

# **CCESR Intern Hub**

CCESR Fellows

# Table of contents

<b>Welcome!</b>	<b>5</b>
<b>I Data Analysis</b>	<b>6</b>
<b>1 Data Analysis at a Glance</b>	<b>7</b>
<b>2 Data Types</b>	<b>8</b>
2.1 Numeric Data . . . . .	8
2.2 Categorical Data . . . . .	8
<b>3 Descriptive Statistics</b>	<b>10</b>
3.1 Centrality . . . . .	10
3.2 Spread . . . . .	12
3.3 Other Descriptors . . . . .	12
3.4 Ecological Community Descriptors . . . . .	12
<b>4 Inferential Statistics</b>	<b>13</b>
4.1 Classic Frequentist Tests . . . . .	13
4.1.1 Assumptions . . . . .	14
4.1.2 Categorical Predictor/s, Numeric Response . . . . .	14
4.1.3 Numeric Predictor/s, Numeric Response . . . . .	15
4.1.4 Numeric Predictor/s, Categorical Response . . . . .	16
4.1.5 Categorical Predictor/s, Categorical Response . . . . .	16
4.2 Bootstrapping . . . . .	16
<b>II R on your Computer</b>	<b>18</b>
<b>5 R Itself</b>	<b>19</b>
5.1 R, the Language . . . . .	19
5.2 R, the Software . . . . .	19
5.3 R Packages . . . . .	20
<b>6 R Studio</b>	<b>21</b>
6.1 RStudio at a Glance . . . . .	21

6.2	R Projects . . . . .	22
<b>7</b>	<b>Optional: Git and Github</b>	<b>23</b>
<b>III</b>	<b>R Programming</b>	<b>24</b>
<b>8</b>	<b>The Basics</b>	<b>25</b>
8.1	Intro . . . . .	25
8.2	Operators . . . . .	25
8.3	Functions . . . . .	28
8.4	Arguments . . . . .	28
8.5	Data . . . . .	29
8.5.1	Values . . . . .	29
8.5.2	Variables . . . . .	31
8.5.3	Data Structures . . . . .	32
<b>9</b>	<b>Next Steps</b>	<b>34</b>
9.1	Packages . . . . .	34
9.2	Subsetting . . . . .	35
9.2.1	Vectors . . . . .	35
9.2.2	Lists . . . . .	36
9.2.3	Data Frames . . . . .	39
9.3	Optional: Flow Control . . . . .	43
9.4	Optional: Writing Functions . . . . .	43
<b>10</b>	<b>Importing Data</b>	<b>44</b>
10.1	Reading Data . . . . .	44
10.1.1	From Your Computer . . . . .	44
10.1.2	tidyverse Function . . . . .	45
10.1.3	From The Web . . . . .	45
10.1.4	Other File Types . . . . .	46
10.2	Checking Data . . . . .	46
10.2.1	The Whole Table . . . . .	46
10.2.2	Individual Columns . . . . .	47
10.2.3	Factors: The Pseudo Data Type . . . . .	48
<b>11</b>	<b>Wrangling Data</b>	<b>49</b>
11.1	Adding Columns . . . . .	49
11.1.1	Base . . . . .	49
11.1.2	tidyverse . . . . .	50
11.2	Pivoting / Reshaping . . . . .	51
11.2.1	Wide to Long . . . . .	51
11.2.2	Long to Wide . . . . .	52

11.2.3	Base . . . . .	52
11.3	String Manipulation . . . . .	52
11.3.1	Base . . . . .	52
11.3.2	tidyverse . . . . .	52
<b>12</b>	<b>Summarizing Data</b>	<b>53</b>
12.1	Describing With Summaries . . . . .	54
12.1.1	tidyverse . . . . .	54
12.1.2	Base . . . . .	56
12.1.3	Making Nice Tables . . . . .	57
12.2	Community Ecology . . . . .	57
12.2.1	Richness . . . . .	57
12.2.2	Diversity . . . . .	58
12.3	Related Topic: Joining Data . . . . .	58
<b>13</b>	<b>Analyzing Data</b>	<b>60</b>
13.1	Making Comparisons . . . . .	61
13.1.1	t-tests . . . . .	61
13.1.2	ANOVA . . . . .	62
13.2	Assessing Relationships . . . . .	63
13.2.1	Correlation . . . . .	63
13.2.2	Linear Regression . . . . .	64
13.2.3	Binomial Regression . . . . .	65
13.2.4	Poisson / Negative Binomial Regression . . . . .	66
13.3	Multivariate Analysis . . . . .	66
<b>14</b>	<b>Visualizing Data</b>	<b>69</b>
	<b>References</b>	<b>70</b>

# Welcome!

This website / HTML book is intended to collect resources for Cedar Creek summer interns doing independent research projects, and present those resources in an easily accessible way.

For now, the focus is primarily on data analysis and using the R programming language.

Much of the content featured is adapted from the work of past CCESR Fellows, including Mariana Cardenas and Bea Baselga.

This site is structured in different parts, which can be read in any order you choose, depending on your needs / what you already know. Currently, the first part goes over data analysis in general, the second part describes R-related software and workflows, and the third part is intended to give a primer in R coding.

**Part I**

**Data Analysis**

# 1 Data Analysis at a Glance

Analyzing your data is usually about transforming long spreadsheets into a form that is relevant to your question/s, and oftentimes including an appropriate statistical approach for inference.

You might use **descriptive statistics**, which is simply *describing* what you observed without presenting every data point, and instead a summary of those data. This can often be helpful in providing a frame of reference to your dataset before looking deeper at trends and comparisons. Alternatively, sometimes descriptive statistics are the main goal - like in surveys of populations and communities (e.g., what is the population size of a certain grass of interest in an old field?). Descriptive statistics include things like the mean and variance, but can also include more niche measures like dispersion.

You could also use **inferential statistics**, which is more about using math or simulation techniques to *infer* some conclusion from the shape of your data. This is directly relevant to when you have an ecological question about cause and effect, associations among variables, comparisons among categories, etc. The results of inferential statistics provide a starting point from which to interpret/discuss an answer to your question. Examples include t-tests and linear regression.

When using both of these types of statistics, you should be mindful of **data types**, which are the form that variables take. For example, the height of a tree is number, but the species of a tree is a category. This contrast is obvious, but there are subtle differences that can be important for how you describe, assess, and plot your data.

## 2 Data Types

First, let's go over different types of data:

### 2.1 Numeric Data

Any data that can be described with numbers or have quantifiable relationships between values is numeric. But! There are multiple types of numeric data. The most important distinction is **discrete** vs **continuous**.

**Discrete** numeric data is data where not every value is possible, but you can still quantify specific differences among the possible values - the major example being integer values (1, 2, 3, the rest). Most programming languages will refer to this type as integer or int. Examples might include number of ants on a log.

**Continuous** numeric data is data where every value is possible! So this is basically all real numbers, including decimals (1.0, 1.1, etc.). Many programming languages will refer to this as simply numeric data, but lower level languages might use “float” or “double”. Examples might include the biomass of ants on a log. Note: measures that consist of very large integer values are approximately continuous.

Other things to consider with numeric data is whether the scale of measurement is bound by any values. For example, the number of or biomass of ants on a log cannot be less than zero. In addition, percentages and proportions are bound by 0 and 100 and 0 and 1 respectively. These limitations can lead to special considerations when performing inferential statistics.

### 2.2 Categorical Data

Any data for which the values have no specifically quantitative difference among them is categorical. Again there is one majorly important distinction: **nominal** vs **ordinal**.

**Nominal** data is data where categories have no ranking or order, like the species of ants on a log.

**Ordinal** data is data where categories have some order, like your top 5 favorite breakfast cereals. But wait! You may be thinking - “isn't this quantitative?” Well yes and no. The difference between ordinal data and discrete numeric data is that you can't really quantify the



exact difference between ordinal data values. Say there is a go-kart race between Mario, Luigi, and Peach. The place that each finished would be ordinal, e.g., Peach got 1st and Luigi 2nd, but you wouldn't be able to say how much faster Peach was than Luigi. The time it took for Peach and Luigi each to finish the race would be a numeric variable, and there would be a specific value difference between them.

## 3 Descriptive Statistics

Now, let's discuss how to describe your data:

### 3.1 Centrality

You'll often want to describe the central tendency of your data - around where are the values centered?

**Mean** - the average of the values, or the sum of all values divided by the number of observations

**Median** - the value at which half of the observations are greater, and the other half are less

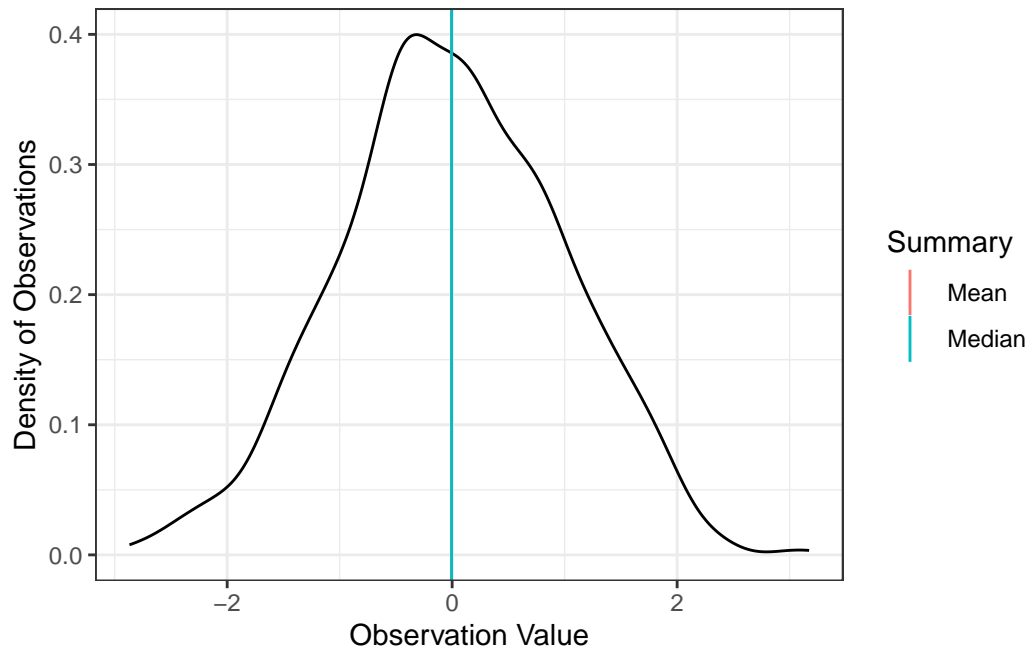
**Mode** - the most commonly observe value

Usually, the mean is a a perfectly adequate descriptor. You can use it on continuous numeric data, discrete numeric data (though the mean value will often be unrealistic), or even ordinal rankings.

