able to interpret a histogram and box-whisker plot.
7. Be able to construct an appropriate figure legend.

3.8 Assignment

Now, let's play around a little more with R!

Let's also try describing the petal lengths of 3 different plant species: Milkweed, Bluestar and Pectis using the dataset that you downloaded.

Please put your results in a document and submit to your TA in Canvas! Be sure to include:

1. Summary statistics for your flower species (mean, median, and se)
2. Plots that you create
3. Appropriate figure legends for each plot, including a description of the data that a box plot shows.

Chapter 4

Basic statistical testing

4.1 Hypothesis testing review

Ahhhh the scientific process: A researcher makes observations about the natural world, generates a hypothesis to test, tests the hypothesis, rejects or fails to reject the hypothesis, and reports these findings. A simple, yet glorious process that has led to incredible discoveries! Statistical hypothesis testing answers very simple questions, while scientists work in very complex knowledge environments. For beginning researchers, it is important to understand what statistics can tell us, and how this builds information to address amazing and very interesting research questions. Hypotheses allow us to articulate exactly what we are testing in statistics and to organize thoughts and analyses. In the data description module, we discussed that most statistical tests are designed to answer one of two questions: 1) Are there differences between these groups? and 2) Are there relationships between these variables? Let's connect these concepts directly to hypothesis testing.

4.2 Downloads for this module

Download this file to the folder you created for this lab:

- Download the R file.

As you read the tutorial, follow along in the R code. Then, use the code to finish your assignment.

4.3 Objectives

We will highlight the basic components of a statistical test with a very simple statistical 'tailed' test. Today, we will:

1. State a hypothesis
2. Calculate a test statistic
3. Determine the p-value
4. Interpret results

The **null hypothesis** (H_0) is a statement about a population parameter that would be interesting to reject. The null hypothesis typically asserts that there is no effect or relationship or that results will not deviate from established knowledge.

For instance:

- The mean height of giraffes in captivity and in the wild do NOT differ.
- The incidence of toenail fungus is the SAME in the control group and the group given anti-fungal medicine.
- There is NO relationship between sea grass height and the number of sea snails.
- The mean human body temperature is 98.6 degrees Fahrenheit.

Null hypotheses are paired with **alternative hypotheses** (H_A) that represent ALL other possibilities other than that stated in the null hypothesis.

For instance:

- The mean height of giraffes in captivity and in the wild differs.
- The incidence of toenail fungus is different in the control group and the group given anti-fungal medicine.
- There is a relationship between sea grass height and the number of sea snails.
- The mean human body temperature is not 98.6 degrees Fahrenheit.

Note that the null hypothesis is very **specific**, while the alternative hypothesis is **general**. Statistical tests are designed to either reject or fail to reject the **null** hypothesis.

4.4 Tailed tests

Tailed tests are very simple tests for comparing means or proportions, and are great for illustrating the basic components of a frequentist statistical analysis.

Let's walk through an example! Handedness is common in humans. Around 90% of humans preferentially use their right hand. You've been watching your cat, Geraldo, play with his toy mouse, and you notice that he preferentially uses his right paw to bat the mouse around. You start to wonder if cats display handedness like humans! You run around visiting cats and observing their paw usage and determine that 14 cats of the 18 you observe appear to be right-pawed, while only 4 preferentially use their left paw. Is this enough evidence to suggest that cats display 'handedness' or did this pattern just emerge by chance?

4.4.1 Generate hypotheses

What is the null hypothesis in this case? Remember it must be specific so that we can either reject or fail to reject the null!

H_0: Left and right-pawed cats are equally frequent in the population (i.e., Cats are not right or left-pawed).

Note that this null hypothesis is very specific. If we would describe this mathematically, we would say that we expect half the cats (9 / 18) to use their right paws and half to use their left paws. If we express this as a proportion, (p), we are testing whether $p = 0.5$. Very specific.

What is the alternative hypothesis?

H_A: Left and right-pawed cats are not equally frequent in the population.

Note that the alternative hypothesis is very broad and encompasses all other possibilities. Because, in theory, we could observe proportions below 0.5 (1/18 cats are right-handed) or above 0.5 (16/19 cats are right-handed), we refer to this as **two-tailed**.

In a two-tailed test, the alternative hypothesis includes parameters on both sides of the value specified by the null hypothesis.

Before we move on, can you think of an example of a 1-tailed test?

Don't look before guessing!

Here are some examples:

- Did you score better than the class average?
- Is the time to getting to the student union less than 10 minutes when you avoid driving through campus?

Note the difference with the 2-tailed test: we are only interested in values in one direction. In the first example, we are interested in whether your score is >80% (class average). In the second, we are interested in whether your drive time is <10 minutes (time to the student union driving through campus from your house).

How could you phrase the first example to be a 2-sided test?

Don't look before guessing!

It could be something like this: Was my score different than the class average? In this case, your score could be higher or lower.

4.4.2 Calculate a test statistic

We generated our hypotheses. Let's calculate a test statistic. What is a test statistic?

A test statistic is a quantity calculated from that data used in statistical analysis to evaluate the null hypothesis.

For this simple test, our test statistic will be 14, since this is the number of right-handed cats we observed. We want to ask whether observing 14 out of 18 cats using their right paw is truly different from the null (9 out of 18 cats using their right paw), or did this pattern occur by chance in our sample?

4.4.3 Determine the p-value

Frequentists statistics is based on the concept of statistical distributions. If we run many trials, we can determine the likelihood of certain events occurring by chance. We refer to the patterns of occurrence of trials as frequency distributions. Let's illustrate using the data above. A cat can either be right-handed or left-handed (in this case there are no ambidextrous cats). To determine the likelihood that our pattern arose by chance, we conduct numerous trials like a coin toss. We would randomly flip a coin 18 times and record the outcome of each trial; heads being right-pawed, and tails being left-pawed.

In class:

- Each student took a coin.
- Flipped the coin 18 times and recorded number of heads.
- Divided by total number of trials (in this case students) to derive relative frequency for each event.

If we would flip the coins many times, we probably generated something that looks like the figure generated by running the code below (run in your R code to generate the figure). This is referred to as a probability distribution and expresses the relative frequency of particular events occurring. Frequentist statistics derives its name from this probability distribution.

```
## pdf
##   2
```

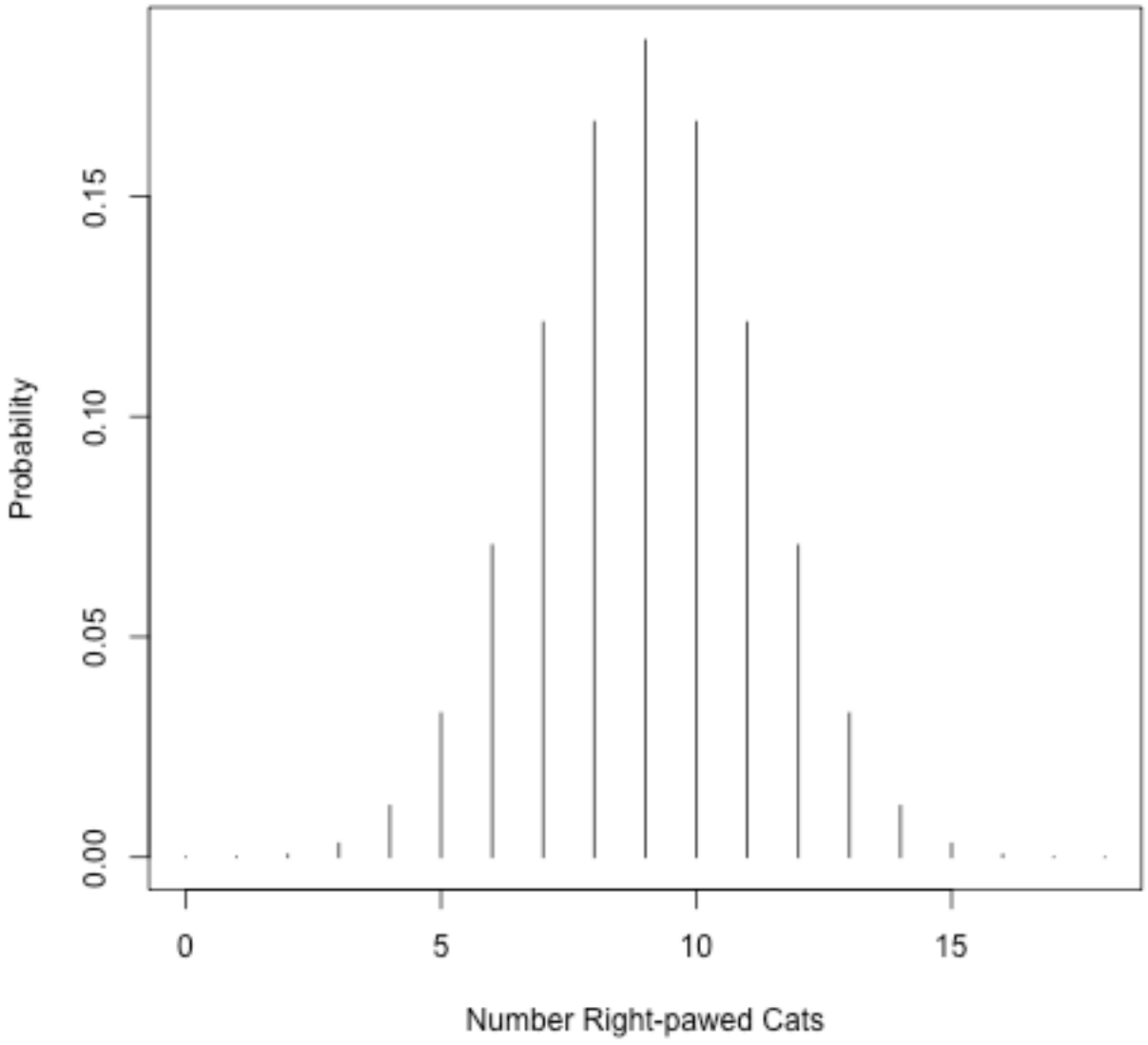


Figure 4.1: Binomial probability distribution

Here is a table of those probabilities:

```
probability <- dbinom(paws, 18, 0.5)
N <- 0:18
pawtable <- cbind(N, probability)
pawtableF <- as.data.frame(pawtable); pawtableF
```

```
##     N  probability
## 1   0 3.814697e-06
## 2   1 6.866455e-05
## 3   2 5.836487e-04
## 4   3 3.112793e-03
## 5   4 1.167297e-02
## 6   5 3.268433e-02
## 7   6 7.081604e-02
## 8   7 1.213989e-01
## 9   8 1.669235e-01
## 10  9 1.854706e-01
## 11 10 1.669235e-01
## 12 11 1.213989e-01
## 13 12 7.081604e-02
## 14 13 3.268433e-02
## 15 14 1.167297e-02
## 16 15 3.112793e-03
## 17 16 5.836487e-04
## 18 17 6.866455e-05
## 19 18 3.814697e-06
```

Take a look at the chart. What is the probability of observing right pawedness in 14 out of 18 cats? Is this difference from the null (9 cats are right-handed, no better than random) big enough to reject the null? To determine this we calculate a p-value.

A p-value is the probability of obtaining the data that we observe if the null hypothesis were true.

In this case, we will generate a p-value for a two-tailed test. To do this, we will add the probabilities of observing 14 right pawed cats or more by chance AND for the possibility of observing 4 or fewer right paws, which would indicate left pawedness. P-value = Pr[14] + Pr[15] + Pr[16] + Pr[17] + Pr[18] + Pr[4] + Pr[3] + Pr[2] + Pr[1] + Pr[0]

Recall that we can add these probabilities up, since, in our example, we say that being right or left pawed are mutually exclusive events.

```
pvalue <- (0.0117 + 0.0031 + 0.0006 + 0.00007 + 0.000004)*2; pvalue
```

```
## [1] 0.030948
```

We generate a P-value of 0.031.

In most sciences, we have agreed on a threshold of 0.05 for establishing statistical significance. If p-values are less than or equal to 0.05, we reject the null hypothesis. If larger, we fail to reject.

The significance level, α, is a probability used as a criterion for rejecting the null hypothesis.

This significance level is important. In the sciences, we would rather err on the side of *not* identifying a pattern, rather than saying there is a pattern, when there it doesn't actually exist. This concept is included in the in the discussion of *statistical errors**.

A Type I error is when you reject a true null hypothesis. You are saying there is a difference, when actually, if could perfectly measure your focal population, there is no difference. By establishing a significance level (α) of 0.05, we are saying that we are willing to accept that 5% of the time, we will say there is an effect when there isn't one.

A Type II error is failing to find a pattern or a difference when there actually is one. If you reduce your α to reduce your likelihood of making a Type I, you increase the likelihood of committing a Type II error.

The probability of committing a Type II error is more challenging to quantify and is related to the concept of **statistical power**. A study with high **power** has a low likelihood of committing a Type II error. Statistical power depends on several things, including **sample size**, the magnitude of the **effect** of the treatment, and **variation** within the sample. A study with a LARGE sample size, a BIG treatment effect, and SMALL variation within samples will have high statistical power. We will talk about calculating statistical power later.

4.5 Drawing conclusions from statistics

In this case, we **reject** the null hypothesis, and state that our results **support** the alternative hypothesis that there is handedness in cats. Note that we 'support' the alternative hypothesis, rather than saying that there is pawedness in cats or accepting the alternative hypothesis. Statements about statistics are phrased to reflect that we are dealing in probabilities, and there is always a chance that our findings are incorrect. Additionally, statistical tests specifically test the null hypothesis, not the alternative (for which there are often many possibilities).

What would we say if we didn't reject the null? We would state that we **failed to reject** the null hypothesis. Failing to reject the null indicates that our sample did not provide sufficient evidence to conclude that the effect exists, but lack of evidence doesn't prove that the effect does not exist. For this reason, we never accept the null.

4.6 Reporting results

When we report findings, we will provide:

1. A statement of findings
2. The test statistic
3. The P-value
4. A description of differences, if differences exist
5. A visualization of differences, if differences exist

Here is how we might report the written results of our previous test:

Cats displayed higher levels of handedness than expected by chance (t = 14, p = 0.03). Around 78% of cats preferentially use their right paw (**Fig. 1**).

Note that it is common to include up to 2 decimal places in the results statements.If p-values are less than 0.01, it is common to report the p-value as $p < 0.01$.

4.7 Assignment

We will now apply the tailed-test to a new question! You think that your dog, Rupert, prefers blue to red. You want to know: Do dogs prefer the color blue?

Provide the following information in a document and turn into your TA:

1. What is your null hypothesis?
2. What is your alternative hypothesis?

You invite all your friends with dogs, place both red and blue balls on the ground, and see if how many times the dogs select the blue ball out of a series of 10 trials. Ten of friends with dogs participate and you record the number of times that the dogs select the blue ball. Run the t-test in R!

3. Provide a corrected crafted results statement describing the outcome of the test. Please correctly structure your results statements, and make sure that you have included the 5 components mentioned above.

