CCESR Intern Hub

CCESR Fellows

Table of contents

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 form 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 exmaple 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 valuable 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 th 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

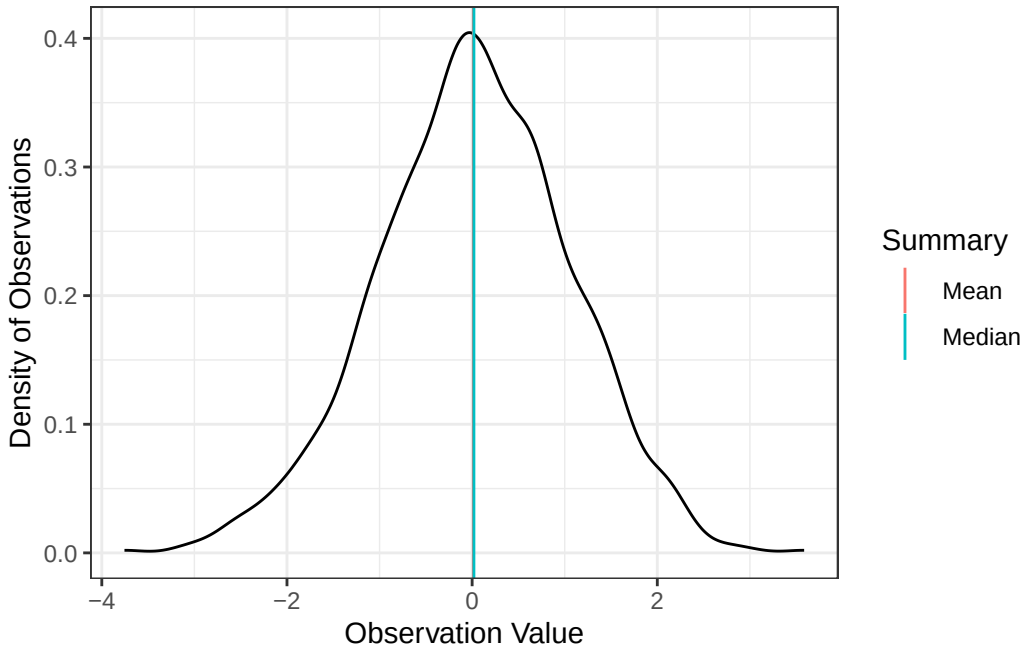
Mode - the most commonly observe value

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.

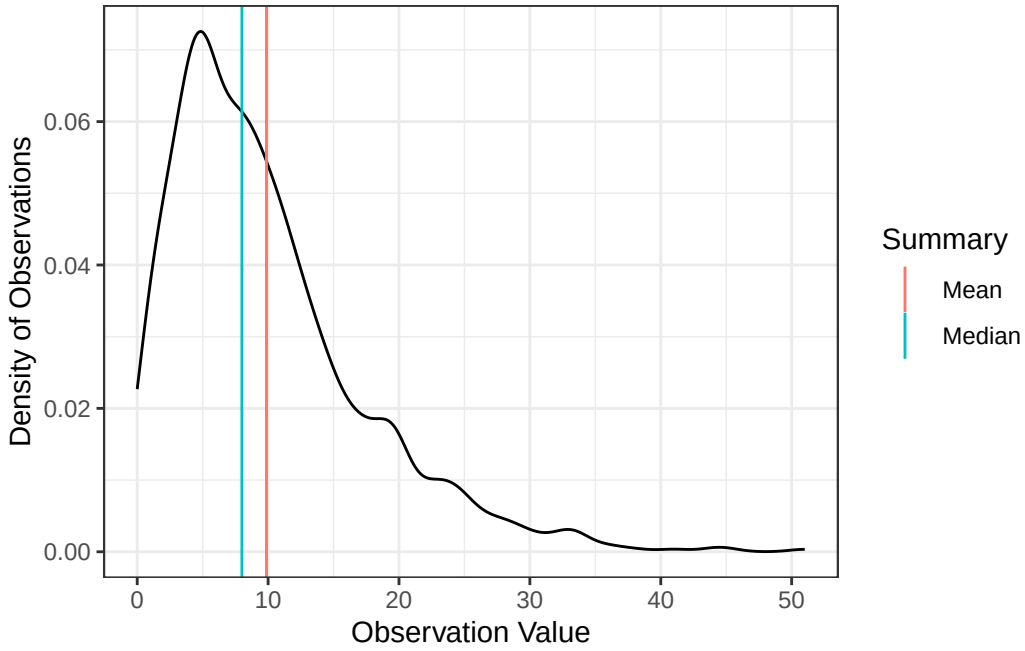
When might you prefer to use the median over the mean?

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:

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

($\sum$ means "sum of")

The differences are squared to get rid of negative differences, because other wise 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 of 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.

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 though 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., C02 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 eachother 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 divcey 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 mean 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.

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 likelihod 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 generat 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 (Image from Bea Baselga):

image

1- Code 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.

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

9 Importing Data

10 Wrangling Data

11 Summarizing Data

12 Analyzing Data

13 Visualizing Data

References