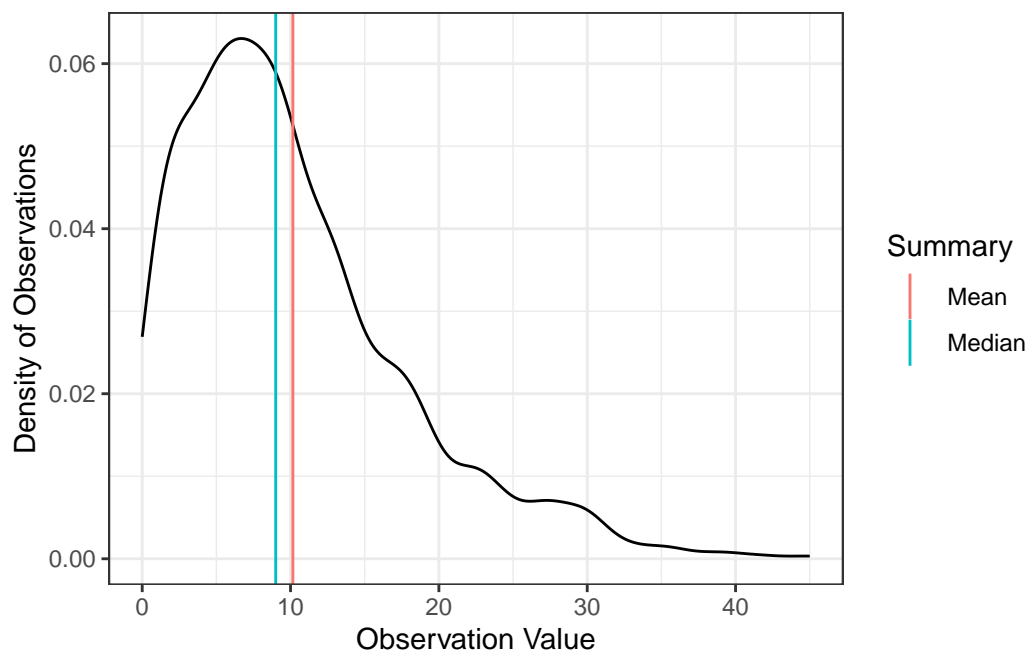
#### **When might you prefer to use the median over the mean?**

When the data is skewed such that there are many small values and a few big values, the mean might be inflated by those large values, and thus overestimate the central tendency in some contexts.

When data is roughly normally distributed, the mean and median are roughly the same:



But when data are skewed, the median may be a better estimate of the central tendency:



## 3.2 Spread

You also might be interested in how varied your data is, how much it deviates from the central tendency. This can be done with the following:

**Variance** - how variable is the data? Measured as the average squared difference between observations and the mean:

$$Variance = \frac{\sum (Observation_i - Mean)^2}{Number of Observations}$$

( $\sum$  means “sum of”)

The differences are squared to get rid of negative differences, because otherwise everything would cancel out and our variance would be zero!

**Standard Deviation** - the square root of the variance. This is useful because it is in the same units as the original measurements!

## 3.3 Other Descriptors

Another descriptor that may prove useful is the **dispersion**, or the variance divided by the mean. This provides an estimate of how skewed the data is - for example, the first plot above has very low dispersion, while the second plot has high dispersion.

## 3.4 Ecological Community Descriptors

Many of you are interested in describing the species composition of a community. Here's a few common descriptors:

**Species Richness** - this is just the number of different species present.

**Species Diversity** - this is an index that takes into account the richness as well as the relative abundances of each species. E.g. Shannon's Diversity Index, where higher numbers mean more species more evenly distributed.

**Species Evenness** - this is an index that estimates specifically how evenly distributed species abundances are. E.g., Pielou's Evenness, which ranges from 0 to 1, with 1 meaning that each species has equal numbers.

Note: these measures can apply to any taxonomic distinction, e.g., family richness, order diversity, etc.

## 4 Inferential Statistics

Now, let's think about how to use your data to answer your questions. There are a couple approaches statisticians use, and we will talk about frequentist statistics, where probabilities are thought of like relative frequencies. There is also Bayesian statistics, which is a bit more complex, so we will skip it for now.

Within frequentist statistics, we can run various tests to see how variables are related, which typically make some assumptions about the data. We can also do something called bootstrapping, which makes no assumptions, but can be simplistic from some perspectives.

In any case, we are hoping to estimate two main values:

**Effect Size:** How related are two variables? How different are two means? How much does one variable affect another?

**p-value:** What is the chance of observing data like yours (or something more extreme) if there was no relationship among the variables?

The p-value can be a bit tricky, but know that it **isn't** the probability that there is no relationship. P-values are used to draw conclusions from test results, a traditional guideline is that if the p-value is less than 0.05, the results are "significant." Some statisticians bristle at this arbitrary and binary system, so it's often best to report both the effect size and the actual p-value, so readers can interpret for themselves. The smaller the effect size, the weaker the relationship, the smaller the p-value, the stronger the evidence for the relationship.

Obviously, there are entire classes taught on this stuff (which may have taken or will take!), but we are thinking in just the basics for now.

### 4.1 Classic Frequentist Tests

Now let's go over some statistical tests! For this section, it can be useful to remind ourselves of the variables involved in a research question:

**Independent / Explanatory / Predictor Variable:** this is either what you are manipulating in an experiment or what your study is designed to capture variation in (e.g., CO<sub>2</sub> at BioCON, species richness at BigBio).

**Dependent / Response Variable:** these are what you measure or observe throughout your study, generally hypothesizing that they will differ among the levels of your independent variable (e.g., aboveground biomass in BioCON or BigBio).

#### 4.1.1 Assumptions

We should mention what these tests generally assume about your data.

First, they assume that your **data are independent**. This just means that no two observations of your data are more related to each other in a way that isn't accounted for by a variable. Say you were comparing mean tree height between two forests - individual tree heights in the same forest would be independent, but two measures of the same tree on different days would be non-independent.

Second, they assume that the **errors are normally distributed**. This is a bit more confusing without a statistical background. An example may be illustrative - in the tree height example above, we assume that the individual tree height are normally distributed about the mean. Without getting too much into the weeds, if you collect enough data (i.e., 30+ observations), these errors will likely be approximately normally distributed. However, things get dicey when we deal with data that is not continuous like tree height, for example, discrete count data - more on that below.

Third, they assume **homogeneity of variance**. This is another complicated one, but it means that the variance of the errors doesn't change with the independent variable. In the tree example, we are assuming that the variance of the differences between observed tree heights and the forest mean does not change between forests.

Data that break the first assumption are difficult to deal with outside of accounting for the non-independence factor (which can severely reduce the size of your sample), but failing to meet the second or third assumptions generally leads to transforming data or using alternative tests.

#### 4.1.2 Categorical Predictor/s, Numeric Response

##### 4.1.2.1 Two Predictor Categories

When you are comparing numeric values from two groups, you can use a **t-test** to compare their means. T-tests can be **paired** when each observation in one group is specifically linked to an observation in the other group (e.g., masses of sibling plants in separate treatments) which can be more powerful. When the variance of values in each group changes, you can do a **t-test with unequal variance**.

The effect size here is the difference between means.

#### 4.1.2.2 More Than Two Predictor Categories

If you have more than two groups/categories, you can use a **Analysis of Variance** or **ANOVA**. This will tell you if the means of each group are equivalent, or if there is at least one inequality. You can test for pairwise comparisons among the groups with **Tukey' test**. If you have multiple categorical predictors, you can do **two-way or three-way ANOVAs**. Tests with more than three categorical predictor variables are uncommon and harder to interpret.

The effect sizes are the pairwise difference in means.

#### 4.1.2.3 Ordinal Predictors

When your predictor variable is ordinal, the quick and easy way to analyze it would be to convert the predictor to a numeric integer data type and proceed from there. However this is imprecise...

*This section is under construction*

### 4.1.3 Numeric Predictor/s, Numeric Response

#### 4.1.3.1 Simple Association

When all you are interested in is whether two numeric variables are related to one another, not cause and effect, you can do a **correlation test**. **Pearson's correlation** is generally applicable for continuous data. **Spearman's correlation** is good for when you are dealing with data with non-normal distributions, like count data (it also works for ordinal data!).