4. For your figure, please be sure to properly craft a legend. Recall from last lesson, the key parts of the figure legend:
5. A description of what the figure is showing you.
6. For complex figures with multiple panels, an orientation to the structure of the figure
7. Explanation of any symbols, colors, and/or lines
8. Description of how variance is quantified
9. Definitions of axes or units, if unclear
10. Acknowledgment of data source, if data source requires attribution
11. Implications of the figure (optional)

Chapter 5

Selecting statistical tests

5.1 A foray into statistics

We've talked about data types and describing data. This week we are applying concepts that we learned last week to select (and run) a statistical analysis. When conducting an analysis, a dataset will contain one or more **response variables** (aka y-variables, dependent variables) and one or more **explanatory variables** (aka x-variables, independent variables). Explanatory variables are used to explain variation in response variables. The data types of the explanatory and response variables determine which statistical test we use.

Pick the explanatory and response variable

Imagine that we want to identify areas that support high numbers of plants from the genus *Mitella*. We hypothesize that *Mitella* occurrence is positively related to water supply (i.e., the number of *Mitella* plants goes up as water availability increases).

Answer

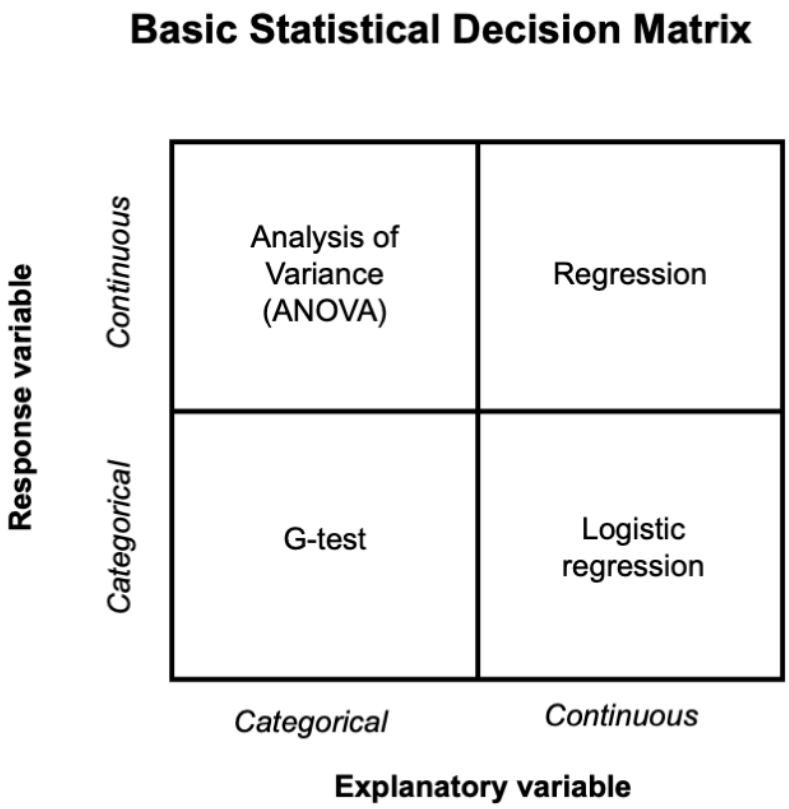
1. **Response Variable (Y)**: This is the variable we are trying to explain or predict. In the case of *Mitella*, it could be the number of *Mitella plants* in a specific area.
2. Explanatory Variable (X): This is the variable that we believe influences or explains the variation in the response. In your example, this would be water availability.

5.2 Downloads for this class

- Download the R file

5.3 Selecting statistical analyses

Below, you will see a very simple decision matrix for selecting statistical analyses. Let's walk through how to use it to select statistical tests.



5.3.1 Using the decision matrix

Both variables are categorical: Use a G-test (or chi-square test) to see if the frequency of categorical outcomes differs across categories of the explanatory variable. **Example**: If we were studying whether *Mitella* plants are found more frequently in shaded vs. sunny areas (categorical), we would use a G-test.

Explanatory variable is categorical, response variable is continuous: Use a t-test (or ANOVA for more than two categories). **Example**: If water availability is grouped into categories (e.g., "low," "medium," "high") and you

want to test its effect on *Mitella* abundance (continuous response variable), a t-test or ANOVA would be appropriate.

Note: ANOVA or ANalysis Of VAriance is just like a t-test except you are comparing more than 2 groups, in this example we would use an ANOVA for comparing "low," "medium," "high" watering regimes, but a t-test for testing "watered" vs "not watered".

Explanatory variable is continuous, response variable is continuous: Use linear regression. **Example**: If you measured the amount of water in milliliters and counted the number of Mitella plants, linear regression could help determine the relationship between water availability and Mitella abundance.

Explanatory variable is continuous, response variable is categorical: Use logistic regression. **Example**: If you're modeling the presence or absence of Mitella plants (categorical response) based on continuous water availability, logistic regression is the correct choice.

5.4 Statistical tests

Let's practice these statistical tests in R!

5.4.1 G-test (Categorical Response and Categorical Explanatory Variable)

Scenario: You are studying bird nest preferences in a forest. You want to know if birds prefer nesting in different tree species (oak, pine, or maple).

- `Response variable: Nest presence (Yes/No)`
- `Explanatory variable: Tree species (oak, pine, maple)`

Run the code in R.

5.4.2 t-test and ANOVA (Continuous Response and Categorical Explanatory Variable)

5.4.2.1 t-test

Scenario: You want to compare the weight of two species of frogs (Species A and Species B) to see if there's a significant difference in weight between them.

- `Response variable: Frog weight (grams)`
- `Explanatory variable: Species (A or B)`

5.4.2.2 ANOVA (Continuous Response and Categorical Explanatory Variable with More Than Two Categories)

Scenario: You are studying the growth of plants in three different habitats: Desert, Forest, and Wetland. You want to compare plant height across these habitats.

- `Response variable: Plant height (cm)`
- `Explanatory variable: Habitat (Desert, Forest, Wetland)`

5.4.3 Linear Regression (Continuous Response and Continuous Explanatory Variable)

Scenario: You are studying how the number of fish in a river changes with water temperature. You want to model the relationship between water temperature (°C) and fish count.

- `Response variable: Fish count`
- `Explanatory variable: Water temperature (°C)`

5.4.4 5. Logistic Regression (Categorical Response and Continuous Explanatory Variable)

Scenario: You want to study the probability of a specific bird species' presence in different areas based on elevation. The response is whether the bird is present or absent.

- `Response variable: Bird presence (Yes/No)`
- `Explanatory variable: Elevation (meters)`

5.5 Assignment

Here are the descriptions of 4 datasets (they are loaded in your R file). For each one, select the correct statistical test, then modify the code from your examples and run the test.

5.5.1 Turn-in

1. Results statements including:

- The correct test statistics

2. Associated figures with correct figure legends.

Chapter 6

Ecological sampling

6.1 Background

Ecology is the study of organisms and their relationship to the environment. With infinite time and capacity, we could measure every organism, every characteristic of the environment, every physiological function that affects the way an organism responds to the environment, and every gene that underlies those physiological functions in order to understand ecological systems. In practice, such detailed measurements are time-consuming and impractical. For that reason, we use statistics to account for the fact we are always missing information when we conduct ecological studies.

In statistics, a **population** refers to the all units of the thing that you are interested in (i.e., all Suriname frogs, all species in a marshland, all grains of sand, all aspen leaves from a genotype found in southern Arizona). Note that the term 'population' in statistics differs from the term population in population ecology, where a population refers to a group of individuals in a particular area that interbreed. Statistics accounts for the fact that we never perfectly measure the 'true population' or the all units of interest. Luckily, by properly applying statistics, we can learn practically anything about almost any population using **samples**!

A **sample** is a subset of the population that we measure to infer something about the true population. In order to avoid erroneous conclusions about the population, our sample must be **representative** of the population of interest and **unbiased**. As an example, imagine that you were interested in whether coat color in cats differed between house cats and feral cats. To select the house cat sample, you randomly select house numbers, visit the house and record coat color, thus collecting a random sample. However, to survey feral cats, you go to several cat colonies at night and record the first cat that you see, which are always white or tan. The sampling strategy for feral cats introduces bias,

because darker cats are harder to see at night. This causes you to overestimate the number of light-coated feral cats, and underestimate dark-coated feral cats, resulting in the erroneous conclusion that a greater proportion of feral cats are light-colored compared to house cats. Experiments must be carefully planned to reduce bias.

We can conduct statistical analysis until the cats come home (ha!), but if your sample is biased, our results will always be meaningless. In the cat example, it was pretty obvious that the researcher was introducing bias, BUT it is REALLY easy to introduce bias in ecological and social research on accident! Imagine that you looking at fire effects on vegetative communities in the Sonoran. In high severity burn areas, there are thickets of cat's claw (a pokey plant). Without proper field sampling protocols, it is very tempting to avoid establishing plots in the cat claw thickets, thus not capturing true differences in vegetation along burn severity gradients. Let's talk about several types of appropriate sampling strategies.

6.2 Sampling approaches

6.2.1 Random sampling

In order to reduce bias, researchers **randomize** sampling. **Random sampling** is when every item within the focal population has an equal chance of being selected. In research, random sampling can be applied to selecting experimental

subjects, assigning individuals to treatments, or identifying plot locations. It is **REALLY** easy to introduce bias in ecological and social research on accident if you do not use a random sampling technique! Imagine that you are looking at fire effects on vegetative communities in the Sonoran. In high severity burn areas, there are thickets of cat's claw (a pokey plant). Without proper field sampling protocols, it is very tempting to avoid establishing plots in the cat claw thickets, thus not capturing true differences in vegetation along burn severity gradients. In practice, researchers use number generators, like those on your phone, or within computer programs, like ArcGIS, to randomly place sampling points. Here, we've included a random number sheet to use to randomly array plots. A random number sheet contains random numbers that someone generated in advance to assist in the field.

We can quickly and easily generate such a sample in R, using the sample function.

```
sample(1:100, 10, replace=FALSE)
```

```
##  [1] 46 54 35 94 79 24 87  7 96 23
```

```
#1:10000 = numbers to chose among
#number of random numbers you wish to generate
#to replace or not (in other words do you wish for the same number to be selected multiple times)
```

6.2.2 Stratefied random sampling

To make a sample representative of the population, you will want to capture the typical state of the population of interest. This is challenging, since prior to collecting data, you do not know the typical state of the population. With an understanding of ecology, however, and precisely describing your research question, you can improve the representation of your sample without a lot of specific a priori (beforehand) knowledge of the target population. One typical approach is referred to as **stratified random sampling**, in which you ensure that you are proportionately sampling from major habitat types or features. In the example in **Fig. 1**, a random sample of the study area overrepresents the forested habitat relative to the grassland habitat. To account for this, the researchers adjust the sampling technique, such that plot locations occur in both of the major habitats proportionally. Since grasslands make up 55% of the study area, 55% of the points would be randomly located in the grassland area. Since there are twenty plots, 11 are placed within grasslands (0.55*20). The remaining 9 plots are then randomly allotted to the forested habitat.

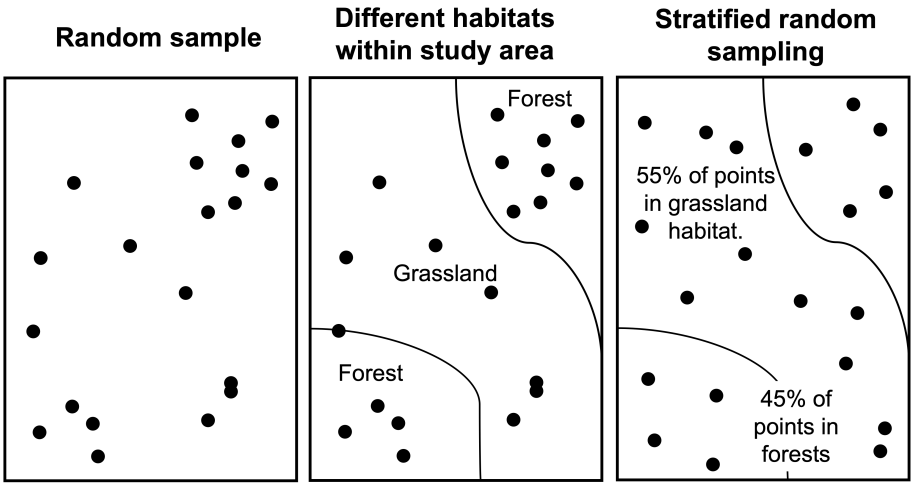


Figure 1. Random versus stratified sampling.

6.2.3 Gridded random sampling

In complex, multi-species systems, another approach to improve coverage of random sampling is to randomly place plots within a grid (**Fig. 2**). This is to ensure that you capture species, which may have an array of distributions. **Distribution** in plant ecology refers to the spatial arrangement of a species or organisms across the landscape. Depending on system dynamics, species may be dispersed, randomly arrayed, or clumped, thus a gridded approach can help capture species, no matter their spatial orientation (**Fig. 2**).

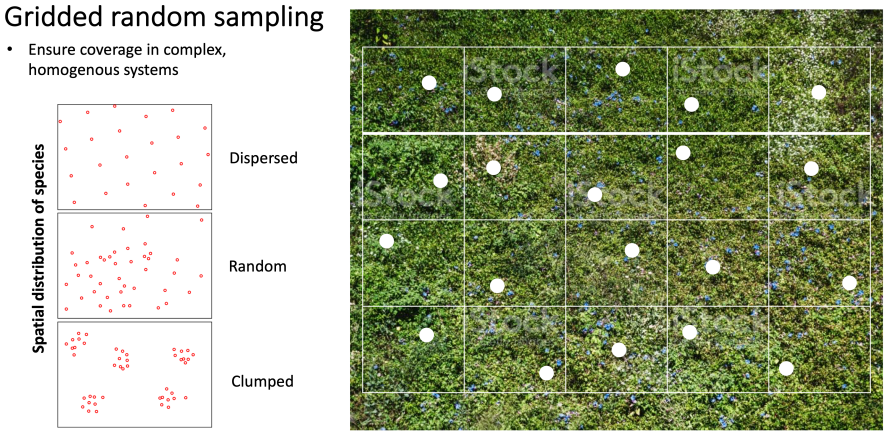


Figure 2. Randomly placing plots within a gridded region helps maximize

the likelihood to capture species dynamics in complex, multi-species systems, composed of species with a variety of spatial distributions.

6.2.4 Random cluster sampling

Random cluster sampling randomly select groups (aka clusters) within a population. This sampling design is used commonly in ecology, when we select random locations for plots, then measure all individuals within those plots. If for instance, we are interested in Ponderosa Pine growth rates on the Coconino National Forest, we would randomly assign points across Pondo habitat on the Coconino. At each point, we would set up a plot in which we measure Ponderosa Pines within an 11.71m radius plot. **Why wouldn't we just go out to a point and measure 1 tree to create a totally random sample?** The plots are randomly assigned (yay!), but the trees within the plots are not **independent**. In other words, we might expect measures of trees within plot A to be more similar to each other than they are to trees within plot B, due to differences in microsite characteristics, genetic similarity among co-occurring trees, or site history (logging, fire). Luckily, we can account for this non-independence, as long as the plots are random!