The effect size here will be a correlation coefficient ranging from -1 to 1, with -1 means an inverse relationship, 0 means no relationship, and 1 mean a direct positive relationship.

#### 4.1.3.2 Cause and Effect

When you are suppose a causal relationship between numeric variables, you can use a **linear regression**. This will use linear algebra or maximum likelihood estimation (don't worry about it) to find the best fit line that describes the relationship between two variables; where the sum of the squared distances from the observations to the line is minimized. You can also include multiple predictor variables to perform **multiple linear regression** AKA **multivariate linear regression**.

When your response variable is count data, the assumptions of simple linear regression are usually unmet, so you can use generalized forms like a **Poisson regression** or a **Negative Binomial Regression**.

The effect sizes here are the parameter coefficients, i.e., how much does the response change for on unit increase in the predictor? Note: these are not straightforward for Poisson and negative binomial regression, so ask your mentor.

#### 4.1.4 Numeric Predictor/s, Categorical Response

##### 4.1.4.1 Binary Response

When your categorical response is only two categories (e.g., presence or absence), you can use a **binomial regression** AKA **logistic regression**. This works similarly to linear regression, but the effect sizes are measured in log odds, which is difficult to interpret, but can be transformed to estimating how the probability of one category value over the other increases with a variable.

##### 4.1.4.2 Multiple Response Categories

**Multinomial regression** (*under construction*)

#### 4.1.5 Categorical Predictor/s, Categorical Response

**Chi-square test** (*under construction*)

### 4.2 Bootstrapping

One alternative to these classic tests has no assumptions: bootstrapping. Essentially, it involves using the sampled data to simulate more samples, and compare your observations to those simulations.

**Empirical Bootstrapping** is where you take your actual observations and shuffle which value is associated with which observation. For example, you could take measurements of tree heights from two forests, and randomly assign forest ID to each measurement.

**Parametric Bootstrapping** is where you summarize your observed data and use it to generate simulated data. For example, you could calculate the mean and variance of tree heights in two forests and then generate simulated forests of trees through random pulls from a normal distribution with the appropriate mean and variance.

With both approaches, you simulate a large number of simulated datasets (1000+), and then calculate whatever you are interested in for each of those simulations, and compare the calculation from the observed data to the distribution of simulated values. For example, if you empirically bootstrap the two forests of tree heights 1000 times, and then calculate difference



in means for each you will have 1000 mean difference values. The proportion of those simulated values that are equal to or more extreme than your observed mean difference is your p-value!

# **Part II**

## **R on your Computer**

## 5 R Itself

R is both a programming language and an application that you can install to your computer.

### 5.1 R, the Language

R is a programming language designed for statistical computing, and is often the language of choice for scientists. R is also used for data science in some business, tech, and health contexts ( but many prefer Python in those areas).

As a programming language it is essentially an expandable collection of functions with syntax to perform tasks, and it could be written in any text editor. However, in order for your computer to interpret the language, it needs some software.

### 5.2 R, the Software

The R application allows you to run R code on your computer, and comes with a basic “console” window where code is run and output is printed, as well as a basic script editor where you can write code to run.

You can download the application from here:

<https://cran.r-project.org/>

If you are asked to select a mirror, simply select the nearest one (I believe Iowa State should work).

If you have a Windows machine, it should be fairly straightforward to simply download and install the “base” R from the link.

If you have a Mac, you will want to select the .pkg file that matches your processor type: x-86 for Intel processors (mostly Macs pre-2020), arm64 for Macs with the M1 or M2 chip (most Macs post-2020).

If you are using Linux, you know more than me.

## 5.3 R Packages

As mentioned above, R is *expandable*. You can add more functionality to R by installing packages. Packages contain more options of code to use to process and analyze data, and also do many other things.

Packages can be installed through writing R code, or by clicking some buttons in RStudio. Then they will live in a directory that was built when you installed R for auxiliary packages.

We will discuss more about installing packages in the R coding section.

## 6 R Studio

While you can use R with just the basic application, it is much easier and beginner-friendly to use RStudio, which is an integrated development environment or IDE. This is just an application that provides a suite of features to make programming easier for users. In fact, I'm typing this in RStudio *right now*!! Note: you must have the R application installed to use RStudio, as it relies on the R application to interpret R code.

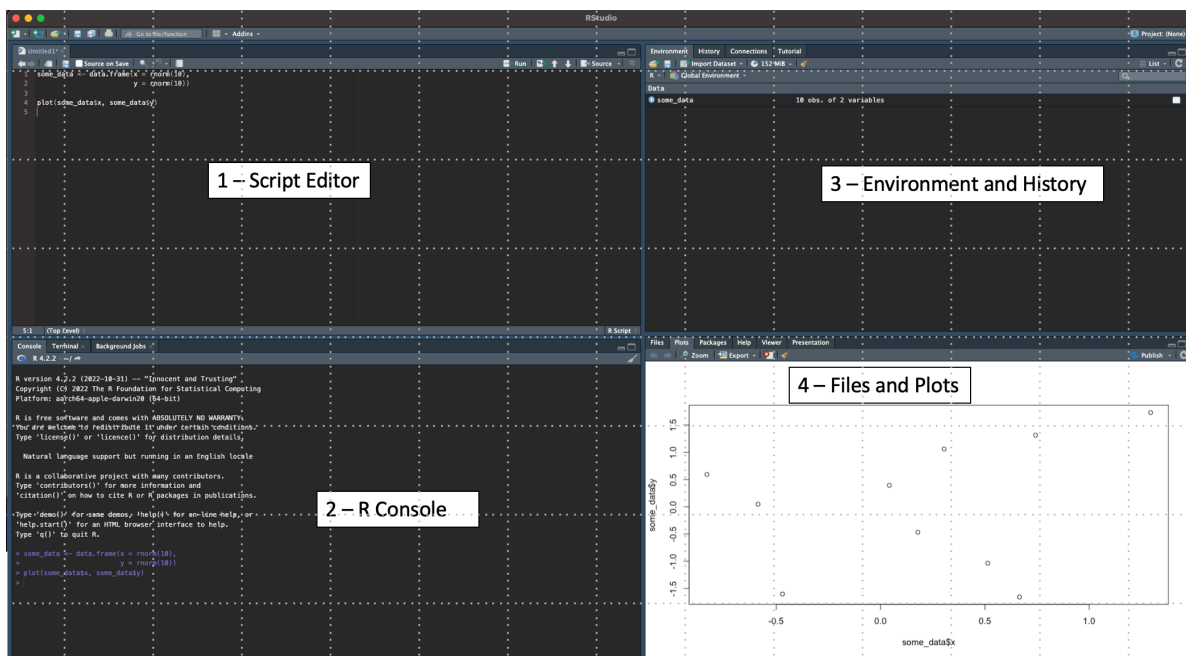
You can download and install RStudio from here:

<https://posit.co/download/rstudio-desktop/>

Which should be more straightforward than downloading and installing R.

### 6.1 RStudio at a Glance

If you open up RStudio, you will see something like this:



**1- Script Editor:** Here is where you will write code! You can create an R script (a text document to save code in) with the file tab, and write what you need in the resulting script. It is **highly** recommended to use scripts, because then you can save your code for later, and troubleshoot errors easier. From this window, you can highlight code and run it with the “Run” button on top, or with Ctrl + Enter / Cmd +Enter.

**2- R Console:** Here is where the action happens - code will run here, and text output, warnings and messages will be displayed. You can also type code into the console, but that is only recommended for installing packages, entering credentials, rendering documents, and things of that nature. Don’t type your data processing or analysis code into the console, use a script instead! There’s also a terminal tab if you ever need to perform shell commands.

**3- Environment and History:** Here you can find a list of the variables and data you have loaded into your “workspace” or “environment” in the Environment tab. These are objects you can do stuff with with code. You can also click the History tab to see the code you have run thus far.

**4- Files and Plots:** Here is where any figures you draw will pop up (and you can save them from here as well). There is also a Files tab that allows you to navigate through your file directory (helpful with projects, described below). The Packages tab shows which packages you have installed and loaded (you can also click “Install” at top to easily install new ones!). Finally, the help tab is where you can search for the documentation on any R function.

## 6.2 R Projects

It is highly recommended to use R Projects when working with RStudio. Projects are essentially just subdirectories in your file folders, but they come with a special .Rproj file that RStudio can read and use. This helps you organize your work, and makes your code more easily portable.

You can create a new projects from the File tab at upper left, or in the project dropdown menu at upper right. You can just create one in a new directory. Then you can select a name and where you want to save it.

There are many different types of projects - this book/website is one!

If you want to backup your work with version control or collaborate with others using git and GitHub, you will need to use projects. (Well, technically you don’t need to, but you’d be doing many things manually).

## 7 Optional: Git and Github

If you are interested in:

1. Backing up your code using a version control system that allows you to roll back changes and monitor incremental progress

*and/or*

2. Sharing your code and collaborating with others

You may like to try using git (a program for your computer) and GitHub (a website that hosts code projects).

We won't go into detail here, but Jenny Bryan's excellent introduction and tutorial on the topic can be found here:

<https://happygitwithr.com/>

# **Part III**

## **R Programming**



# 8 The Basics

## 8.1 Intro

I'm sure those of you reading this come from a wide variety of backgrounds regarding computer programming - some of you may be very familiar with it, others total novices. Some of you may love computing, others might hate it. If you're apprehensive about learning R, or if you find yourself struggling with it - don't worry! Scientific computing presents a challenge at some point to everyone who does it. Just remember a few things:

1. Everyone makes mistakes.
2. Don't be afraid to ask questions!
3. Don't compare yourself to others, compare you today to you yesterday.
4. Everyone is constantly learning new things, including those who seem like experts.

That said, learning a programming language is a little like learning a human language, except there's a much smaller vocabulary and the grammar is *very* strict. And where human language has parts of speech like nouns and verbs, R has a certain syntax as well. Some of the main components of the R language are **operators**, **functions**, **arguments**, and **data**.

## 8.2 Operators

Operators are short symbols that tell the computer to do certain simple things. You are already familiar with many operators - the **math operators** like `+`, `-`, `*`, and `/`. R at its simplest is a calculator:

```
## This is block of R code! Anything that starts with # is a comment, and doesn't run.  
  
## adding  
2 + 2
```

```
[1] 4
```

```
## subtracting  
5 - 4
```

```
[1] 1
```

```
## multiplying  
3 * 3
```

```
[1] 9
```

```
## dividing  
6 / 2
```

```
[1] 3
```

There are a couple other math operators too:

```
## exponentiate with ^  
3^2
```

```
[1] 9
```

```
## find the remainder with the modulus, %%  
10 %% 3
```

```
[1] 1
```

```
## perform integer division with %/%  
10 %/% 3
```

```
[1] 3
```

But math operators aren't the only type! There are also the closely related **comparison operators**, which will return TRUE or FALSE instead of calculated numbers:

```
## equals, ==  
2 + 2 == 4
```

```
[1] TRUE
```

```
## does not equal, !=  
2 + 2 != 4
```

```
[1] FALSE
```

```
## greater than, >  
5 > 4
```

```
[1] TRUE
```

```
## less than, <  
5 < 4
```

```
[1] FALSE
```

There are also greater than or equal to ( $\geq$ ) and less than or equal to ( $\leq$ ).

You can combine comparisons with **logical operators** - and ( $\&$ ), or ( $\mid$ ), and not ( $!$ ):

```
## and: are both true?  
(3 > 2) & (4 > 3)
```

```
[1] TRUE
```

```
## or: is at least one true?  
(2 == 1) | (4 < 3)
```

```
[1] FALSE
```

```
## not: is this false?  
!(2 == 1)
```

```
[1] TRUE
```

There are few other important operators, but they will make more sense once we talk about the other parts of R.

## 8.3 Functions

Functions are words (though not necessarily real words) or letters that instruct the computer to perform more complicated tasks. They generally are followed by parentheses ().

```
## here's a function that returns the current date  
Sys.Date()
```

```
[1] "2023-07-25"
```

```
## and here is a function that returns the date with the time  
Sys.time()
```

```
[1] "2023-07-25 12:44:13 CDT"
```

No you may be thinking - “this is pretty basic” and “what are the parentheses for?”, which brings use to arguments!

## 8.4 Arguments

Arguments are values or objects that go inside the parentheses of functions to specify what you want the function to do. This is what gives functions their power. Arguments are separated inside a function by commas.

```
## the sum function can sum many numbers  
sum(1,2,3,4,5)
```

```
[1] 15
```

In the function above, each number is acting as an argument. In this case, the arguments don't have names. Oftentimes a function's arguments will be explicitly named, and to specify what you want those arguments to be, you use the = operator.

```
## this function pulls values randomly from a normal distribution specified in the arguments  
## n specifies how many numbers to return, and mean and sd specify shape of the distribution  
rnorm(n = 10, mean = 5, sd = 1)
```

```
[1] 5.520307 5.324284 6.935449 5.147405 6.310083 6.831964 3.502630 3.750079  
[9] 4.381058 3.033246
```

Operators are actually a special type of function that can be used with syntax that is more intuitive for them. You can also use them in the same way as most functions by surrounding them with back ticks, `.

```
## here we use the + operator in a much more confusing context  
`+`(2, 2)
```

```
[1] 4
```

```
## it is equivalent to  
2 + 2
```

```
[1] 4
```

## 8.5 Data

We are using the word data here to broadly encompass **values** (like the numbers we were using above, both with operators and as arguments), **variables** (stored values), and **data structures** (organized collections of values).

### 8.5.1 Values

Values are much like the data types we discuss in the data analysis section. In fact, the different types of values R can deal with are called data types as well!

In R, values can be numeric, character, or logical (among other, more specific types).

```
## numeric values are numbers!
```

```
2
```

```
[1] 2
```

```
2.5
```

```
[1] 2.5
```

```
## character values are letters, words, phrases (often referred to as "strings")
```

```
"a"
```

```
[1] "a"
```

```
"apple"
```

```
[1] "apple"
```

```
"there is a worm in my apple"
```

```
[1] "there is a worm in my apple"
```

```
## note: character values or strings must be surrounded by "" or '' for R to interpret the
```

```
## Logical values are TRUE or FALSE (you've seen these above)
```

```
TRUE
```

```
[1] TRUE
```

```
FALSE
```

```
[1] FALSE
```

There are other types of values too: missing values (NA and NaN), infinite values (Inf and -Inf), and something that indicates empty (NULL).

## 8.5.2 Variables

Variables are named values that are stored in the “environment”, or the workspace that R can access to perform its tasks. In order to store a value as a variable, you need to use a special kind of operator called an **assignment operator** (<- or =). As I mentioned variables have names, which are unquoted text.

```
## store 2 as a variable called x  
x <- 2  
  
## R returns no output here because you're just storing a value  
## but you can return the value by calling the variable  
x
```

[1] 2

```
## store 3 as a variable called y  
y <- 3  
  
## you use variables with operators  
x + y
```

[1] 5

```
## store a character value  
string <- "hello"  
  
## math doesn't work on strings
```

Technically, you can use = in place of <-. This is why the equals operator is ==. I generally use <- to prevent any confusion between assignment and comparison.

### 8.5.2.1 Naming Rules

Variables have rules about how they can be named:

1. No special symbols other than `_` and `.`
2. You can't start with a number or `_`.
3. They can't be special words that R interprets differently. You can enter `?Reserved` in your console to see a list.

### 8.5.3 Data Structures

Data structures are collections of values with some sort of organization, and also saved in the environment. Plot twist: the variables above are a the simplest data structure, the **scalar**, which is just a single value.

The next data structure is the **vector**, which is a collection of values of the *same data type*. We can store them much like variables.

```
## we use another operator, :, to create a sequence of integers from 1 to 10  
my_vector <- 1:5
```

```
my_vector
```

```
[1] 1 2 3 4 5
```

```
## you can also create vectors with the combine function, c()  
my_other_vector <- c("a", "b", "c")
```

```
my_other_vector
```

```
[1] "a" "b" "c"
```

The next data structure is called a **list**. A list is a collection of values like a vector, but they can be of any data type, or data structure. You can have a list of numeric values and character values, a list of vectors, or even a lists of lists! Every other complex data structure is technically a list with special attributes and/or rules.

```
## you can create lists with the list function  
my_list <- list("a", 1, 2:4)
```

```
my_list
```

```
[[1]]
```

```
[1] "a"
```

```
[[2]]
```

```
[1] 1
```

```
[[3]]
```

```
[1] 2 3 4
```



```
## can also use the combine function, but it will default to a vector when data types are  
my_other_list <- c("b", 2)
```

Finally, the most common special type of list you will use is the **data frame**. A data frame is a list of vectors that are arranged in a table, much like an excel spreadsheet. Each of the vectors will be named as a column, and all must be the same length. The position of a value in a vector is its row in the data frame.

```
## we can make a data frame with the data.frame function  
my_data <- data.frame(letter = c("a","b","c"), # each column has a name  
                      number = c(1, 2, 3),  
                      vowel = c(TRUE, FALSE, FALSE))
```

```
my_data
```

	letter	number	vowel
1	a	1	TRUE
2	b	2	FALSE
3	c	3	FALSE

Next, we will extend these concepts a bit further!

## 9 Next Steps

Now that we have the basic “parts of speech” of R down, we can move toward what we can do with them.

What follows are a few unconnected topics that will prove as useful background to working with data in the later sections.

### 9.1 Packages

Packages are collections of R functions that people write to make tasks easier. One of the strengths of R is that countless programmers have taken the time to assemble functions of use in their respective fields, and shared them with the world. For example the “vegan” package contains a number of functions geared towards community ecology, like calculating diversity indices. You could calculate a diversity index with just the base R, but it would be more difficult and take longer.

You can install packages in at least two ways:

1. You can use the following code, with the package names in quotes (this is one of the few times where using the console is recommended, because you only need to install a package once):

```
install.packages("PACKAGE NAME HERE")
```

2. Or you can use the packages tab in RStudio. In the lower right panel, there should be a packages tab in between “Plots” and “Help”. Once there, there is an “Install” button. When clicked a window will appear allowing you to search for packages to install.

But Installing packages does not make them automatically accessible to you. When R boots up, it only loads its base functionality by default, so you have to load any packages that you want to use for a given R session. You can do this with the following code (with the package name not in quotes):

```
library(PACKAGE NAME)
```

The code for loading packages should be saved in your `r` script, because it will need to be done every time you open R.

There is a family of packages that is very popular called the “tidyverse.” The aim of the tidyverse is to make data manipulation and visualization streamlined and efficient. Some people are very opinionated about whether you should use the tidyverse or base R, but in my opinion, it’s mostly silliness. If you only want to dip into R and don’t plan to use it much in the future, you may as well just pick up the specific functions you need to use and not worry about much else. If you’d like to continually use R for data analysis, but don’t plan on getting deep into it, getting a handle on the tidyverse may be a good idea. If you want to really get into R, I would recommend learning how to do things in base R (as well as tidyverse functions).