6.3 Plot shapes, sizes, and transects

Deciding on the shape and size of your sampling unit depends on the species or feature that you are trying to measure. **Plots** can be *ANY* shape, but usually the shape of the plot is either a square or circle for simplicity – why would you sample using a hexagon? Often in forestry, plots are circular, marked with a central post, which foresters attach logging tape and rapidly measure trees that fall within a certain radius from the central post (**Fig. 3**). Since trees are large, this helps foresters quickly collect data on the species composition and structure of forests. For smaller organisms, like understory species, **quadrats** (small plots, often square and 1 m2 in size) are often used, since many smaller organisms occur in this area – if plots were too large, then data collection would be too time-consuming. In some cases, researchers are interested in how certain ecological variables differ as a function of distance from a feature. In these cases, researchers will often use a **transect** – a linear feature – and collect variables of interest along it. Any of these sampling shapes and sizes can be combined or adapted to measure ecological features to answer the question of interest.

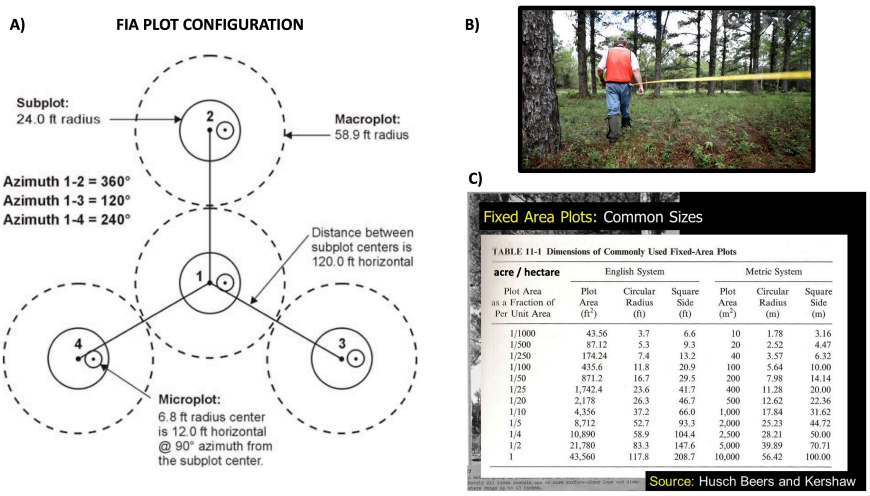


Figure 3. A) The standard plot configuration for the Forest Inventory and Analysis dataset, which includes data on all of our forested lands in the US. Forestry plots are often circular (A), allowing foresters to attach loggers tape to a central plot marker (B) and quickly measure trees within these fairly large plots (C). Plots must be large in order to include enough trees to describe stand characteristics.

6.4 Reducing sampling error

What is **sampling error**? Sampling error is the difference between the true estimate of a population and the measurements that researchers collect on a sample. Error happens by chance and is unavoidable – it can be thought of as noise within the data. Error is different from bias, because it is non-systematic. For instance, imagine two people are measuring cactus heights for a demographic study. Error in height measurement is introduced by many things – the shakiness of each person's hands, the amount of degradation and stretch in various measuring tapes. **Bias**, on the other hand, would be introduced if person 1 only measures the small cacti, or always mis-reading the measuring tape and measuring heights 5 cm less than their actual height. Bias should always be avoided, and error reduced as much as possible. Larger samples are less affected by chance and so will have lower bias. In ecology, we refer to the number of independent units being measured as **replicates**. The more replicates, the less sampling error!

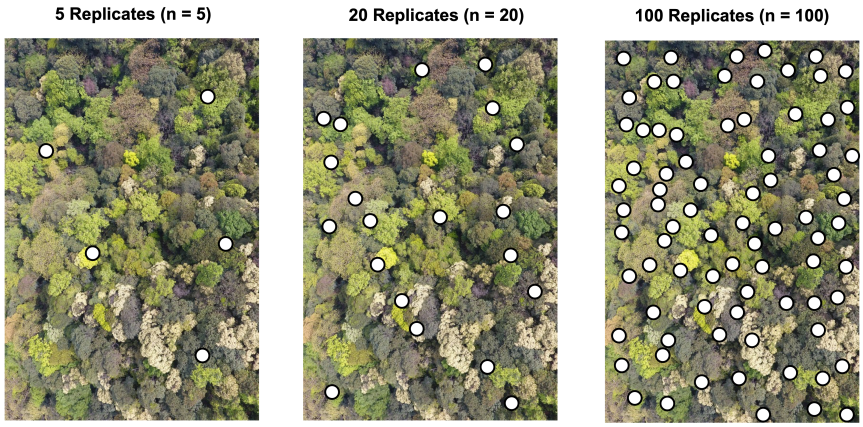


Figure 4. As the number of replicate plots increases so does the accuracy at which we can estimate parameters for this forest stand.

6.5 Variables of interest

Finally, depending on the research question, there are a number of different variables that you might want to measure. In ecology, you may want to measure the number of different species in an area to look at diversity patterns, or collect data on size or growth to look at performance, or monitor individuals after a disturbance to look at mortality. We will collect different forms of data throughout the semester, but the following principles will always apply:

1. Samples should be random and representative
2. Sampling methodology – plot shape and size – should reflect the organism or ecological feature that you are measuring
3. More replication is better, since it reduces sampling error

6.6 Assignment

6.6.1 Downloads for this class

Please download for class:

- Random number table for compass bearings (optional)
- Land propery maps
- City plan

- Description of Sinclair Wash This information can help you develop your sampling plan.

6.6.2 Sampling Q & A

Answer the following questions.

1. You are working in an agroforestry system attempting to understand where a species of finch occurs across this landscape. The habitat is patchy, with forests comprising 80% of the landscape and fields covering the remaining 20%. You decide to use random stratified sampling to capture finch occupation in both systems proportionally. Given this, how many of your 40 plots will occur in forested habitat? Select one:
2. 20
3. 30
4. 32
5. 38
6. You are interested in how the abundance of frogs varies across habitats moving away from a local stream. In order to look at frog abundance as a function of distance from stream, you:
7. Place 1m2 quadrats along the stream bank.
8. Create a grid along stream habitat and randomly assign plots to occur within the grid cells.
9. Place a 50 m transect running perpendicular to the stream, starting at the stream bank and moving away.
10. Place a 50 m transect along the stream parallel to the running water.
11. Which of the following sampling schemes will have the lowest sampling error? Select one:
12. plot per hectare
13. 5 plots per hectare
14. 10 plots per hectare
15. Unable to determine using this information

Review: Last week you learned about different data types, list the data type for each variable below (continuous, nominal):

1. Growth

2. Survival (yes, no)
3. Foliar nitrogen content

6.6.3 Sampling activity

During this semester, we are going to monitor the impacts of the **Rio de Flag Flood Control Project** on the recovery of Sinclair Wash (also called Clay Avenue Wash) to be compiled over time and provided to the City of Flagstaff. This project has the dual benefit of restoring important ecological habitat along the Rio de Flag, as well as reducing flood risk in historically minority neighborhoods in Flagstaff. The project is a collaboration between the City of Flagstaff and the US Army Corps of Engineers and is projected to take 20 years to complete and cost $122 million dollars.

There are several nice resources to learn about this project, including this resource page. Please watch the following informational video on the project (feel free to watch in groups): A Southside Story. Also, please visit the city project page.

Once you have looked over this information and the downloads provided, if the weather allows, walk out to Sinclair Wash and orient yourself to the site. Walking directions. If the weather is poor, check out what the wash looks like on google and use this view to create your plan. Navigate to the Aerial view of Sinclair Wash, and zoom in until you can see the Rio, which follows the tree line and path (path is hard to see here). Create your plan - you can even import a screenshot into powerpoint or use other applications to map out your sampling strategy.

Write out a brief description of a sampling strategy for determining how restoration is affecting revegetation along eroded banks along Sinclair Wash. Be sure to mention how you would reduce bias, collect a representative sample, and reduce error. Please also describe plot shapes and sizes. Then, draw or make a map of your sampling strategy.

6.6.4 Deliverables

Please turn in:

1. The answers the the questions.
2. A written description of your sampling strategy
3. A picture / drawing of your sampling strategy

Chapter 7

Natural selection

7.1 Background

The living world is rich with diversity. This diversity is often examined at the level of species, with over 1.5 million species described and estimates of the total number of species on Earth ranging from 5 to 100 million (estimates vary greatly; see figure below). Yet diversity also exists within species, populations, social units, and family lineages. Variability in the features of organisms that we see today is put into even greater perspective when one considers the vast amount of diversity that occurred in previous generations. Only a fraction of this remains recorded in the fossil record.

Fossils suggest that the number of species has increased over time, although some relatively brief periods of decrease are also recorded. Records also suggest that characteristics have become more complex over time. There are numerous hypotheses that could explain the origin and diversification of species. Some attribute it to divine intervention, some to extraterrestrial influence, and still others to simple physical and chemical processes that occurred in our primordial backyards.

Regardless of what caused the origin and initial diversification of life, ecologists are generally curious about the sources of diversity within modern populations, and knowing what causes this diversity to be maintained, increased, or eliminated from populations. Indeed, the entire field of conservation biology is devoted to understanding and preserving natural variation within and among groups of organisms.

Key Terms: (you'll want to be sure to use these in your lab report!)

Natural selection: the process by which organisms better adapted to their environment are able to survive for longer and produce more offspring.

Fitness: an organism's relative likelihood of surviving long enough to pass on its genes. In other words, those who contribute most to the next generation exhibit the highest fitness.

Selective environment: an organism's environment plays a role in selecting favorable traits, and traits that get selected for may be different depending on the nature of each environment.

Selective agent: a biological or physical agent that imposes selection, and thereby determines which individuals pass on genes and which do not.

7.2 Downloads for this class

- Download the CSV file
- Download the R file

7.3 Objectives

To understand how natural selection operates and characteristics within and among populations change over time, and to answer the following research question (especially in your lab report):

7.4 Research question:

Does the selective environment affect the composition of future generations?

7.5 Methods

Materials

Each group of three should have: 1 piece of patterned fabric 7 different colors of dots (approximately 80 of each color, kept in separate piles) Green bowl (for discarding dots) Note-taking materials

Procedure

1. With your group members, read through the entire procedure. Based on what you now know about the experiment, formulate a hypothesis. Write it down for inclusion in your lab report.

2. Each group of three to four should select a fabric pattern of its choice (it's OK if two or more groups use the same pattern) and get *seven* piles of differently colored paper dots (keep colors separate; colors include black, hot pink, green, pink/peach, dark yellow, white, blue).

3. Count out 12 dots of each color, and mix them in a single dish.

4. Spread the fabric flat on a table and randomly sprinkle the 84 dots onto the fabric. Be sure to disperse the dots all over the fabric, not just in a small group on one part of the fabric.

5. On the spreadsheet provided here, record the frequency of each type of colored dot.

6. Imagine each group member is a red-tailed hawk, and that each colored dot is a mouse with a particular fur color. Each mouse runs free in its respective environment, while each hawk preys on the mice. Hawks can only prey on one mouse at a time. A hawk must "fly" the mouse back to its nest (another culture dish) each time it captures one. The hawks must hunt quickly – you will only have 45 seconds to remove 56 mice!

7. In the first year of predation, the hawks remove 56 mice from the population (thus, 28 mice remain in the population). Group members will remove the first 56 dots they see in 45 seconds. If you need more than 45 seconds, keep going until you remove 56, just try to go as quick as you can while following the instructions!

8. Remove the 28 survivors from the fabric as well. Each of the survivors "overwinters" and has two offspring of the same color. Count out these new mice and incorporate them into the population. [If this requires hole-punching more of a particular color, feel free to do so.] After they've been incorporated, record the frequency in Table 1 of each fur color in the population (note: the population size after the new mice are incorporated should again be 84, but the frequencies of the fur colors might be different than those recorded in Step 4 of this investigation). Again, randomly sprinkle the 84 dots onto the fabric.

9. Repeat Steps 6 -8 three times, recording the frequency of fur colors for each generation. After the third iteration of hawks preying on the population, count the remaining mice (those that would make up the fourth generation) and record in Table 1. Next to your first graph, make a bar graph showing the frequency of each type of colored dot in the final generation. Copy the two graphs into your lab report.

10. Compare the frequencies of fur colors (dot colors) in the first generation to those in the fourth generation.

11. If the frequencies of fur color changed in the population over time, then within your group and with everyday language describe the phenomenon that was responsible for the change.

12. Graph your results using the R file that you downloaded

7.6 What to turn in

Lab Report Due in 1 week on Canvas

Your final product will be a lab report containing only research question, hypothesis, methods, and results (Format 1) as described in the "Written Lab Report Guidelines" here: Lab report guidelines

Only use the data from your group of hawks and habitat. Compare the distribution of mice fur colors between the first, second, third and fourth generation to answer the research question and address how your results aligned with your hypothesis. Describe your data collection methods (outlined above) and your results. **You must include one table and one set of figures based on today's results.**

Next week, you will turn in a completed lab report to your instructor through Canvas in which you have made your best effort to write up those components of a lab report. Then, your instructor will provide you with feedback by the following week's class.

Chapter 8

Population ecology

8.1 Lab set-up (week 1)

In a few weeks, we will complete a lab on the population dynamics of duckweed (*Lemna minor*) in microcosms. Today, we will set up your experiment. Next week, you will collect data on your populations and maintain your experiment.

8.1.1 Downloads for this lab

- Download the CSV file
- Download the R file

8.1.2 Objectives

To investigate population dynamics of duckweed using microcosms.

8.1.3 Materials

1. Microcosms (i.e., clear plastic containers)
2. *Lemna minor* individuals
3. Water
4. Small labels or markers
5. Light source (natural sunlight or artificial light)
6. Data recording sheets

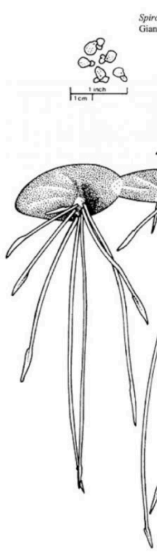
First, we will explore growth rates of populations with two different initial population sizes. Working in groups of 3 – 5, follow the procedure below. Then, proceed to setting up your experiment in which nutrient levels are varied!

8.1.4 Procedure

1. Fill two microcosms with artificial pond water, 200 ml. Mark the 200 ml water level on the cup so that you can refresh the culture solution to the same volume.
2. Place two healthy Lemna plants in one of the cups. Place 15 health Lemna plants in the second cup.
3. Because a plant can consist of one or more thalli, it is necessary that you now count the number of thalli in each cup. One thallus is any leaf unit that is over 1.5 mm. Record these data in the Day 0 column of Table 1.
4. Place the cups under fluorescent or near natural lights for a period of two weeks, check them periodically to refill the cups to the 200 ml line.
5. Count the number of thalli in each cup on Day 7 and Day 14. These counts are already reflected in the datasheet for this lab, but you will need to record population size going forward

Here is some information on *L. minor* to assist you!