You can install the tidyverse suite with:

```
install.packages("tidyverse")
```

## 9.2 Subsetting

In the last section we introduced data structures. Now let’s talk about what you can do with them.

### 9.2.1 Vectors

The individual elements of a vector can be accessed with bracket operators - `[` and `]`. You can refer to an element by its index, or its numeric place in the sequence of elements (e.g., the 1st, the 10th, etc.). It’s important to note here that R starts counting at 1, while many other programming language start counting at 0 (e.g., Python). This is another thing that people are opinionated about, and if you put your mind to it, *you can be too!* Anyway, here are some examples:

```
## let's create a vector of the first five letters of the alphabet
my_vector <- c("a","b","c","d","e")
my_vector
```

```
[1] "a" "b" "c" "d" "e"
```

```
## now let's return the 5th element
my_vector[5]
```

```
[1] "e"
```

```
## we can return multiple elements with c()
my_vector[c(2,4)]
```

```
[1] "b" "d"
```

```
## or as a series with :
my_vector[2:4]
```

```
[1] "b" "c" "d"
```

You can also use negative numbers to exclude values from what's returned:

```
## lose the last element
my_vector[-5]
```

```
[1] "a" "b" "c" "d"
```

```
## everything but the last element
my_vector[-1:-4]
```

```
[1] "e"
```

### 9.2.2 Lists

Subsetting vectors is fairly straightforward, but subsetting lists can be tricky. Since lists have multiple levels of organization, they use both the `[]` operators and the `[[[]]` operators. Single brackets give you the list element, and double brackets give you *what the list element contains*. Let's demonstrate:

```
## create a list
my_list <- list(c("a","b","c"), "d", "e")
my_list
```

```
[[1]]  
[1] "a" "b" "c"
```

```
[[2]]  
[1] "d"
```

```
[[3]]  
[1] "e"
```

```
## grab the first list element  
my_list[1]
```

```
[[1]]  
[1] "a" "b" "c"
```

```
## grab what's contained in the first list element (in this case a vector)  
my_list[[1]]
```

```
[1] "a" "b" "c"
```

```
## another example with a scalar  
my_list[2]
```

```
[[1]]  
[1] "d"
```

```
my_list[[2]]
```

```
[1] "d"
```

```
## you can also subset what you have subsetted:  
my_list[[1]][1]
```

```
[1] "a"
```

```
## but if you try subsetting a list element, it won't work the same way  
my_list[1][1]
```

```
[[1]]  
[1] "a" "b" "c"
```

```
## this is because [] returns the list element as a list of length 1, therefore [1] gives
```

This distinction can be difficult to understand, but don't worry! It takes time. The best analogy I've seen is from Hadley Wickham here:

<https://adv-r.hadley.nz/subsetting.html#subset-single>

You can think of a list as a train, every list element is a train car, and each has its own contents. Single brackets give you the train car/s, and double brackets gives you what's inside a single train car. And even a single train car is a train (or a list). Also note:

```
## you can grab multiple list elements with []; this give a list with two elements  
my_list[1:2]
```

```
[[1]]  
[1] "a" "b" "c"
```

```
[[2]]  
[1] "d"
```

```
## list elements can be named and indexed by their name as well  
named_list <- list(first = 1:3, second = 10)  
named_list
```

```
$first  
[1] 1 2 3
```

```
$second  
[1] 10
```

```
named_list["first"]
```

```
$first  
[1] 1 2 3
```

### 9.2.3 Data Frames

Subsetting data frames is a little easier to get a handle on, you just need to think in two dimensions. When using single brackets to subset data frames, you need to specify the index of the row and the column separately and in that order. You separate each index number by a comma inside the brackets. Check it out:

```
## create data frame
my_data <- data.frame(letter = c("a","b","c"), # each column has a name
                      number = c(1, 2, 3),
                      vowel = c(TRUE, FALSE, FALSE))

my_data
```

	letter	number	vowel
1	a	1	TRUE
2	b	2	FALSE
3	c	3	FALSE

```
## grab the element in the 2nd row, 1st column
my_data[2,1]
```

```
[1] "b"
```

```
## you can also grab a whole row or column by leaving one side of the comma blank
my_data[2,]
```

	letter	number	vowel
2	b	2	FALSE

```
my_data[,1]
```

```
[1] "a" "b" "c"
```

```
## (subsetting a row gives you a data frame, subsetting a column gives you a vector)
```

But data frames also have named columns! Let's use that to our advantage. You can specify a column's name instead of its index in brackets, like for a list, or you can use the \$ operator.

```
## subsetting by name in brackets  
my_data[, "vowel"]
```

```
[1] TRUE FALSE FALSE
```

```
## subsetting by name with $ (notice no quotes)  
my_data$vowel
```

```
[1] TRUE FALSE FALSE
```

```
## the downside of $ is that you can't grab more than one column like with brackets  
my_data[, c("letter", "vowel")]
```

```
letter vowel  
1      a  TRUE  
2      b FALSE  
3      c FALSE
```

```
## subsetting multiple columns gives you a data.frame
```

```
## you can use $ with named lists too  
named_list$first
```

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

```
## you can mix subsetting operators if you ever need to  
my_data$vowel[1]
```

```
[1] TRUE
```

```
my_data[1,]$vowel
```

```
[1] TRUE
```

You can also use brackets to select rows by value, not index. You just need to use some comparison operator in a statement that resolves as TRUE or FALSE.



```
## grab the consonant rows
my_data[my_data$vowel == FALSE,]
```

```
letter number vowel
2      b         2 FALSE
3      c         3 FALSE
```

```
## grab the rows before the third
my_data[my_data$number < 3,]
```

```
letter number vowel
1      a         1  TRUE
2      b         2 FALSE
```

```
## you can combine criteria
my_data[my_data$number < 3 & my_data$vowel == FALSE,]
```

```
letter number vowel
2      b         2 FALSE
```

Now the reason that we talked about packages in between data structures and subsetting is because the tidyverse (specifically, the dplyr package) has more functions for subsetting: `filter` and `select`. `Filter` works much like grabbing rows by value, and `select` works like grabbing columns by name. Let's look at some examples:

```
## load the tidyverse
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
```

```
v dplyr      1.1.1      v readr      2.1.4
v forcats   1.0.0      v stringr   1.5.0
v ggplot2   3.4.2      v tibble    3.2.1
v lubridate 1.9.2      v tidyr     1.3.0
v purrr     1.0.1
```

```
-- Conflicts ----- tidyverse_conflicts() --
```

```
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
```

```
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
## filter for consonants  
filter(.data = my_data, vowel == FALSE)
```

```
letter number vowel  
1      b      2 FALSE  
2      c      3 FALSE
```

```
## select letter related columns  
select(.data = my_data, letter, vowel)
```

```
letter vowel  
1      a  TRUE  
2      b FALSE  
3      c FALSE
```

```
## you can also exclude columns  
select(.data = my_data, !number)
```

```
letter vowel  
1      a  TRUE  
2      b FALSE  
3      c FALSE
```

```
## note: selecting a single column will give a data frame, not a vector  
select(.data = my_data, number)
```

```
number  
1      1  
2      2  
3      3
```

```
## another tidyverse/dplyr function, pull, will give just a vector  
pull(.data = my_data, number)
```

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

As you can see, filter, select and pull are versatile, consistent and powerful. However, they lack one key ability: assignment. You can use brackets and \$s to assign things:

```
## assign a new value to a data element (NA means missing value)
my_data[3,2] <- NA
my_data
```

	letter	number	vowel
1	a	1	TRUE
2	b	2	FALSE
3	c	NA	FALSE

```
## create a whole new column with $ (vector must be of same length as the number of rows)
my_data$new_column <- c("some", "new", "data")
my_data
```

	letter	number	vowel	new_column
1	a	1	TRUE	some
2	b	2	FALSE	new
3	c	NA	FALSE	data

## 9.3 Optional: Flow Control

## 9.4 Optional: Writing Functions

# 10 Importing Data

The previous chapters focused on R’s functionality, but you might not be feeling any closer to working with *your* data in R. We’ll transition to that now! The first step is getting your data into your R environment, so that you can use it with R’s functions.

## 10.1 Reading Data

Importing data into R is often referred to as reading data, as that is what the computer is doing, it’s reading the contents of a file (usually a text file). Most ecologists and data scientists work with a text file called a Comma Separated Value file, or csv. This is a small file that’s easy for computers to read where each column is separated by a column, and each row by a new line. You can save excel files as csv from the “Save As...” menu, and you can specify csv as the type when downloading a Google sheet.

### 10.1.1 From Your Computer

So when you have files you want to read locally on your computer, the first thing you need to think about is what’s called the “working directory”. The working directory is the folder on your computer where R will look for files when prompted, and also where it will save output.

You can check your current working directory:

```
## return current working directory  
getwd()
```

```
[1] "/Users/kit/Documents/UMN/Research/cedar_creek_projects/ccesr_intern_hub"
```

You can also set your working directory manually

```
## change working directory  
setwd("some/different/folder")
```

Or, in RStudio, you can click the Session dropdown menu at the top of the window, then “Set Working Directory”, then “Choose Directory.”

If you use an R Project (**highly recommended**), you don't have to worry as much about this. If you have a project open, the working directory will be automatically set to the folder that contains the .Rproj file that is created when you create a project. See Section 6.2 for more info!

When you're in an R Project, or have a csv you want in your working directory, you can read it into your environment like so:

```
## read data
my_data <- read.csv("the name of your file in quotes", header = TRUE)
```

The read.csv function creates a data frame from the csv you specify, and then the <- assigns it to “my\_data.” The “header = TRUE” argument tells R to interpret the first line of the csv as the column names.

### 10.1.2 tidyverse Function

The readr package in the tidyverse family also has its own data reading functions.

```
## load tidyverse
library(tidyverse)

## read data (assumes header by default)
my_data <- read_csv("name of your data in quotes")
```

These functions are pretty similar, with one exception: read.csv gives you a data frame, but read\_csv gives you a “tibble.” What is a tibble? It's another special type of list, much like a data frame, but with a few differences. IT was designed to work more consistently with tidyverse functions. One important difference between data frames and tibbles that when you subset an individual column with the brackets ([]), data frames will give you vectors, and tibble will give single column tibbles. This has caused me confusion when writing functions, but you may not run into it.

### 10.1.3 From The Web

You can also read files directly from the web. If you have your data in Google sheets, you can create a URL for R to import it directly. Simply go to the File menu, click “Share” and then “Publish to web”. In the box that pops up, you will need to select the file type as “.csv”, not web page. Then save the URL that it gives you!

For demonstration, I've created a few data sheets that you too can import into R by copying the following code:

```
## put the url of the data in quotes
fake_mammals <- read.csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vQ9mfx88nM33PC6W

fake_insects <- read.csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vT0snHMdsxzzkxt
```

These two files will be used throughout the next chapters. The first is some made-up data of some mammal captures at 6 sites across forest and savanna habitats (with mass and parasite info), and the second is made up sweep-netting data from the same sites.

### 10.1.4 Other File Types

Under construction....

## 10.2 Checking Data

Now that you have data, you will want to look at it!

### 10.2.1 The Whole Table

You can look at a whole data frame by clicking on its name in the “Environment” pane in RStudio (upper right), or with the `View()` function:

```
View(fake_mammals)
```

You can also just look at parts:

```
## check top 6 rows
head(fake_mammals)
```

	site	site_type	species	mass_g	tick_count	helminth_mass_mg
1	a	forest	White-footed mouse	20	0	512
2	a	forest	White-footed mouse	24	10	365
3	a	forest	White-footed mouse	23	2	0
4	a	forest	White-footed mouse	19	0	608
5	a	forest	White-footed mouse	25	12	109
6	a	forest	Deer mouse	22	3	456

```
## check bottom 6 rows
tail(fake_mammals)
```

	site	site_type	species	mass_g	tick_count	helminth_mass_mg
43	f	savanna	White-footed mouse	21	0	408
44	f	savanna	White-footed mouse	25	1	197
45	f	savanna	White-footed mouse	24	0	152
46	f	savanna	Deer mouse	20	0	508
47	f	savanna	Deer mouse	22	2	496
48	f	savanna	Meadow vole	23	NA	56

You can also take a look at the structure of the data with `str()`, which will tell you how many rows (observations) and how many columns (variables), as well as the type of each column.

```
## check structure
str(fake_mammals)
```

```
'data.frame':  48 obs. of  6 variables:
 $ site          : chr  "a" "a" "a" "a" ...
 $ site_type     : chr  "forest" "forest" "forest" "forest" ...
 $ species       : chr  "White-footed mouse" "White-footed mouse" "White-footed mouse" "Wh...
 $ mass_g        : int   20 24 23 19 25 22 22 21 23 20 ...
 $ tick_count    : int    0 10 2 0 12 3 2 0 NA NA ...
 $ helminth_mass_mg: int   512 365 0 608 109 456 521 432 20 129 ...
```

## 10.2.2 Individual Columns

You can also take a look at individual columns with the `$` operator, and get quick summaries with `summary()`:

```
## summarize mammal masses
summary(fake_mammals$mass_g)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
17.00	20.75	23.00	22.60	25.00	28.00

```
## summarize helminth masses
summary(fake_mammals$helminth_mass_mg)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	119.5	341.5	306.5	501.2	713.0

Note that for the second summary (helminth mass), it tells you how many NA's, or missing values, there are.

### 10.2.3 Factors: The Pseudo Data Type

Unless you specify, csv reading functions will assume the data type of each column in a data sheet (numeric, character, etc.). Any categorical variable will be considered a character type generally. But the way character types are stored in computer memory does not lend itself well to statistical analysis. To remedy this, R has a special data type called the factor for categorical data. A factor is made up of two parts - the levels, which are stored to the computer as integers, and the labels, which are character strings that we can read as the category names.

In our mammal data, we probably want mammal species, site type, and site all to be factors. We can convert them with the `as.factor` function!

```
## convert species to factor
fake_mammals$species <- as.factor(fake_mammals$species)
## convert site type to factor
fake_mammals$site_type <- as.factor(fake_mammals$site_type)
## convert site to a factor
fake_mammals$site <- as.factor(fake_mammals$site)

## check it out!
summary(fake_mammals$species)
```

Deer mouse	Meadow vole	White-footed mouse
15	9	24

```
## also do for insects
fake_insects$site <- as.factor(fake_insects$site)
fake_insects$site_type <- as.factor(fake_insects$site_type)
```

There are also similar functions for converting data types to numeric (`as.numeric()`) and character (`as.character()`).



# 11 Wrangling Data

Now that we have our data in, let's play with it!

Setup:

```
## read data
fake_mammals <- read.csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vQ9mfx88nM33PC6W

fake_insects <- read.csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vT0snHMdsxzzzkxt

## convert species to factor
fake_mammals$species <- as.factor(fake_mammals$species)
## convert site type to factor
fake_mammals$site_type <- as.factor(fake_mammals$site_type)
## convert site to a factor
fake_mammals$site <- as.factor(fake_mammals$site)

## also do for insects
fake_insects$site <- as.factor(fake_insects$site)
fake_insects$site_type <- as.factor(fake_insects$site_type)
```

## 11.1 Adding Columns

One simple thing you may want to do is add columns to your data, which may be calculations from existing columns.

### 11.1.1 Base

In base R, we have already kinda done this. You can assign something to a new column with the \$ and <- operators.

For the insect data, let's say we wanted to calculate average temperature at a given site based on the recorded high (temp\_hi) and low (temp\_low):

```

## calculate mean temp
fake_insects$temp_mean <- (fake_insects$temp_hi + fake_insects$temp_lo)/2

## check it out
fake_insects$temp_mean

```

```
[1] 21.25 23.25 19.25 26.75 24.00 28.00
```

## 11.1.2 tidyverse

In the tidyverse, adding new columns is done with the mutate function:

```

## load tidyverse
library(tidyverse)

-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.1.1      v readr      2.1.4
v forcats    1.0.0      v stringr    1.5.0
v ggplot2    3.4.2      v tibble     3.2.1
v lubridate  1.9.2      v tidyr      1.3.0
v purrr      1.0.1
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become

```

```

## mutate a new column
fake_insects <- mutate(.data = fake_insects, ## specify data
                       temp_mean_mutated = (temp_hi + temp_lo)/2) ## calculate new column

## this column should be the same for all six rows (a TRUE should be returned for each)
fake_insects$temp_mean == fake_insects$temp_mean_mutated

```

```
[1] TRUE TRUE TRUE TRUE TRUE TRUE
```

As you can see, using mutate mean you have to write the name of data frame fewer times.

## 11.2 Pivoting / Reshaping

You also may need to transform your data between the wide and long formats (recall the data management slides, which will be added to this book eventually). I find that the pivot functions from `tidyr` in the `tidyverse` are easier to use, so we will go over those. But by all means, if you prefer base, go for it! I'm just less familiar with using base R for this.

### 11.2.1 Wide to Long

The insect data is partly in wide format: we have a column for each order of insect, where the count is implicitly the values in the cells. It will be easier to work with if we make one "count" column and one "order" column. We can do this with the `pivot_longer` function:

```
## lengthen the order count data
long_insects <- pivot_longer(data = fake_insects,
                             cols = c(hymenoptera, lepidoptera, coleoptera, diptera,
                                       odonata, hemiptera, orthoptera, ephemeroptera,
                                       tricoptera, plecoptera),
                             names_to = "order",
                             values_to = "count"
                             )

## check it out
head(long_insects)
```

```
# A tibble: 6 x 8
  site site_type temp_hi temp_lo temp_mean temp_mean_mutated order      count
<fct> <fct>      <dbl>  <dbl>    <dbl>          <dbl> <chr>      <int>
1 a     forest      23.5    19      21.2          21.2 hymenoptera    9
2 a     forest      23.5    19      21.2          21.2 lepidoptera    3
3 a     forest      23.5    19      21.2          21.2 coleoptera    16
4 a     forest      23.5    19      21.2          21.2 diptera      29
5 a     forest      23.5    19      21.2          21.2 odonata       4
6 a     forest      23.5    19      21.2          21.2 hemiptera    10
```

Now we have multiple rows for each site, one for each order! You may not believe me, but this will make things easier down the line.

For reference, the `cols` argument specifies which columns you want to pivot. The `names_to` argument names the column that will contain the pivoted column names, and the `values_to` argument names the column that the cell values will be placed.

### 11.2.2 Long to Wide

You'll occasionally want to turn long to wide as well. `pivot_wider` works for this

```
## widen our long data  
wide_insects <- pivot_wider(data = long_insects,  
                             names_from = "order",  
                             values_from = "count")
```

This is the inverse of what we just did - we made a column for each value in the column given to the `names_from` argument, the values of which are pulled from the column given to the `values_from` argument.

### 11.2.3 Base

You can pivot data in base R with `reshape`, but this section is under construction...

## 11.3 String Manipulation

Under construction....

### 11.3.1 Base

### 11.3.2 tidyverse

## 12 Summarizing Data

Now let's describe our data!