Lemna minor, commonly known as duckweed, is a small aquatic plant that belongs to the Lemnaceae family. *Lemna minor* is typically found in quiet, freshwater environments such as ponds, lakes, slow-moving streams, and marshes. It prefers still or slow-flowing water bodies where it can float on the surface. *Lemna minor* plays several ecological roles in aquatic ecosystems. It provides habitat and food for various microorganisms, insects, and small aquatic animals. The dense mats of duckweed also offer shade, affecting water temperature and reducing the penetration of light into the water. This can influence the growth of other aquatic plants beneath the duckweed mat. The presence and abundance of *Lemna minor* can serve as an indicator of water quality. Its sensitivity to nutrient levels and environmental conditions makes it a useful tool for assessing the health of aquatic ecosystems. In some cases, excessive growth of duckweed may signal water quality issues, such as nutrient pollution.



8.2 Maintaining your microcosms and measuring population growth (week 2)

1. Check the microcosms and refill the cups to the 200 ml line.
2. Count the number of thalli in each cup on Day 7 and Day 14. Record these data in the population size column of your datasheet.

8.3 Final data collection and analysis (week 3)

In this lab, we will collect data from your microcosms and analyze your results. First, we will learn a little more about microcosm studies in ecology, your model organism, and some population ecology background.

8.3.1 Microcosm studies of population dynamics

Microcosms in ecological sciences are small-scale experimental systems that replicate natural ecosystems. Researchers use microcosms to study ecological interactions, nutrient cycling, and other ecological processes in a controlled environment. Here we will use microcosms to examine population dynamics of duckweed (*Lemna minor*) to gain insights into core concepts in population ecological, such as population growth and carrying capacity.

In class, we discussed how populations have the capacity to grow **exponentially**, but resource availability eventually limits population growth. Light, space and nutrients are all examples of resources that may be limited within ecosystems and constrain population growth. Here, we will explore populations growth using the model species, *L. minor*. A **model organism** is a species used in scientific research to represent a broader biological phenomenon, serving as a convenient and well-understood subject for studying fundamental biological processes.

Over the last few weeks, you have been establishing and maintaining microcosm experiments with duckweed (experiment modified from Population growth: Experimental models using duckweed (Lemna spp.); University of Toronto, Toronto CANADA).

Few plants are suitable for studying continuous population growth because most plants have life cycles with discrete jumps in population size, their reproduction is seasonal and they respond to changes in population density by changing size and shape instead of population number (Harper, 1977). However, free-floating aquatic plants such as duckweeds (*Lemna* spp) or water ferns (*Azolla* and *Salvinia*) undergo continuous growth and therefore are excellent models for quantifying aspects of population growth (Clatworthy and Harper, 1962; Harper, 1977).

These plants are stemless and have only one to four leaf-like structures called thalli (singular = thallus), if they are flowering plants, or fronds, if they are ferns. Roots from the thallus hang free in the water. Duckweeds can reproduce by flowering and setting seed (sexual reproduction) but seldom do. More commonly they reproduce asexually by producing a new thallus or frond directly from an old one. When a new thallus has grown large enough and has roots, it breaks loose from its parent plant and grows on its own as a separate plant. The growth of a population can be followed by counting thalli or measuring changes in biomass (dry weight).

If a pond or lab beaker is inoculated with one or two thalli and conditions are favorable, the plants commence exponential growth (Fig. 1, Phase I). The growth rate of the population under these conditions is density independent; the population grows unimpeded by resource limitation or competition. We can estimate the intrinsic rate of growth (r- see the equations on following pages) by measuring the uninhibited growth of low-density populations.

As thalli accumulate, the population becomes crowded and limited by the available resources. For a period, growth appears constant (Fig. 1, Phase II) as the width and thickness of the mat of floating plants increases. Eventually the beaker or pond fills with floating plants (Fig. 1, Phase III) and the population reaches a steady state (see the following equations). At this point, for every new thallus that appears, an existing one is shaded and dies, i.e., the population size is stable. The logistic growth curve (Fig. 1) illustrates all three Phases.

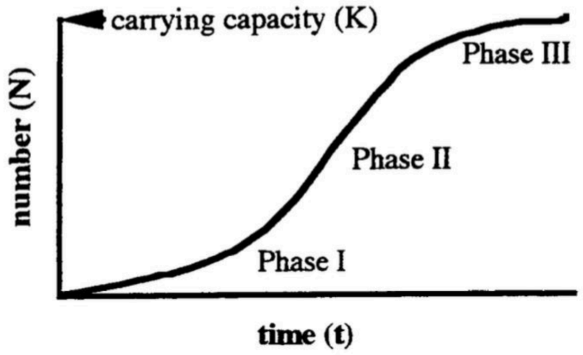


Figure 1. Number of individuals through time in a population with carrying capacity (K).

8.3.2 Population dynamics basics

Using your own or class data (if your experiment has failed), graph the average (mean) number of thalli (N) as a function of time for the cultures that started with two plants and the cultures that started with 15 plants.

The three equations shown below describe growth of populations:

Exponential population growth (expressed by the instantaneous rate of increase, r):

$$\frac{dN}{dt} = rN$$

where: - N is the population size, - r is the intrinsic growth rate, and - t is time.

When we are looking over a discrete period of time, we can calculate **Geometric population growth rate**, described by the equation:

$$N(t+1) = N(t)e^{rt}$$

where: - $N(t)$ is the population size at time t, - r is the intrinsic growth rate, and - t is time. - e is the base of the natural logarithm (constant)

The factor by which a population increases in one unit of time (ert) is the finite growth rate of the population ( ), from:

$$N(t+1) = N(t)e^{rt}$$

We take the natural logarithm of both sides:

$$\log N(t) = \log N(0) + rt$$

This equation now represents a linear relationship between $\log N(t)$ and t, where: - The slope of the line is r (the intrinsic growth rate), - The intercept is $\log N(0)$ (the log of the initial population size).

By plotting $\log N(t)$ vs. time t, the slope of the line provides a direct estimate of r.

When resources are finite, we can rearrange the exponential growth equation to include the carrying capacity of the environment. The logistic growth model, which accounts for a population's carrying capacity, is described by the equation:

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right)$$

where: - N is the population size, - r is the intrinsic growth rate, - K is the carrying capacity, and - $\frac{dN}{dt}$ is the rate of change of the population over time.

Estimating geometric growth

Let's start by plotting population growth through time.

Since population growth rate is exponential, we can plot it as log (N) over time. Plot log N as a function of t for the cultures that started with two plants. The slope of a line drawn through the mean log N at Day 0, Day 7, and Day 14 would approximate r. Make the same graph and calculations for the cultures that started with 15 plants.

Now, let's calculate the finite rate of increase, λ, using the equation:

$$\lambda = \left(\frac{N_{t+1}}{N_t}\right)^{\frac{1}{t}}$$

where: - N_{t+1} is the population size at the final time point (e.g., Day 14), - N_t is the population size at the initial time point (e.g., Day 0), and - t is the total time (e.g., 14 days).

As the population of Lemna in your cup grows, the rate of growth will slow down. When the population reaches the carrying capacity of the cup, the growth rate of the population will be 0 (dN/dt = 0). If you plot the geometric growth rate (, calculated above) for each cup, as a function of population size (Nt), you should have a linear plot where the y intercept (where N = 0) would approximate r and when = 1, n = K. We derived that information by rearrange your formula for carrying capacity:

Estimating r: - The intercept of the linear model (where $N = 0$) is an approximation of the intrinsic growth rate r.

Estimating K: - The carrying capacity K is estimated by solving the equation where $\lambda = 1$ (i.e., when the population growth rate reaches zero):

$$K = \frac{1 - \text{intercept}}{\text{slope}}$$

Plot your data. What is your estimated carrying capacity (K)?

8.4 Assignment

Please turn into your TA, your:

1. Estimates of r for both populations & associated figure
2. Estimates of finite population growth
3. Estimate of carrying capacity & associated figure
4. Summarize the findings and draw conclusions about the factors influencing population dynamics and discuss the implications of the study for understanding population ecology in natural ecosystems. Be sure to describe any differences in growth rates between the two microcosms and discuss why you might be seeing that pattern.

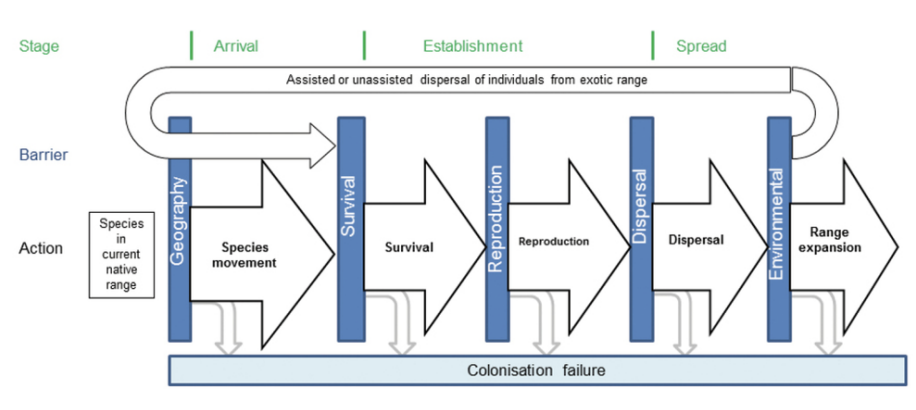
Chapter 9

Invasive species

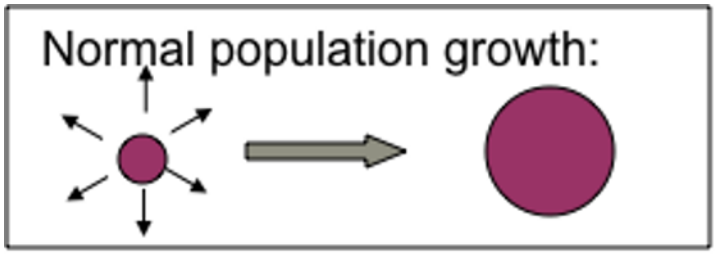
9.1 Introduction

In this lab, you will consider the distribution of a non-native species that has become widespread in our region. You will collect field vegetation data according to common methods in ecology and answer discussion questions after your data collection. We will be meeting near the South Commuter Lot / Disc Golf Area on campus for this lab. Download location information

Biological invasions are considered one of the primary threats to biodiversity and ecological functions today. Any species that has been introduced by humans (either accidentally or on purpose) to a location it didn't reach on its own is considered a non-native species. Other terms you might hear include exotic and alien. A non-native species that exerts a negative impact on native species is termed **invasive**. Not all non-native species become invasive: in fact, it is estimated that out of all species that are introduced, about 10% are able to establish reproducing populations in place, about 10% of those escape human-dominated areas and spread through natural areas, and about 10% of those become invasive. **Figure 1** illustrates the phases of invasion.



As non-native species establish and spread, their populations tend to follow predictable growth patterns. An initial group of individuals will reproduce and demonstrate population growth in the immediate area, slowly expanding the infestation as shown in **Figure 2**. Occasional long-distance dispersal events will establish new populations, each of which then demonstrates the same process and expands its occupied area.



Invasive species are known to exert a range of impacts, depending on the invading organism and on the characteristics of the invaded system. Worldwide, invasive species include vertebrates, invertebrates, and plants. Invasives have been known to outcompete native species, reducing native biodiversity in the invaded locale. Thick infestations of invasive plants, where no native plants are able to persist, are known as monocultures. Other invasives act as predators or herbivores, directly damaging or preying upon native species. Such effects have resulted in extinction (for example, predation by invasive cats and rodents has resulted in extinction of ground-nesting birds on islands). Other invasive species transform ecological conditions or functions: for example, invasions may alter soil pH or stability, water table accessibility, pollinator abundance, etc.

Agencies that manage forests and other natural areas must survey and monitor invasive plants within their jurisdictions. Monitoring information that is collected includes: total area infested, density or cover of the infestation, spread rate from year to year, environmental characteristics of inhabited sites. This information can help agencies predict further spread in the future. Other in-

formation that is often collected is focused on understanding the impacts of the invader, including the implications of the infestation for native species, soil characteristics, moisture availability, and ecological functions such as species interactions.

In this lab, we will look at where a non-native species of your choosing occurs, measure its density, and evaluate the presence of young individuals in order to develop some inferences about likely spread rates.

9.2 Downloads for this lab

- Download random number sheet
- Download the CSV file
- Download the R file

9.3 Objectives

To characterize the current infestation of your nonnative focal species near a recreation site between NAU's campus and a 4-lane interstate highway, and see how this infestation is driven by disturbance to the landscape.

9.4 Materials

Each group of 4 or 5 should have:

1. 1 Tape measure (50m)
2. 1 meter tape (or meter stick)
3. 16 Pin-flags
4. Printed lab with datasheets
5. Website or application for identifying invasives
    a. https://nazinvasiveplants.org/
    b. iNaturalist application (free download from your app store)

For a brief tutorial on making iNaturalist observations, please Download this document

9.5 Procedure

9.5.1 Hypothesis development

In this lab, we will be characterizing a nonnative species infestation at our study site and exploring where it occurs. Specifically, we are interested in the question: **Does disturbance increase the abundance of nonnative species?**

Look around the site. Considering the site characteristics you can observe, brainstorm with your group and write a hypothesis for how you think environmental conditions are affecting the distribution of your nonnative focal species in the area. **Record your hypothesis** - you will need it for you lab packet!

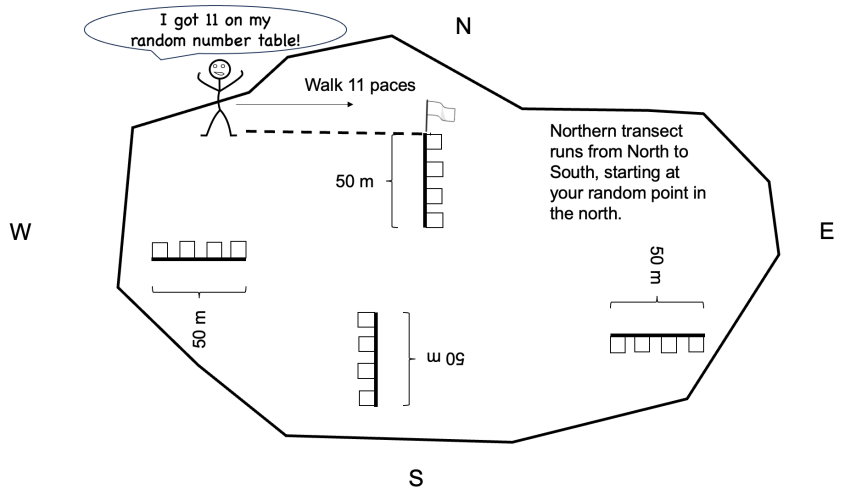
Now, what evidence will you need to accept or reject your hypothesis. Write an if/then statement to guide your work, again recording for your lab packet. Several things to know when you are writing your if/then statement, we will be measuring non-native species density as your response variable. You will be recording both evidence of disturbance (i.e., roads, trash, etc.) as a direct measure of disturbance, and tree cover as an indicator of undisturbed habitat.

9.5.2 Nonnative species density quadrats

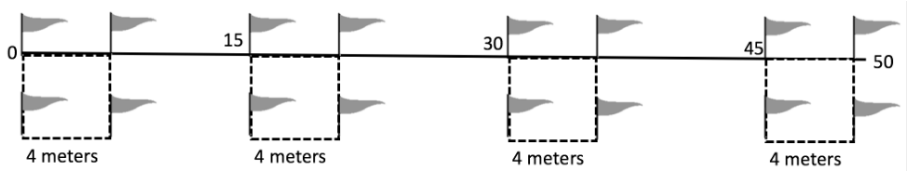
You will set your **quadrats** up along **transects**, or linear sampling features that you create with a tape measure or **transect tape**. Since you suspect that azimuth, or the direction, could influence nonnative species behavior, you decide to establish 4 transects, and place 4 quadrats along each transect.

You divide the field into general directions. Azimuth is measured clockwise from 0° (which corresponds to true north) up to 360°. - 0° or 360° = North - 90° = East - 180° = South - 270° = West

At each end, use this random number sheet or the random number generator on your phone (you will need to choose smaller numbers) to identify a starting point for your transect. As an example, go to the northern part of the field, place your finger anywhere along the random number table. Use only the first two numbers of the 5 digit number. This will indicate the number of paces you will take along the northern end of the field. Then run the transect from the starting point south (north to south; see **Fig. 3**).



Extend the transect tape 50 meters to the south, in a line roughly parallel to the fence on the west side of the study site. You will sample four 4x4m quadrats along this 50m transect: one starting at 0m, one starting at 15m, one starting at 30m, and one starting at 45m.



At each survey point, place a pin-flag there to mark the corner. Continue 4m further down the transect tape and place another pin-flag. Use your meter stick/tape to measure 4m to the east of each of those pin-flags and place another pin-flag at each resulting point, creating a 4x4m quadrat.

Within a quadrat, count the total number of your nonnative focal species plants and record that number. Record the total number of your nonnative focal species plants in the column 'NumberNonNativeSpecies' in this datasheet.

In addition, walk carefully throughout that 4x4m quadrat and look for evidence of disturbance (e.g., animal waste, trails, trash, scat, etc.). If there is disturbance in your plot, record 'Yes' if there is no disturbance record 'No' (be sure to capitalize) in this column, 'DisturbancePresent'. As evidence of more intact habitat (less disturbed), also record the number of trees (taller than the height of the tallest member of your group) that are shading the quadrat in the column, 'TreeNumber'. Repeat all of these methods until you have gathered data for all four 4x4m quadrats along all 4 transects (You will have collected data from 16 quadrats).

9.5.3 Analyses

Once you collect your data and enter it into your csv file, analyze your data using your R file. We are going to conduct a statistical test on your data using linear models. A linear model is a mathematical method used to describe the relationship between one or more predictor variables (also called independent variables) and a response variable (also called the dependent variable). When we fit a linear model, we want to know if the relationship between the predictor variables and the response variable is statistically significant. This helps us determine whether the predictors have a meaningful effect on the response, or whether the observed relationships could just be due to chance.

9.5.3.1 1. Hypothesis Testing for Each Coefficient

For each predictor in the model, we test the following hypotheses:

- **Null Hypothesis (H_0)**: The coefficient is zero, meaning the predictor has no effect on the response variable.
- **Alternative Hypothesis (H_A)**: The coefficient is not zero, meaning the predictor does have an effect on the response variable.

9.5.3.2 2. P-value

The **p-value** tells us the probability of observing the data, assuming the null hypothesis is true. For each predictor: - A **small p-value** (typically < 0.05) suggests that the predictor has a significant effect on the response variable, and we reject the null hypothesis. - A **large p-value** (typically > 0.05) suggests that the predictor does not have a significant effect, and we fail to reject the null hypothesis.

9.5.3.3 3. F-statistic (Overall Model Significance)

The **F-statistic** tests whether the model as a whole is better than a model with no predictors (a model that only includes the intercept). This tests the null hypothesis that all the coefficients (except the intercept) are zero: - A **significant F-statistic** (p-value < 0.05) suggests that the model explains a significant portion of the variation in the response variable.

9.5.3.4 4. R-squared

The **R-squared** value tells us how much of the variation in the response variable is explained by the model. While this isn't a test of significance, a higher R^2 value indicates that the model fits the data better.

9.5.3.5 5. Reporting your results

A results statement has several important components: 1. A description of the statistical results with an appropriate citation on your results (test statistic, p-value). 2. If you have a *statistically significant* effect, a description of the magnitude and/or direction of the relationship, which often cites a figure.

Here is an example of what your results statement should look like: The position of the transect affect nonnative species density ($F = 3.45$, $p = 0.02$). Quadrats located on the southern side of the field had higher levels of invasion (Fig. 1).

9.6 Assignment

When you are finished, report out on the primary research question that we addressed in lab: Does disturbance affect non-native plant density?

In your brief report, provide:

1. Your hypothesis for the question
2. Your if/then statement describing the patterns that you expect to see in order to identify the information that you need to gather
3. A results statement for each statistical test
4. A figure with an appropriate legend
5. Interpretation of your results, answering the question: Does disturbance affect non-native plant density? Did you support or refute your hypothesis?

Put this in a document and turn it into your TA!

Chapter 10

Water quality

10.1 Background

Stream ecology is best understood in the contexts of physicochemical and biological processes. The physical properties of streams include the slope, aspect, elevation and temperatures of the stream environment. The chemical properties of streams include a large set of characteristics; such as alkalinity, pH, nutrients - such as nitrogen and phosphorus, and the gasses dissolved in stream waters, including dissolved oxygen.

In this lab we will collect water samples at Francis Short Pond in order to document effects of Rio de Flag Flood Control Project on water quality to be compiled over time and provided to the City of Flagstaff. We will evaluate the current water quality of Francis Short Pond to determine whether it falls within standards established by environmental protection agencies worldwide.

10.1.1 History of Francis Short Pond

Some people think Frances Short Pond began as a water storage pond for the Santa Fe Railroad's steam trains. In the 1920s a dam was built upstream on the Rio de Flag to create a swimming and skating area. But over time the area got filled with trash, sediment and street sweepings dumped in by the city.

About 40 years ago (in the late 60s), the city of Flagstaff was considering paving over the pond and making it into a parking lot. At the time, Aztec Street ran right beside the pond's banks, connecting Cherry Avenue and North Thorpe Road. A local teacher rallied support and convinced the city council to preserve Frances Short, then worked with students to plant cottonwoods, Bebb willows, and junipers near the pond's banks and create the island in the middle of the water.

In more recent years high vegetation levels (algae, bulrushes etc.) have affected water quality, habitat, and recreational and educational uses of the pond. The city underwent one major restoration project in 2005 where it drained the pond and dredged it. That grant project also added a sediment cleanout area, a small spillway and a wetland to filter some of the storm runoff that enters the pond. An aeration system helps keep the pond livable for fish and other wildlife.

The city completed a similar process in 2015, removing vegetation and restoring the trail around the pond. The overgrown vegetation is a problem because during the times when those plants aren't photosynthesizing, such as at night or when the pond is covered in ice, they respire, pulling oxygen from the water. That lack of oxygen has caused fish die-offs in recent years.

The Arizona Department of Game and Fish regularly stocks the pond with hundreds of pounds of catfish, largemouth bass and bluegill sunfish. The pond also has nonnative Siberian elm and Russian olive trees that volunteers have tried to address in the past. The pond is supplemented with reclaimed water (in addition to whatever comes down the Rio de Flag channel).

10.1.2 Rio de Flag Flood Control Project

The **Rio de Flag Flood Control Project** is a large-scale infrastructure initiative in Flagstaff, Arizona, aimed at reducing the risk of flooding in the area, particularly in downtown Flagstaff and the Southside neighborhood. The project was developed in response to the recurring flood risks posed by the Rio de Flag, a waterway that runs through the city and is prone to flooding during heavy rain events.

- **Purpose**: To mitigate the flood risks for over 1,500 structures, including homes and businesses, and prevent potential damage estimated in the millions of dollars.
- **Scope**: The project includes upgrading stormwater infrastructure, creating detention basins, and improving the flow capacity of the Rio de Flag channel.
- **Partners**: The project is a collaboration between the City of Flagstaff and the U.S. Army Corps of Engineers.
- **Status**: While portions of the project have been completed, such as some channel improvements, the full implementation has been delayed due to funding and design challenges. Full completion is expected to significantly reduce the risk of flooding in Flagstaff.

Francis Short Pond plays an important role in the Rio de Flag Flood Control Project, including:

1. **Stormwater Retention**: Francis Short Pond helps capture stormwater runoff from nearby areas, acting as a retention basin. This helps reduce

the volume of water flowing into the Rio de Flag during heavy rain events, contributing to flood control by temporarily storing excess water.

2. **Flood Mitigation**: The pond reduces the pressure on the Rio de Flag's natural channel by holding stormwater, which helps manage water flow and reduce the risk of downstream flooding. By retaining stormwater, it lowers the volume of water that reaches the Rio de Flag during peak rain events, which is crucial in preventing flood events in urban areas like downtown Flagstaff and the Southside neighborhood.
3. **Water Quality Improvement**: The pond can help improve city water quality by acting as a buffer zone for sediments and pollutants before they enter the Rio de Flag system. As a retention basin, it allows suspended sediments, nutrients, and pollutants to settle out of the water column, preventing them from being washed downstream into more sensitive areas.

Given this, how do you think Rio de Flag Flood Control Project will affect the water quality at Francis Short Pond?

10.2 Key Water Quality Terms and Concepts

- **Dissolved Oxygen (D.O.)** – Oxygen that is dissolved in water; the most important indicator of water body health for the support of aquatic ecosystems.
- **Effluent**: Wastewater (sewage) that has been treated at a wastewater treatment plant. In the US, treated effluent is discharged.
- **Nitrate(NO3-)**: A nitrogen-containing organic molecule that is found in fertilizer and can be readily used by plants; excess nitrate in water can cause eutrophication. Other sources of nitrates include municipal and industrial waste water, septic tanks and private sewage disposal systems, urban drainage and decaying plant debris.
- **Phosphate(PO4-)**: A phosphorus-containing organic molecule that is derived from rocks or detergents; excess phosphate in water can cause eutrophication. Because phosphorus tends to attach to soil particles, it moves into surface-water bodies as a result of erosion and runoff into the water.
- **pH**: A measurement of how acidic or basic a solution is; technically, it is the concentration of hydrogen ions (H+) in a liquid. Perfectly neutral = pH of 7. Natural changes in pH occur with interactions between the water source and surrounding rock. pH can also vary with precipitation (especially acid rain) and wastewater or mining discharges.
- **Total dissolved solids (TDS)** Inorganic salts (principally calcium, magnesium, potassium, sodium, bicarbonates, chlorides, and sulfates) and some small amounts of organic matter that are dissolved in water. High TDS readings can be the result of water running through a region that has rocks with a high salt content. Human-caused sources include agricultural

and urban runoff, wastewater discharges, industrial wastewater and salt that is used to de-ice roads.

10.3 Water Quality Monitoring

The Environmental Protection Agency (EPA) and the Arizona Department of Environmental Quality (AZDEQ) set safe levels for naturally-occurring chemicals and contaminants in Arizona's waters. Water quality standards exist for drinking water, surface water, and wastewater effluent and vary depending on the water source and the intended use. Of the three sets of standards, drinking water standards are the strictest. The EPA currently regulates approximately 90 contaminants that occur in drinking water but adds to the list of regulated contaminants as we learn more about the effects of various chemicals on human health.

Water is perhaps our most precious natural resource in the arid Southwest. High-quality water is important in many aspects of our lives, from providing cities and towns with clean drinking water to supporting native fish and other wildlife species in rivers and creeks found in the region.

10.4 Downloads for this lab

Sometimes we conduct this lab outside at Francis Short Pond if weather permits. It takes a half an hour to walk from Physical Sciences to Francis Short Pond, however you are welcome to drive or bike. Plan to meet a half an hour after class starts!

- Get Directions to the location
- Download the map of the meeting place
- Download the excel file ## Hypothesis generation

The research question for this lab is: Does the current water quality in the wash fall within standards established by environmental protection agencies worldwide?

Given what you know about Rio de Flag Flood Control Project, generate a hypothesis to the research question.

How will you know if your hypothesis has been proved correct? Write an if/then statement!

10.5 Methods

Divide into four groups, and make sure that the backpack kit that you have has instructions for each of the tests listed below.

1) nitrate (N)
2) phosphorus (P)
3) stream water temperature (T)
4) stream water pH (pH)
5) dissolved oxygen (dO)
6) total dissolved solids (TDS)

GET NITRILE EXAM GLOVES FROM YOUR LAB INSTRUCTOR BEFORE HANDLING ANY CHEMICALS After you have the necessary number of gloves, carefully follow the instructions for each water quality test, detailed on the laminated cards. It is a good idea to read over the instructions for each test before you run the test, to make sure you don't forget or miss any steps. Gather all solid waste into the black trashbag provided, and place all liquid waste into the appropriately labeled waste bottles provided – it is very important that we remove all chemicals and trash from the pond so that we are not contaminating the environment.

10.5.1 Record and compare your results

Record the results of your tests in the excel file. Compare your results with other groups to ensure you all conducted the tests properly. If your numbers are way off, consult with your lab instructor. Now, compare the results you found with the global water quality standards (no need for a statistical test). Is the Francis Short Pond water within standard ranges for water quality?

10.6 Assignment

Turn into your TA:

1. Your hypothesis
2. Your if/then statement
3. Your table with water quality results
4. Your answer to the question: Is the Francis Short Pond water within standard ranges for water quality?

Chapter 11

Species interaction project

11.1 Introduction

When you look at a given ecological community (whether it's a ponderosa pine forest, a meadow, or a garden), you are looking at the combined results of thousands of interactions. Interactions can occur between species (i.e., interspecific interactions) or within (among the same) species (i.e., intraspecific interactions). Studies of ecological interactions are important in a range of disciplines, including agriculture, disease ecology, restoration, forestry and fisheries.

In this lab, you will use an experimental approach to explore various types of ecological interactions. This lab will require you to develop scientific questions, generate hypotheses, and design tests that will allow you to gather evidence to support (or fail to support) those hypotheses. This is a multi-week lab. On the first day of the lab, you will select and initiate your experimental tests. You will also begin to plan a final poster based on this lab. Over the next six to seven weeks, you will visit your experimental plants to measure their growth.

Organisms interact with one another to obtain food, protection, transportation, growth substrate, and other requirements. In interactions, the service or substance exchanged between organisms is known as "currency." Like an exchange of money, currency exchanges can be two-way (where both organisms receive something beneficial) or one-way (where only one organism receives something beneficial. Furthermore, one-way exchanges can exert a negative effect on the other organism (as in the case of predation) or may have no effect on the other organism. The characteristics of ecological communities are determined by interactions between organisms. That is, ecological interactions dictate what species are present in a given environment and in what numbers, what species are able to colonize or invade a particular location, and sometimes even whether species become extinct.