Setup:

```
## load tidyverse
library(tidyverse)

## read data
fake_mammals <- read.csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vQ9mfx88nM33PC6W

fake_insects <- read.csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vT0snHMdsxzzzkxt

## convert species to factor
fake_mammals$species <- as.factor(fake_mammals$species)
## convert site type to factor
fake_mammals$site_type <- as.factor(fake_mammals$site_type)
## convert site to a factor
fake_mammals$site <- as.factor(fake_mammals$site)

## also do for insects
fake_insects$site <- as.factor(fake_insects$site)
fake_insects$site_type <- as.factor(fake_insects$site_type)

## lengthen the order count data
long_insects <- pivot_longer(data = fake_insects,
                             cols = c(hymenoptera, lepidoptera, coleoptera, diptera,
                                       odonata, hemiptera, orthoptera, ephemeroptera,
                                       tricoptera, plecoptera),
                             names_to = "order",
                             values_to = "count"
                             )
```

## 12.1 Describing With Summaries

You'll often want to give simple, illustrative information about the data you collected. The tidyverse is great for this!

### 12.1.1 tidyverse

The package dplyr in the tidyverse has two wonderful functions: `group_by()` and `summarize()`. You can also use the British spelling, “`summarise()`”, but I use a z because *it's what the founders would have wanted*.

Before we use these however, we need to introduce a very useful operator, the pipe: `%>%`. This operator directs data into the first argument of a function, which allows you to chain functions together efficiently. Let's try an example with the filter and select subsetting functions (see Section 9.2.3):

```
## grab only the forest sites from the insect data
forest_sites <- fake_insects %>% ## take fake_insects and pipe it into filter...
  filter(site_type == "forest") %>% ## filter only forest rows, pipe into select
  select(site) ## select only the site column
## the whole pipe chain is assigned to "forest_sites"
```

```
forest_sites
```

```
  site
1    a
2    b
3    c
```

```
## this is the same as
forest_rows <- filter(fake_insects, site_type == "forest")
forest_sites <- select(forest_rows, site)
```

Note: base R also has a pipe, `|>`. It's newer and mostly the same as `%>%`, so I just haven't transitioned.

Now, let's try with `group_by()` and `summarize()`! Let's say you wanted the total number of insects caught at each site (be sure to have pivoted your insect data in the last chapter!):

```
## summarize total insect catch
insect_counts <- long_insects %>%
```

```

group_by(site) %>% ## group observations
summarize(total_insects = sum(count)) ## sum all insects

insect_counts

```

```

# A tibble: 6 x 2
  site total_insects
  <fct>         <int>
1 a             84
2 b             87
3 c            136
4 d             60
5 e             61
6 f             51

```

As you can see, the summarize function works a bit like the mutate function: you

Note that group\_by doesn't visibly change your data, but it changes some attributes that the computer can see when it runs the summarize function. If you forgot which sites are in which type of habitat, you could also include that variable in the group\_by arguments (since it doesn't subdivide the sites, it won't change the calculation).

```

insect_counts <- long_insects %>%
  group_by(site, site_type) %>%
  summarize(total_insects = sum(count))

```

You can also calculate means and variances! You can use the mean, var, and sd functions. Let's try for each order across all sites:

```

order_summary <- long_insects %>%
  group_by(order) %>%
  summarize(count_mean = mean(count), ## you can do multiple summaries at once
            count_var = var(count),
            count_sd = sd(count))

head(order_summary)

```

```

# A tibble: 6 x 4
  order count_mean count_var count_sd
  <chr>         <dbl>    <dbl>    <dbl>
1 coleoptera    15.8     43.4     6.59

```

2	diptera	17	163.	12.8
3	ephemeroptera	5.33	171.	13.1
4	hemiptera	6.17	42.6	6.52
5	hymenoptera	8.5	9.1	3.02
6	lepidoptera	5.67	5.87	2.42

You could also do this separately by site type:

```
orders_by_habitat <- long_insects %>%
  group_by(site_type, order) %>%
  summarize(count_mean = mean(count), ## you can do multiple summaries at once
            count_var = var(count),
            count_sd = sd(count))
```

`summarise()` has grouped output by 'site\_type'. You can override using the `groups` argument.

```
head(orders_by_habitat)
```

```
# A tibble: 6 x 5
# Groups:   site_type [1]
  site_type order      count_mean count_var count_sd
  <fct>     <chr>         <dbl>     <dbl>     <dbl>
1 forest   coleoptera         20.7        17.3        4.16
2 forest   diptera            28.3         9.33        3.06
3 forest   ephemeroptera     10.7       341.        18.5
4 forest   hemiptera          11.7        14.3        3.79
5 forest   hymenoptera        10.3         2.33        1.53
6 forest   lepidoptera         4.33         2.33        1.53
```

Finally, you can save summarized output with `write.csv()` or `write_csv()`:

```
## save summary to your working/project directory
## first argument is data, second argument is filename
write.csv(orders_by_habitat, "summary of orders")
```

## 12.1.2 Base

aggregate function (under construction...)



### 12.1.3 Making Nice Tables

the gt package is good for this (under construction....)

## 12.2 Community Ecology

Averages and variances are all well and good but what about ecological measures?

### 12.2.1 Richness

You may be interested in how many insect orders are represented in each site.

Let's do it in a pipe chain!

```
## order presence
order_richness_site <- long_insects %>%
  mutate(presence = as.numeric(count > 0)) %>% ## create binary presence column
  group_by(site, site_type) %>%
  summarize(order_richness = sum(presence))
```

`summarise()` has grouped output by 'site'. You can override using the  
`.groups` argument.

```
order_richness_site
```

```
# A tibble: 6 x 3
# Groups:   site [6]
  site site_type order_richness
<fct> <fct>         <dbl>
1 a     forest         8
2 b     forest         7
3 c     forest         9
4 d     savanna         6
5 e     savanna         7
6 f     savanna         6
```

I calculated the presence column by checking if each value is positive ( $> 0$ ), which returns a logical TRUE or FALSE, and then if you convert a logical variable to a numeric variable, TRUEs become 1s and FALSEs become 0s. Nifty!

## 12.2.2 Diversity

vegan package (under construction...)

## 12.3 Related Topic: Joining Data

Sometimes with summaries, you will want to connect them to other pieces of data. Here we have some insect counts by site, and some mammal data by site. Let's connect them! We can use the "merge" function from base R or the "join" functions from the tidyverse.

With merge:

```
## grab only the site and total columns from insect_counts
## this prevent doubling the site_type column
merged_data <- merge(insect_counts[,c("site", "total_insects")], fake_mammals, by = "site")

## look at the new column in your data
str(merged_data)
```

```
'data.frame':  48 obs. of  7 variables:
 $ site          : Factor w/ 6 levels "a","b","c","d",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ total_insects : int  84 84 84 84 84 84 84 84 84 84 ...
 $ site_type     : Factor w/ 2 levels "forest","savanna": 1 1 1 1 1 1 1 1 1 1 ...
 $ species       : Factor w/ 3 levels "Deer mouse","Meadow vole",...: 3 3 3 3 3 1 1 1 2 2 .
 $ mass_g        : int  20 24 23 19 25 22 22 21 23 20 ...
 $ tick_count    : int   0 10 2 0 12 3 2 0 NA NA ...
 $ helminth_mass_mg: int  512 365 0 608 109 456 521 432 20 129 ...
```

With join:

```
## there are different join functions for different contexts
## left_join keeps every row from the first data frame and adds any matching rows from the
## second. it works in most cases
## inner_join and full_join can also be useful
joined_data <- insect_counts %>%
  select(total_insects, site) %>%
  left_join(fake_mammals, by = "site")

## look at it
str(joined_data)
```

```

tibble [48 x 7] (S3: tbl_df/tbl/data.frame)
 $ total_insects    : int [1:48] 84 84 84 84 84 84 84 84 84 84 84 ...
 $ site             : Factor w/ 6 levels "a","b","c","d",...: 1 1 1 1 1 1 1 1 1 1 1 ...
 $ site_type        : Factor w/ 2 levels "forest","savanna": 1 1 1 1 1 1 1 1 1 1 1 ...
 $ species          : Factor w/ 3 levels "Deer mouse","Meadow vole",...: 3 3 3 3 3 1 1 1 2 2 .
 $ mass_g           : int [1:48] 20 24 23 19 25 22 22 21 23 20 ...
 $ tick_count       : int [1:48] 0 10 2 0 12 3 2 0 NA NA ...
 $ helminth_mass_mg: int [1:48] 512 365 0 608 109 456 521 432 20 129 ...

```

## 13 Analyzing Data

Now let's draw some conclusions about our data, and maybe answer questions!

Setup:

```
## load tidyverse
library(tidyverse)

## read data
fake_mammals <- read.csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vQ9mfx88nM33PC6W

fake_insects <- read.csv("https://docs.google.com/spreadsheets/d/e/2PACX-1vT0snHMdsxzzzkxt

## convert species to factor
fake_mammals$species <- as.factor(fake_mammals$species)
## convert site type to factor
fake_mammals$site_type <- as.factor(fake_mammals$site_type)
## convert site to a factor
fake_mammals$site <- as.factor(fake_mammals$site)

## also do for insects
fake_insects$site <- as.factor(fake_insects$site)
fake_insects$site_type <- as.factor(fake_insects$site_type)

## lengthen the order count data
long_insects <- pivot_longer(data = fake_insects,
                             cols = c(hymenoptera, lepidoptera, coleoptera, diptera,
                                       odonata, hemiptera, orthoptera, ephemeroptera,
                                       tricoptera, plecoptera),
                             names_to = "order",
                             values_to = "count"
                             )
```

## 13.1 Making Comparisons

First off, let's just do some simple comparisons.