Ecological interactions are often classified by an interaction sign (Table 1). The sign is a reflection of the impact of the interaction on the reproductive potential of each participating organism. Interactions that are mutually harmful are delineated with two negative signs (-,-). Interactions that benefit one organism while being detrimental to another are delineated with a negative and a positive sign (-,+). Interactions that are mutually positive are delineated (+,+); and so on.

Understanding the sign of a given interaction can be important for predicting its effects on the community. If a new species is introduced to a site and acts as an herbivore, for example, it may be expected to decrease the reproductive potential of one or more plants in the area. This in turn can alter the abundance of those plants, and perhaps their competitive interactions or ability to interact with their pollinators or with other herbivores.

Over the next six to eight weeks, you will visit your experimental plants in the greenhouse every week to measure their growth. At the end of that time, you will perform final measurements on your plants and dispose of them. Toward the end of the semester, you will turn in a poster reporting your results.

11.2 Objectives

1. To explore ecological interactions using plants as model organisms.
2. To become more familiar with (and comfortable with) the scientific method of inquiry.
3. To practice scientific writing.

11.3 Week 1

11.3.1 Downloads for this lab

Please download:

- A reading about plant species interactions
- Experimental design worksheet for turn-in
- An example data sheet

This is a datasheet for an experiment with 12 pots and two species, with each pot containing species A & B. You may need to adjust the structure of the dataset depending your experimental design

We will be working in the NAU greenhouse. For those of you who have not visited the greenhouse yet, here is a mapped location:

- View NAU greenhouse location on Google Maps

11.3.2 Materials

Each group of two should have:

- 2 six-pack of pots
- Potting soil
- Seeds
- Labels
- Permanent marker (for pot labeling)
- Data sheet

11.4 Overall procedure

1. This lab will run between 6 and 8 weeks. The first week will be devoted to hypothesis generation, experimental design, and planting. During the following weeks, while the class is completing other lab assignments, students should arrange with one another to ensure that the plants from this experiment are measured at least once per week until the end of the experiment. At that time plants will be removed from pots and their root and shoot lengths measured and biomass placed in a drying oven.

2. In this lab, you will evaluate ecological interactions in very small, microcommunities: greenhouse pots. Considering your previous experience and the knowledge you bring to the lab, consider what might happen if you plant certain combinations of seeds in a particular pot and why a particular pattern might emerge.

3. Available to you are multiple plant species. By sowing these plants in pots, separately or together, you can create small-scale communities. Working in groups, develop one to two scientific questions relevant to ecological interactions among or within the species at hand. To do this, begin by brainstorming what sort of ecological interactions might occur if various combinations of seeds are used. Then, develop a question about an ecological interaction that is of interest to you and might be expected to occur among these species. Write down your question(s). You may want to google species that you use in your project in order to inform your question and hypotheses.

4. Develop a hypothesis for your scientific question. Hypotheses are possible answers to your questions. Write down your hypotheses. You'll include these in your poster. Check with your instructor to make sure you've appropriately developed your hypotheses.

5. Consider the type of evidence that might support each of your hypotheses ("if… and … then…"). Develop experimental tests (plant species/number of

seeds) that would give you the evidence for each of your hypotheses. Ensure that the experiments you design are scientifically robust (controlled, quantifiable, unbiased, repeatable, falsifiable, etc.).

Note: Since we have limited space, we won't be able to ensure independence of our replicates

6. Carry out your experiments. Each group of students will have the use of at least 12 small pots, labeled with your names. Note that most questions/hypotheses will require more than one pot (for example, pots for controls and treatments). Thus, the number of questions and tests carried out may vary among pairs of students. Carefully record the species and number of seeds used on a white label inserted into each pot so you know for sure what each pot contains. Each white tag should include: Your group's initials and lab section number. Tray # and Space # and what's in each pot and the quantity.

Example:

RBR, CA and EC – Lab Section 002 Tray # 1, Space #2

Radishes x 3

Here is an example label:

Note: the greenhouse is open between 7:00 am - 3:00 pm & you will need to get the greenhouse code from your instructor.

11.5 Set up your experiment

Now that you have generated your question, hypothesis, and identified what evidence you need to gather (if/then statement), set up your experiment. When planting, ensure sufficient light: smaller seeds should be buried approximately ¼ inch below the soil surface, and larger seeds approximately ½ inch. Take a photo of your pots once they are planted and the labels are in place.

11.6 Creating your Data Sheet

1. You should create a data sheet with your group members to successfully record your data for the next 6 - 8 weeks. Do this by modifying the example data sheet that we have provided.
2. Here are a few suggestions for recording data:

- If no germination, record "No" or 0.
- If dead, record "No" under the survival column or 0.
- If multiple shoots per species, you may want to record the average height.

11.7 Turn-in week 1

1. Name of group members
2. Date of planting
3. Question/s
4. Hypothesis/es
5. Evidence (if/then statement)
6. Drawing of your experimental design
7. Your data sheet and a description of what you will measure each week

11.8 Weeks 2 - 5

During this time, you will be maintaining your experiment and continue to collect data.Make a schedule with your group members to ensure that someone will measure growth of the plants at least once per week over the next seven weeks. The same person should also photograph the plants each time measurements are made. To measure each plant, place the end of a ruler against the soil surface. Gently pull the plant to its full height and measure the tallest point intersected along the ruler by the plant. ALL MEASUREMENTS ARE IN MILLIMETERS!

Note: the greenhouse is open between 7:00 am - 3:00 pm. You will need to get the pass code from your TA!

11.8.1 Analyzing and visualizing your data

Refer back to lab manual chapters 3 and 4 for a reminder about descriptive and inferential statistics. To run statistical tests on your data, refer to lab manual chapter 5: Selecting statistical tests.

11.9 Final week

11.9.1 Downloads

Please download:

- Poster instructions

- Peer evaluation

Once you have all of your data, you will prepare a poster with your group detailing your experiment and its outcome. For the poster you will follow standard scientific writing style (see handout). Using the statistical lessons and R scripts that you have received earlier in the semester, analyze your data and create figures. Remember the correct methods for reporting results and writing figure lessons.

11.9.2 Turn-in

Submit your **poster as a PDF** to your lab instructor through canvas (or via email, if instructed by your TA). Only one person in your group needs to submit. Each group member must also submit a **peer evaluation**, assessing the contribution of each group member.

The poster is in format C, which means that the poster must include:

- 5 sources cited (look at other poster examples to see how to cite in text without using a lot of space)
- An introduction, methods, results, discussion, and conclusion section.
- Include least one figure or table **and** one photo of your plants.

Other notes:

- You do not need an abstract on this poster.
- 15% of your grade will be determined by your group mate's assessment of your participation level.

Chapter 12

Species diversity

12.1 Introduction

Diversity and the measurement of diversity are central to many issues in ecological research as well as for applying ecology to real world problems. Every textbook in ecology devotes considerable description and explanation of species diversity, species richness, and species evenness. Community ecologists use measures of diversity to study and explain ecological patterns in many different types of communities.In terrestrial ecosystems, litter decomposition has important effects on processes such as nutrient cycling and community structure. Decomposition is affected by the type and quality of litter, climate, the edaphic conditions (including soil temperature, hydration, and chemistry), and the community of decomposer organisms (Swift et al. 1979).

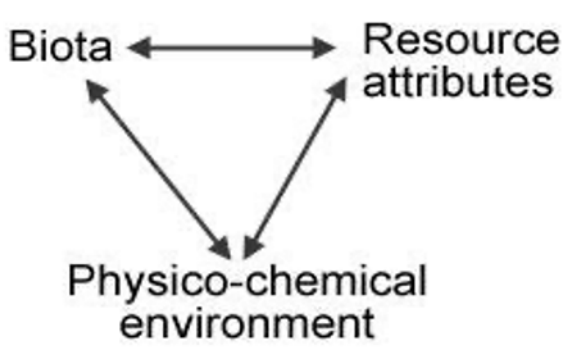


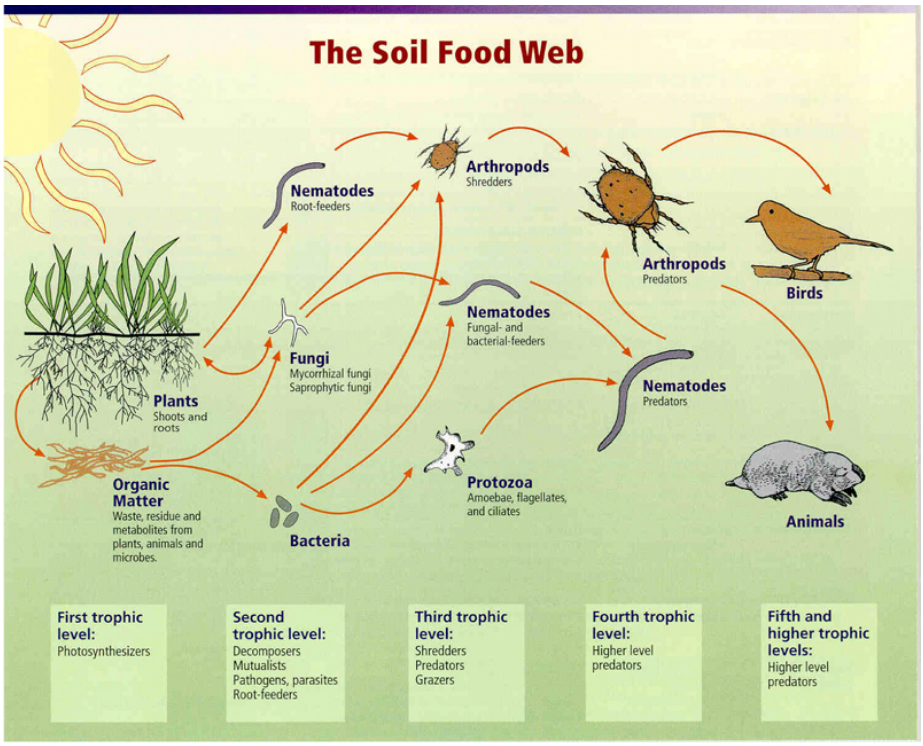
Figure 1. Interactions among factors that control litter decomposition (from Swift et al. 1979).

This model shows the relationships among the three factors that govern litter decomposition rates: the Biota (structure and activity of the biotic soil food webs, i.e., microbes, invertebrates, vertebrates), the Physico-chemical environment (climate, habitat, edaphic factors, i.e., contributions from the non-living

environment); and Resource Attributes (primarily plant species diversity and tissue chemistry, i.e., contributions from the living environment).

Many studies have shown how both the living and the non-living environments affect soil community structure and diversity (Swift et al. 1979, Elliott et al. 1980, Ingham et al. 1982, Freckman & Virginia 1997). For example, decomposition of plant litter that is high in lignin and/or low in nutrients and is therefore difficult to decompose (resource quality) leads to dominance by fungal-feeding groups in the soil food web (namely, some taxa of nematodes, mites and Collembola), whereas easily broken-down litter is decomposed primarily by bacteria, which is reflected higher up the food chain (Coleman & Crossley 1996). And soil community diversity is at least partially determined by plant community diversity (Siemann et al. 1998). So in this case, the living environment is determining the soil community.

On the other hand, recent work suggests that composition and biodiversity of soil organisms itself may have a greater effect on decomposition than has been previously recognized (González and Seastedt 2001, Wardle and Lavelle 1997, Wardle et al., 2003), especially in tropical ecosystems. So in this case, the soil biota is the driving force of the Physico-chemical environment and therefore the Resource Attributes in the Swift et al. (1979) model above. On yet a third hand, the soil Biota can directly affect the Resource Attributes. De Deyn et al. (2003) showed that soil fauna enhanced succession and diversity in a grassland community.



Relationships between soil food web, plants, organic matter, and birds and mammals
Image courtesy of USDA Natural Resources Conservation Service
http://soils.usda.gov/sqi/soil_quality/soil_biology/soil_food_web.html.

Soil invertebrates play important roles in soil communities. Some directly consume **detritus**, others consume **detritivores**, whereas others are higher-level **carnivores** that can indirectly control decomposition by their effects on lower levels of the food web (see Soil Food Web figure on the next page). The classic study of detrital food webs was conducted by Gist and Crossley (1975), showing which invertebrate groups are detritivores and which are carnivorous. Smith and Smith (2001) has a good description of the major findings of that study.

Soil invertebrates are clearly affecting litter decomposition rates, soil aeration, nutrient mineralization, primary production, and other ecosystem services related to soil ecosystem function and agroecological conservation (e.g., Six et al. 2002). With interest in global climate change has come the realization that soil biota may strongly affect soil CO2 sequestration and release, which is a critical variable in climate change models. Agroscientists and restoration ecologists have found that soil biota play critical roles in toxic chemical and metal mobility and remediation; they directly affect disturbed ecosystem recovery/ ecological restorations that occur after fire, UV-B exposure, post-urbanization, and herbicide-stressed soils (e.g., Lal 2002). Bioprospectors carry out the search for novel antibiotics and other drugs among the billions of soil microorganisms.

Soil invertebrates are also recognized for their role in mediating or determining belowground interactions among plants. Because they are often prey for vertebrates such as birds and mammals, they have vital roles in the food chains that include those animals. Take notes on the compost process from the videos linked to below. You will be able to reference this video in your lab report when discussing the methods for making compost. You will be turning this in, along with photos of the experimental set-up.

Here are the video links:

- Watch video 1 on YouTube
- Watch video 2 on YouTube
- Watch video 3 on YouTube
- How to compost

12.2 Week 1

12.3 Downloads for this lab

Directions to the compost area

12.4 Objectives

Determine whether arthropod diversity varies with the depth of compost. **Research Question**: How does soil arthropod diversity vary with depth of compost?

Record your hypothesis to the research question, along with how you will know if you hypothesis has been proven correct (if/then statement). You will turn this into your TA!

12.5 Methods

Study Site(s): The class will be measuring soil invertebrate diversity in samples collected from the NAU campus compost facility.

1. Collect soil from the NAU campus compost facility. Working together, be sure to collect soil from different depths (surface, deep) within the pile and be sure to track at which depth the soil was collected. If weather is inclimate, your TA may have collected soil for you.

2. Working in groups of 2-3, prepare your funnels: Remove a PVC pipe container. Cut a section of cheesecloth (two layers thick), and attach it tightly to the narrower end of the PVC pipe with rubber bands. Use masking tape to label the container with your compost age, lab section number and group member initials.

3. Scoop compost into the PVC pipe container until it is approximately half full. Make sure there are no solid clumps of compost in the container; you may need to gently break apart any clumps. Place the container into its place above the funnel and below a lightbulb.

4. Pour approximately 1 inch of half water, half 90% isopropyl alcohol into a 20-mL vial (3/4 full). Label the funnel with the same information as the PVC pipe. Place the vial beneath the funnel.

5. Screw in the light bulb above your funnel to turn it on. This is a behavioral extraction technique: many soil invertebrates will move away from light, and these organisms will end up falling through the funnel and into the alcohol. Your instructors will also monitor the liquid in each of the small vials throughout the week when you are not in lab. Moisture from the extractors may condense in those vials and increase the depth of their liquid, while diluting the alcohol.

When you return to lab next week, you will examine the contents of the 20-mL vials to observe the organisms that have been collected, and identify them using keys. You will then perform data analysis, calculating diversity metrics.

12.6 Turn-in week 1

Please turn-in:

- Notes from the video of the compost process
- Your hypothesis and if/then statement
- A photo of the Tullgren / Burlese funnel set-up

12.7 Week 2

12.8 Downloads for this lab

Please download this information:

- Data sheets for recording arthropod diversity

- Taxonomic key for arthropod identification
- Excel files to calculate diversity
- Reading on Shannon's Index
- Discussion questions

In this lab we will be analyzing the data from the compost arthropod lab to determine values for species richness, evenness, and diversity. We will compare two metrics of soil invertebrate biodiversity: We will be using the Shannon's index (H´) to calculate species diversity. This is one of the most common measures of species diversity in ecology, though there are others. Shannon's index takes into account both richness (the number of species) and evenness, or how evenly individuals are distributed among species. A large value of H´ denotes high biodiversity. Shannon's index is advantageous over simply counting the total number of different species, because a simple count is affected by sampling effort (plot size and total number of individuals sampled): the larger the sample, the more rare species you find. H´ is superior because it is calculated from proportions (%), as you will see, and rare species contribute very little. Therefore, this index is relatively insensitive to the random inclusion or omission of rare species that happens with any sampling effort. See the background information on Shannon's index for more information.

12.9 Objectives

Identify arthropods and calculate diversity of compost collected from two different depths.

12.10 Methods

You will be identifying invertebrates (Arthropods) from the compost using dissecting microscopes and dichotomous keys. Use the dichotomous key provided to make your best guess at the taxon for each organism in your sample. Record abundances in each sample using the table below.

12.11 Turn-in week 2

Please turn in your discussion questions.

12.12 Acknowledgements

Material in this lab is based mainly on: Boyce, Richard L. 2005. Life under your feet: measuring soil invertebrate diversity.http://tiee.esa.org/vol/v3/experiments/soil/abstract.html.

Chapter 13

Species Distribution Modeling

13.1 Introduction

Species distribution models (a.k.a.; ecological niche models, habitat models) relate environmental predictors like climate, elevation, or soil characteristics to species presence or abundance. These relationships are used to project likelihood of occurrence across space, by calculating the likelihood of occurrence across the study area using values associated with raster maps of the environmental variables in the model (**Fig. 1**). Most SDMs are correlative models that mathematically describe observed patterns of occurrence, and that do not incorporate underlying mechanisms in model projections. Understanding the limitations of correlative models (discussed below) is important for deciding when to use these models *and* interpreting your results.

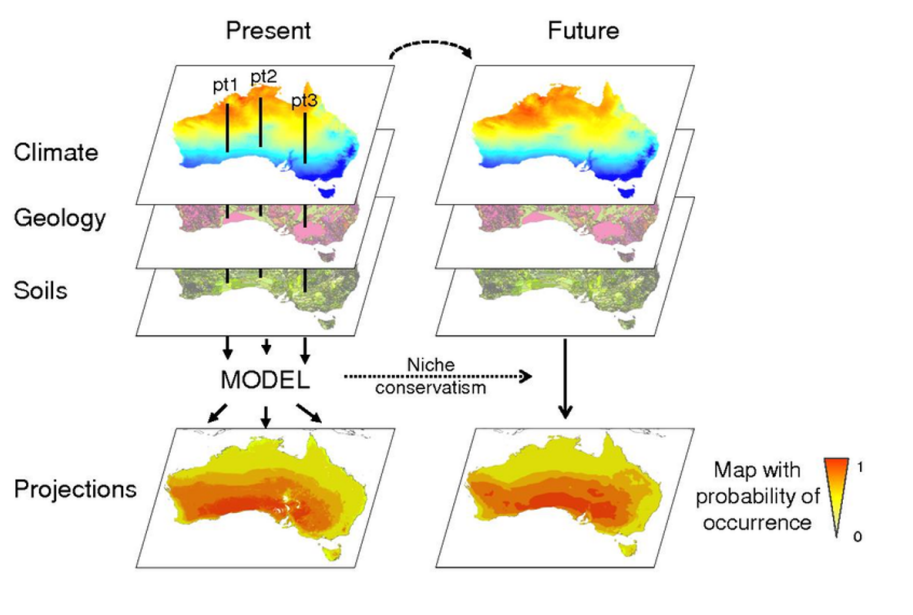


Figure 1. Raster layers are stacked to predict likely habitat.

13.2 A brief review of niche theory

In spatial and population ecology, we define a species' niche as all the conditions under which populations of a species maintain growth rates that are at or exceed replacement rates. The niche is often defined in n dimensional space, since so many factors contribute to the performance of a species (**Fig. 2**). In ecology, there tends to be a lot of confusion about spatial niche concepts, since ecology students often first learn about species niches from an evolutionary standpoint. In evolutionary ecology, a species' niche is defined by a suite of traits possessed by an organism related to how this species 'makes its living' (attains food, nutrients, water). However, niche concepts and definitions are inherently related, and broadly describe the role of species within ecosystems.

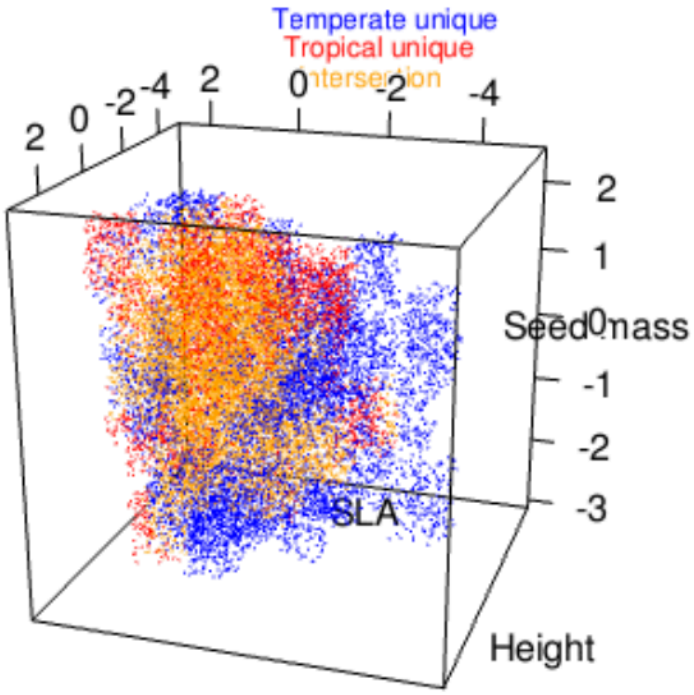


Figure 2. Species niches are complex. In theory, we could describe the niche of a species by adding N number of axes and create a cloud representing all the habitats where a species could persist.

Ecologists break a species niche into two components: the **fundamental niche** and the **realized niche**. The fundamental niche is similar to the Grinnellian niche concept (*Introduced by Joseph Grinnell in is 1917 paper, The niche relationships of the California Thrasher*) niche concept. The **fundamental niche** of the species describes all the abiotic conditions that a species can physiologically tolerate and maintain population growth at or above replacement rates (we don't count areas where species persist, but are not maintaining themselves; these area are known as demographic sinks). The **realized niche** refers to all of the areas that we actually observe a species on the landscape. The concept emerged out of the work of Elton, who highlighted the importance of species interactions in defining species distributions. The **realized niche**, therefore, reflects the combined effects of the abiotic and biotic environment on persistence across the landscape.

Typically, the **fundamental niche** is larger than the **realized niche**. In other words, a species can be found in a lot of places, but processes such as competition for resources or predation, reduce the total area occupied by a species. In some

cases, the **realized niche** can be larger than the **fundamental niche**. This occurs when positive species interactions, like mutualisms or facilitation, allow a species to overcome some sort of environmental resistance and occupy sites that would be inhospitable for the species without 'help from a friend'. Finally, stochastic events, like disturbances, or landscape features, such as barriers to dispersal, can influence where as a species is found on the landscape.

The distinction between **fundamental** and **realized** niches is important in order to understand the limitation of correlative SDMs, since these models cannot distinguish between the fundamental and realized niche of a species. Recall that species distribution models relate environmental factors to current occupation of a species. In most cases, habitat suitability is predicted using abiotic factors, including climate, soils, topographic information; all of which essentially describe the fundamental niche of a species (i.e., physiological tolerance to abiotic characteristics). However, the data used to build these models quantifies the current occupation of a species on the landscape, or the realized niche.

This presents several issues:

- While we can learn generally about the abiotic factors that affect a species range, it is not a perfect picture of these tolerances, since distribution is affected by a variety of factors not included in the model.
- When a species fundamental niche and realized niche are really different due to factors not included in the model, current habitat projects can be inaccurate. This is a common problem when modeling habitat suitability for wild-harvested species, since they are underrepresented in suitable habitat, because those habitats are targeted for harvest.
- Using these models to predict future suitability should be interpreted skeptically. Future habitat suitability predictions are strong working hypotheses for ecological investigations or management actions. We don't actually expect many species to track their bioclimatic niches, since many of the factors that influence a species range aren't directly measured when building SDMs. SDMs are particularly unreliable when factors that shape a species niche change as a function of climate change. For instance, species interactions shape species distributions, and since species respond idiosyncratically to climate change, represent a major source of uncertainty in model predictions.

These caveats and limitations, particularly related to your data, should be included in the discussion of your results. All of that said, SDMs can provide us with a lot of information with relatively little effort, and are often the best hypothesis on which to base decisions.

13.3 Downloads for this lab

Create a file on your desktop called 'speciesdistributionmodels' and place the following files within it:

- R script for creating the SDM
- Occurrence data
- Worksheet to turn-in

13.4 Objectives

Project the effects of climate change on *Pinus ponderosa* in Arizona.

13.5 Methods

13.5.1 Let's get modeling

Now, let's walk through an example to discuss the various considerations and options for creating SDMs. For this exercise, we will use bioclimatic variables as environmental predictors. Bioclimatic variables are derived from downscaled climate models and created to be more more ecologically relevant compared to simple temperature and precipitation means. Bioclimatic variables are a great first step in model building, but depending on your study species and area, you may need to download finer scale layers or include other factors, like soil data.

Resolution

Raster layers are spatially mapped grids comprised of hundreds, thousands, or millions of cells (aka pixels) with values related to a variable assigned to each pixel. The smaller the pixel, the higher the resolution, but this greatly affects processing speed and may exceed computer storage.

Map projections

Different methods are used to project the 3D earth into a 2D map. We have to specify the projection of the layers used in our models or our data layers will not align properly. In the example below, we will specify a coordinate reference system (CRS), which defines, with the help of coordinates, how the projected map relates to locations on the earth.

A CRS contains the following information:

- Coordinate system: The X, Y grid that defines where a point is located in space.

- Horizontal and vertical units: The units used to define the grid along the x, y (and z) axis.
- Datum: A modeled version of the shape of the Earth which defines the origin used to place the coordinate system in space. You will learn this further below.
- Projection Information: The mathematical equation used to flatten objects that are on a round surface (e.g. the Earth) so you can view them on a flat surface (e.g. your computer screens or a paper map).

Luckily, we can pull all of this information from a spatial object, use the CRS function and reproject our data so that we are working with all data using the same CRS. Let's start by installing and loading the libraries that we will need for our analysis, and by importing both current climate and future climate projections.

For this exercise, we will download bioclimatic variables to characterize current climate. Bioclimatic variables are variables derived from mean, maximum and minimum temperature and precipitation data summarized from weather station data from across the globe, then *interpolated* based on various landscape features, most importantly elevation, in order to assign climatic values to locations with no climate stations. These bioclimatic variables have been created in order to represent climate data in a way that is biologically-relevant. Specifically, these bioclimatic include:

- Annual Mean Temperature (bio1)
- Mean Diurnal Range (Mean of monthly (max temp - min temp); bio2)
- Isothermality (bio3), Temperature Seasonality (standard deviation $\times 100$; bio4)
- Max Temperature of Warmest Month (bio5)
- Min Temperature of Coldest Month (bio6)
- Temperature Annual Range (bio7)
- Mean Temperature of Wettest Quarter (bio8)
- Mean Temperature of Driest Quarter (bio9)
- Mean Temperature of Warmest Quarter (bio10)
- Mean Temperature of Coldest Quarter (bio11)
- Annual Precipitation (bio12)
- Precipitation of Wettest Month (bio13)
- Precipitation of Driest Month (bio14)
- Precipitation Seasonality (Coefficient of Variation; bio13)
- Precipitation of Wettest Quarter (bio16)
- Precipitation of Driest Quarter (bio17)
- Precipitation of Warmest Quarter (bio18)
- Precipitation of Coldest Quarter (bio19).

Additionally, we will download climate projections. Climate projections are generated by Global Climate Models, which predict future climatic conditions

based on complex algorithms describing the atmosphere. In order to project future species distributions, we need to select a particular climate model and a time period for which we are making predictions. Here, we are projecting suitable habitat for the time period 2061-2080. Since several facilities equipped with climate models generate climatic projections, we also have selected the CNRM-CM6-1 modeling group.

Finally, we select the 'socio-economic pathway' utilized by our climate model. The degree of warming that occurs depends primarily on decisions that humans make around fossil fuel use and other climate mitigation strategies. The Intergovermental Panel on Climate Change (IPCC) works with social scientists, legislators and other to generate possible carbon use futures, which are then used to generate climate predictions. Here, we will use the Shared Socio-economic Pathway (SSP) '585'. Take a minute to search SSP 585.

Record the answers to these questions on your worksheet:

Provide a description of SSP 585. How do countries respond to climate change in this scenario? What climate-related technologies are assumed to be used in this future?

Is a low, middle-of-the-road, or high degree of warming predicted in this scenario? If you were to repeat this exercise, which SSP would you choose and why?

13.5.2 Run the R code

Load your R file run and walk through the exercise.

Now that we've loaded our current and future climate models, let's input our occurrence data. For our data on species occurrence, we will use data from the Global Biodiversity Information Facility (GBIF), a repository for species observations and locations derived from multiple sources, including citizen science, herbarium and museum collections. The GBIF data are biased by observer behavior, since many observations are derived from citizen science projects. Humans tend to collect data from easily accessed areas around roads, popular hiking trails or congregating areas. One way to reduce bias in presence only data is to use spatial thinning to reduce weighting observations from heavily trafficked areas more than observations made in other areas. There is no standard thinning distance, but it is typical to require a minimum of 5 km between observations. If you are working at smaller or larger scales, you could reduce

or increase this distance! Also, if you are using presence and absence data or have employed an unbiased sampling strategy to collect data, you can skip this next step. While there are various quality control measures used by GBIF to ensure high data quality, we will run a few other data cleaning codes to remove anomalous points. This step is not necessary if using nonGBIF data! However, when producing your own spatial datasets, some form of data quality control is necessary.

For this exercise, we will investigate whether models predict that Ponderosa pine forests that surround Flagstaff are predicted to persist as climate changes. Let's download occurrence data for Ponderosa pine (*Pinus ponderosa*).