### 13.1.1 t-tests

Let's say we want to compare two groups, like the number of insects caught in forests and savannas. We already created a summary of this in the last chapter:

```
## sum insects by site
insect_counts <- long_insects %>%
  group_by(site, site_type) %>%
  summarize(total_insects = sum(count))
```

``summarise()`` has grouped output by 'site'. You can override using the ``.groups`` argument.

```
insect_counts
```

```
# A tibble: 6 x 3
# Groups:   site [6]
  site site_type total_insects
<fct> <fct>         <int>
1 a     forest           84
2 b     forest           87
3 c     forest          136
4 d     savanna           60
5 e     savanna           61
6 f     savanna           51
```

It seems that there may be a difference! So let's run a t-test to test for a difference in means between two groups (see Section 4.1.2.1). Most stats functions in R can use the formula operator, `~`. This allows us to connect our dependent variable (insect count in this case) as a function of our independent variable (site habitat type): `total_insects ~ site_type`.

```
## run the t test
habitat_comparison <- t.test(formula = insect_counts$total_insects ~ insect_counts$site_ty

## check the output
```

```
habitat_comparison
```

Welch Two Sample t-test

```
data: insect_counts$total_insects by insect_counts$site_type
t = 2.6235, df = 2.1422, p-value = 0.1116
alternative hypothesis: true difference in means between group forest and group savanna is not equal
95 percent confidence interval:
 -24.30218 114.30218
sample estimates:
mean in group forest mean in group savanna
      102.33333      57.33333
```

If we look at the output, it looks like the forest mean was 102.333, and the savanna mean was 57.333, for a mean difference or effect size of 45. The p-value, or how strong the evidence for a relationship is, is 0.1116. This is higher than the traditional threshold for significance, likely because we have a very small sample size (6 total).

Note: if you check the help for the `t.test` function (run `?t.test`), you can find arguments for paired t-tests (`paired`) and unequal variances among groups (`var.equal`).

### 13.1.2 ANOVA

What if we have more than two categories, and we want to see if any two categories have different means? Let us return to the mammal data and compare the mass of helminths (parasitic worms) in different mammal species. We can run an analysis of variance (see Section 4.1.2.2).

```
## run the anova
helminth_comparison <- aov(fake_mammals$helminth_mass_mg ~ fake_mammals$species)

## check the output (now with the summary function)
summary(helminth_comparison)
```

```
              Df  Sum Sq Mean Sq F value    Pr(>F)
fake_mammals$species  2   719721   359861    10.59 0.00017 ***
Residuals            45  1529035    33979
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hey, that's a small p-value (0.00017)! That means we have strong evidence that there is at least one difference among the pairs of species, either between white-footed mice and deer mice, white-footed mice, and meadow voles, or deer mice and meadow voles. We can use a Tukey test to find out more:

```
## run tukey on the anova output
helminth_tukey <- TukeyHSD(helminth_comparison)

## check it out
helminth_tukey
```

```
Tukey multiple comparisons of means
 95% family-wise confidence level
```

```
Fit: aov(formula = fake_mammals$helminth_mass_mg ~ fake_mammals$species)
```

```
$`fake_mammals$species`
```

	diff	lwr	upr	p adj
Meadow vole-Deer mouse	-356.2222	-544.58911	-167.85534	0.0001061
White-footed mouse-Deer mouse	-111.3333	-258.37719	35.71052	0.1698916
White-footed mouse-Meadow vole	244.8889	70.26811	419.50966	0.0039996

At the bottom here we can see the pairwise comparisons. The two mouse species differ in helminth mass by ~111mg, but the difference is not significant. Meadow voles have a significantly different mean helminth mass from both mouse species. I guess the in the data I made up, voles have less helminth mass than mice.

## 13.2 Assessing Relationships

But what if you're not dealing with categorical comparisons? Then we can check for numerical associations.

### 13.2.1 Correlation

We can look for simple associations without cause and effect with correlations (see Section 4.1.3.1). Mice seem to have high helminth loads, so let's check for a correlation between their body mass and helminth mass:

```

## create a subset of only mouse data
## I use the %in% operator to specify that species should be found in a specified vector
## AKA, it could be white-footed mouse OR deer mouse
mouse_data <- filter(fake_mammals, species %in% c("White-footed mouse", "Deer mouse"))

## run correlation with two variables (no formula here)
cor.test(mouse_data$helminth_mass_mg, mouse_data$mass_g)

```

Pearson's product-moment correlation

```

data: mouse_data$helminth_mass_mg and mouse_data$mass_g
t = -10.941, df = 37, p-value = 3.741e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9323983 -0.7711347
sample estimates:
      cor
-0.8740017

```

Here we get an effect size of -0.874 (correlation coefficient), and a p-value of 3.741e-13, which means  $3.741 \times 10^{-13}$ , or  $\ll 0.001$ . This means there is a strong negative relationship observed between mouse mass and helminth mass, and we have very strong evidence for it.

### 13.2.2 Linear Regression

If we want to infer cause and effect we can use linear regression (see Section 4.1.3.2). Let's say we want to know if the number of insects at a site is predictive of mammal mass at a site. First let's join the two data frames like we did in the last chapter:

```

## join our data
mammals_insects <- insect_counts %>%
  select(total_insects, site) %>%
  left_join(fake_mammals, by = "site")

## regress mammal mass on total insects with lm function
## this time I'm specifying the data frame with the data argument
## then I don't have to write it twice
mass_model <- lm(mass_g ~ total_insects, data = mammals_insects)

```



```
## look at the output with summary again
summary(mass_model)
```

Call:

```
lm(formula = mass_g ~ total_insects, data = mammals_insects)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.6018	-1.5936	0.0221	2.0487	4.3982

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	20.15860	1.09484	18.412	<2e-16 ***
total_insects	0.02808	0.01184	2.372	0.0219 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.549 on 46 degrees of freedom

Multiple R-squared: 0.109, Adjusted R-squared: 0.08959

F-statistic: 5.625 on 1 and 46 DF, p-value: 0.02195

If we look at the coefficient table, we can see that the total insect term has an estimate of 0.028, which is our effect size. For every added insect to a plot, the expected average mass of the mammal community goes up by 0.028g. Connected to that effect size is a p-value of 0.0219, which means we have strong evidence for the relationship.

### 13.2.3 Binomial Regression

If your response variable is binary (presence absence), you can use a binomial regression with the `glm()` function. Let's test if mammal mass effects the probability of having ticks attached (`tick_count`).

```
## first make a presence absence variable for ticks
fake_mammals$tick_presence <- as.numeric(fake_mammals$tick_count > 0)

## now do the regression, with the binomial "family"
tick_pres_model <- glm(tick_presence ~ mass_g, data = fake_mammals, family = "binomial")

## check it
```

```
summary(tick_pres_model)
```

Call:

```
glm(formula = tick_presence ~ mass_g, family = "binomial", data = fake_mammals)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7550	-0.4805	0.1027	0.4404	1.4731

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-22.3584	6.7603	-3.307	0.000942 ***
mass_g	0.9857	0.2950	3.341	0.000835 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 54.040 on 38 degrees of freedom  
Residual deviance: 28.181 on 37 degrees of freedom  
(9 observations deleted due to missingness)  
AIC: 32.181

Number of Fisher Scoring iterations: 5

If we look at this like we looked at the linear regression, the `mass_g` term has a very small p-value meaning strong evidence for a relationship. It also has an effect size of 0.9857, meaning that the chance of having a tick increases with body mass. However, the units are in log odds, which are hard to interpret. The reason for this is some stats theory that is beyond the scope of this book.

### 13.2.4 Poisson / Negative Binomial Regression

under construction....

## 13.3 Multivariate Analysis

You can of course use multiple explanatory variables in your analyses. For example, when we regressed mammal mass on insect count, we ignored mammal species. We could include it like

so:

```
## multiple regression
multi_mod <- lm(mass_g ~ total_insects + species, data = mammals_insects)

summary(multi_mod)
```

Call:

```
lm(formula = mass_g ~ total_insects + species, data = mammals_insects)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.1418	-1.1856	0.1335	1.8338	5.1527

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	19.10561	1.36855	13.960	< 2e-16 ***
total_insects	0.03490	0.01272	2.744	0.00875 **
speciesMeadow vole	-0.29446	1.08017	-0.273	0.78644
speciesWhite-footed mouse	1.02943	0.86932	1.184	0.24270

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.545 on 44 degrees of freedom

Multiple R-squared: 0.1505, Adjusted R-squared: 0.09254

F-statistic: 2.598 on 3 and 44 DF, p-value: 0.06421

Now we have multiple terms, and since species is categorical, the effect sizes and p-values are based on comparisons to a reference level (deer mouse in this case because it is first alphabetically).

We can look at the overall significance of species by running an ANOVA with aov, and summarizing the output:

```
## we can use the model object in our aov function to save time, it will take the formula
summary(aov(multi_mod))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
total_insects	1	36.55	36.55	5.643	0.0219 *
species	2	13.93	6.96	1.075	0.3501

```
Residuals      44 285.00      6.48
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looks like there is not much evidence of an effect of species.

You may be confused a bit by this code, but essentially t.tests, ANOVAs, and regressions are all “linear models”, and we are specifying them and looking at them differently with different R functions. To learn more, I recommend taking stats classes!

Finally, you may be wondering which variable to include in your analyses. Model selection is another thing you would learn in stats, but the tl;dr could be: What is your question? Use those variables.

## 14 Visualizing Data

## References