Great! Now we have data! Let's build an SDM. Our first decision point on model construction is based on the response variable. In this case, we have presence-only data; in other words, no one went out to the field to confirm locations where a species is *absent* across the landscape. Since location information often suffers from *absence* of absence data (ha), researchers have found statistical workarounds, which produce amazingly similar results to models fitted with presence or absence data. The method most commonly used to model habitat suitability with presence-only data involves generated numerous background points across your study area to compare with areas where your species is present. Note that your focal species *could* occur at any one of these background points, but that doesn't matter. Essentially your model is characterizing habitat available to the species and the habitat of known occurrence to identify the environmental factors that best distinguish occupied habitat.

Let's breakdown the major model types used for SDMs:

Profile techniques

Profile techniques are simple algorithms that use environmental distance to known sites of occurrence to 'profile' habitat characteristics. These techniques are rarely used any more, so I won't discuss further! **Profile techniques include**:

- Mahalanobis distance
- Ecological niche factor analysis (ENFA)
- Isodar analysis
- Bioclim

Regression-based approaches

You are familiar with regression based approaches from other statistical analyses! All regression approaches build upon standard regression models (**Fig. 3**), but differ in subtle ways to address common challenges to data modeling, like issues of nonnormality or heterogeneity of variance.

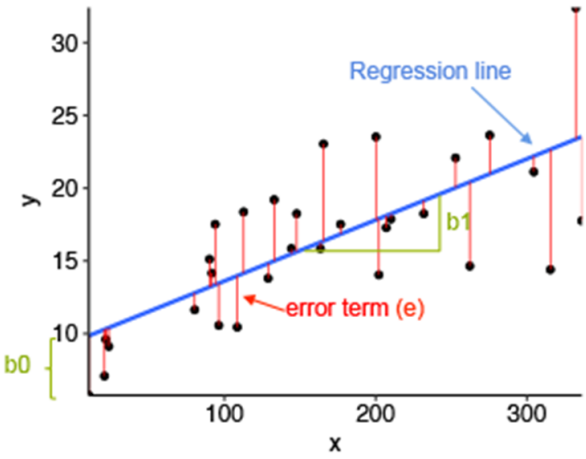


Figure 3. Regression basics. Term y is related to term x. If the slope of the line differs significantly from 0, then there is a relationship between the variables. Error terms are derived from the residuals of the model; how much each individual point deviates from the line of best fit. The best fit is determined by repeatedly mapping lines across the data to identify that which most reduces error.

Regression-based techniques include:

- Generalized Linear Modeling (GLM) (parametric)
- Flexible Discriminant Analysis (FDA) (parametric)
- Multivariate Adaptive Regression Splines (MARS) (nonparametric)
- Generalized Additive Modeling (GAM) (nonparametric)
- **Generalized Linear Models** are a flexible form of regression models. GLMs are 'generalized' by using a link function to relate the linear model to the response variable (which can be binomial, continuous, count data or other) and by relativizing the variance of each model term to its predicted value.
- **Generalized Additive Models** incorporate 'smoothing functions' to allow nonparametric estimates to be generated using a Bayesian approach.
- **Multivariate Adaptive Regression Splines** automatically models data nonlinearities and interactions between variables.
- **Flexible Discriminant Analysis** uses optimal scoring to transform the response variable so that the data are in a better form for linear separation.

Machine learning approaches: Machine learning techniques use training data to 'learn' about the dataset in order to make predictions.

Machine learning approaches include:

- Random Forest (RF)
- Boosted Regression Trees (BRT)
- Maximum Entropy (MaxEnt)

There are other machine learning techniques, like Artificial Neural Networks (ANN), but the list above is most commonly used for distribution modeling!

Random forest and **boosted regression trees** are similar, in that they create different 'trees' by iteratively bifurcating the dataset using predictor factors and identifying the tree that best predicts species occurrence.

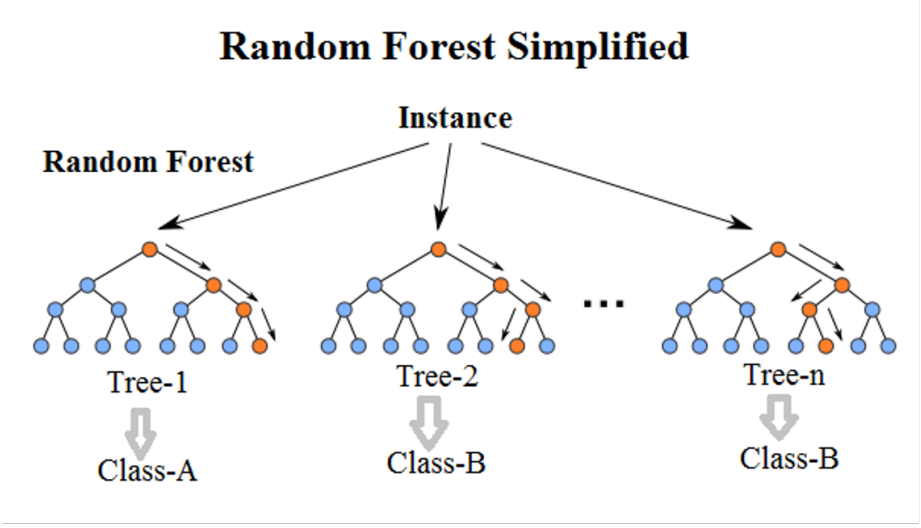


Figure 4. An illustration of random forest tree construction.

MaxEnt models are a little different. According to the principle of maximum entropy, high entropy is when the probability distribution best represents the current state of knowledge about a system, in the context of precisely stated prior data. These models evaluate the set of all trial probability distributions that would encode the prior data and select the distribution with maximal information entropy.

13.5.2.1 Model selection

The world of species distribution modeling is a contentious one! Many leaders in the field have their own 'pet' models that invariably they helped to develop software or methodology for! The general consensus is that each of the

different modeling techniques has various strengths and weaknesses, and they should be combined into ensemble models for habitat predictions. However, other approaches exist. One line of thinking in distribution modeling is to use solely GLMs, spending great care to identify critical predictor variables in a way that is tied to current ecological understanding and that reduces nonlinearities among these variables. By taking these steps, models are created, which in theory, should provide better inferential power for both current and future habitats. For presence only data, maximum entropy models are generally considered an excellent model choice. In my experience, there is no perfect model, rather model accuracy varies from species to species. For this reason, I typically build ensemble models to integrate the strengths of different model types.

13.5.2.2 Build an SDM

Deal with environmental predictor colinearity

In general, it is recommended to avoid having correlated features (variables that have different numbers, but are following the same pattern) in your dataset. Indeed, a group of highly correlated features will not bring additional information to our analyses, but will increase the complexity of the algorithm, thus increasing the risk of errors. Including highly correlated variables in models also, in essence, weights the correlated variables more than independent variables, again leading to less accurate model outputs. In other words, we need to remove highly correlated variables. We will do this by generating Variable Inflation Factor (VIF) values, a measure of collinearity, for all predictor variables. Then, we will remove one of the two correlated variables.

Build the dataframe for the SDM Building the SDM, requires two additional steps. In the first, we assemble the final dataset to be used in the model.

Using the sdmData function, we indicate the following:

- The column that contains presence data.
- The environmental predictors.
- Absence data or how to create background data.

Specify model evaluation parameters

When we build the final model using the SDM function, we specify **replication**. **Replication** is the method used to partition the dataset into training and test data. Ideally, we would have collected completely independent training and test datasets; however, I've never actually seen this done, except for researchers who are investigating SDM methods. Ninety nine point nine percent of the time, datasets are split into test and training datasets. As the names imply, training data are used to build the model, and then test data are used to measure how good our predictions are by quantifying how often our model correctly predicts presence or absence. Splitting or partitioning data into test and training

datasets is often conducted several times, since outcomes may depend on the test or training data used to build and evaluate models.

There are several methods to create training and test datasets. The three available in the package that we will use are subsampling (sub), crossvalidation (cv), bootstrapping (boot). For sub and boot, you must indicate what proportion of test and training data. A 30% test data, 70% training data split is common (test.percent=30). Finally, you will also the models how many times to repeat evaluations using the n equals code. This can eat up a lot of memory, so I typically use an n of 5.

Choosing the evaluation model

Crossvalidation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, in most methods, multiple rounds of crossvalidation are performed using different partitions, and the validation results are combined (e.g. averaged) over the rounds to give an estimate of the model's predictive performance. Crossvalidation does not rely on random sampling, but rather splits the dataset into k unique subsets. This is the preferred method for spatial model evaluation and estimating generalization capability. Note that you have to select the number of 'folds' or data partitions, which is typically set at 5.

Bootstrapping iteratively creates separate datasets from randomly sampling with replacement. Bootstrapping it is not as strong as crossvalidation when it is used for model validation, since it contains repeated elements in every subset. Bootstrapping is typically repeated 30 times in SDM model evaluation!

Subsampling randomly splits the dataset into training and test datasets, but doesn't maintain the independence of the datasets. In other words, due to random sampling, you might wind up with similar training and test datasets in each trial. For this reason, the more structured crossvalidation method is typically preferred.

Build the SDM Once the data is appropriately compiled, we use the sdm function to build the actual model. **Within this function, we specify**:

- The column that contains presence or presence or absence
- The dataframe that we are using (d1)
- The types of models that we are using
- Replication type (cv), number of folds (5), how many times to repeat partitioning (1)

THIS STEP WILL TAKE SOME TIME - JUST LET THE PROGRAM RUN!

The model object (m1) tells you several things. First, it gives a brief summary of the model you ran. You can double check this to be sure that the model did what you told it to do. Here, everything seems fine: We ran a model for one species,

we used two modeling methods, glm and maxent, we used cross_validation with 5 partitions. The model runs were successful (100% each). Finally, we are provided with 4 measures of model performance: AUC, COR, TSS, and Deviance.

What types of models are we using to predict habitat suitability for Ponderosa pines (i.e., machine learning, regression, profile techniques)?

13.5.2.3 Model evaluation explained

AUC stands for Area Under the Curve. AUC refers to a ROC plot, which plots sensitivity over 1 minus specificity. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. **AUC is desirable for SDM model evaluation for two main reasons**:

- AUC is scale invariant.
- AUC is classification threshold invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

This curve plots two parameters:

- True Positive Rate (Sensitivity): the proportion of presences correctly predicted as presence,
- False Positive Rate (1 minus Specificity): The specificity denotes the proportion of absences that are correctly predicted as absence, so the false positive rate indicates how many times the model predicted an occurrence when there was none.

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. As a rule, an AUC > 0.75 indicates a high performing model.

SDMs predict a probability of occurrence across the landscape. Different thresholds can be used to create a cutoff to predict presences or absences. For instance, we could say that if there is a 90% chance or more that a pixel is suitable habitat, then we consider those areas as occupied. We want to identify a cutoff that maximizes true presences, while minimizes false positives (i.e., areas that you incorrectly say contain a population, but don't). You can see that as you decrease

the cutoff, from say 90% to 70%, then your likelihood of correctly predicting presences goes up, BUT so does your likelihood of false positives.

So, let's check out the ROC plot below.

Looking at AUC, a common form of model performance assessment, which is the best performing model?

Generally, there is strong agreement between the test and the training data. As you increase the cutoff threshold, the likelihood that you correctly assign presence goes up, but so does the false positive rate (**Fig. 5**). Note that on the far right hand side of each ROC plot, if the cutoff is high enough, you will have a 100% true positive rate, and a 100% false positive rate (the cutoff is so low, that all habitats are predicted to support the focal species). Alternatively, with a low enough cutoff (left hand side of the ROC plot), you won't have any positives or any false positives! This AUC cutoff will be important for building ensemble models; explained below!

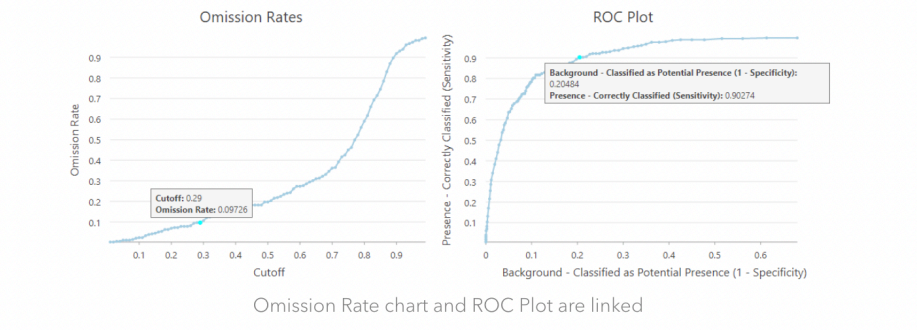


Figure 5 This figure (lifted from an ARCGIS website) shows the relationship between omission rates and ROC plots. So if we want an Omission Rate that is slightly less than 10%, we can use 0.29 as the Cutoff instead, and in this case, we will pick up 20.48% background points as potential presence locations, which is also a good rate.

Like AUC, **True Skill Statistic (TSS)** values are calculated across the range of possible thresholds for classifying model scores.The TSS similarly incorporates sensitivity and specificity comparing models against random, yielding values that range from negative 1 to positive 1, where positive 1 indicates perfect prediction and greater than or equal to 0 indicates a model that performs no better than random. TSS is typically considered a better indicator of model performance for presence only models. **A TSS of 0.5 or higher indicates high model performance.**

Pearson correlation (COR) between the predicted likelihood of presence and the presence or absence testing data. **Deviance** Lastly, if a model is interpreted

as estimating species' probability of presence, rather than just giving an index of habitat suitability, then the model predictions can be evaluated using deviance, defined as 2 times the log probability of the test data.

13.5.2.4 Ensemble model assembly

Finally, we will merge models into an ensemble model. You may want to exclude models that didn't have high predictive performance. We will give higher weights to the models with higher accuracy, in this cause using the TSS score.

13.5.2.5 Investigating model components

We can run code to look at the model components that best predict presence.

According this figure, which climatic variable best predicts habitat suitability for Ponderosa pine?

13.5.2.6 Convert to presence or absense predictions

We use the test statistics to identify a threshold that maximized true positives and reduced false negatives. In order to do this, we will create a new raster and populate it with predictions of presence, using that threshold.

13.5.2.7 Plotting and predictions

Let's take a quick look at the predictions we have created for the current time period.

To predict response of your focal species to future climate, just plug the novel climate conditions into the model! Let's run this model.

Now, let's plot this prediction against our original! First, we'll take a zoomed out look, then we will focus into our region!

Let's convert to predicted occurrence and plot!

According to these maps, how will the amount of suitable habitat for *Pinus ponderosa* in Flagstaff change if climate change continues along the SSP 585 projection?

How certain are you of these projections? Why might these models NOT be accurate?

What type of vegetation do you think might be more common around Flagstaff as climate changes?

Submit your answers to the questions presented throughout this tutorial and the figures that you generated (Occurrence of *Pinus ponderosa* currently and in the future) to your TA.

Chapter 14

ENV 226 Lab Exercises

14.1 Conclusions

Thanks so much for a wonderful semester! I hope that the code and information can be useful in future classes during your time at NAU! Have a wonderful break!