Procedures and Guidelines

Clerissa Copeland, Jason Pither, and Mathew Vis-Dunbar

2021-08-10

Contents

Welcome **7**

Copyright 7

Why Procedures and Guidelines? 8

Structure 8

File and Data Management **11**

1 File and Data Management **11**

2 File Naming **13**

2.1 Quick Reference 13

2.2 What's in a name 14

2.3 An example 16

3 Directory Structures **21**

3.1 Directory Hierarchies 21

3.2 Directory Naming 22

3.3 readme files and data dictionaries 22

3.4 Root folder readme 23

3.5 Data directory readme 24

3.6 Data dictionary 24

3.7 Example BIOL 116 25

3.8 Example BIOL 125 27

4 Tidy data **35**

4.1 Wide Data 35

4.2 Tidy Data 37

4.3 Side by Side Comparison 38

Data Presentation 41

5 Figures & Tables **41**

5.1 Tables 41

5.2 Descriptive & Summary Statistics 42

5.3 Results of Statistical Tests 43

5.4 Figures 44

Writing and Citing 51

6 Markdown **51**

6.1 How Markdown Works 51

6.2 What You Need to Get Started 52

6.3 Prose 53

6.4 Structure 53

6.5 Emphasis and Style 54

6.6 Code 55

6.7 Blockquotes 56

6.8 Lists 56

6.9 Tables 57

6.10 Links 58

6.11 Images 58

6.12 Markdown Flavours 58

7 APA Citations **61**

7.1 In-text Citations 61

7.2 Reference List 62

8 Academic Integrity **65**

9 Copyright **67**

10 Glossary **69**

Welcome

The procedures and guidelines articulated in this document represent accepted standards for the conduct and presentation of student works in the Biology undergraduate curriculum at UBC Okanagan.

These guidelines are modeled on best practices in the life sciences and where necessary, adapted specifically for the biological sciences and student engagement in learning and research.

For Students These are guidelines only. You may be asked to adhere to them directly as part of your coursework or you may be asked to work with a specific implementation of what is suggested here.

For Instructors Any concerns or omissions from these procedures and guidelines can be forwarded to Jason Pither (jason.pither@ubc.ca) or Mathew Vis-Dunbar (mathew.vis-dunbar@ubc.ca)

NOTE This is a living document. Expect that content will be added over time and adapted as needs and circumstances change.

Copyright

Please use the following for citing this document

Copeland, C., Pither, J., Vis-Dunbar, M. (2021). *Procedures and Guidelines*. https://ubco-biology.github.io/Procedures-and-Guidelines/

All source files are available at https://github.com/ubco-biology/Procedures-and-Guidelines.

Why Procedures and Guidelines?

When we talk about procedures and guidelines, we're very much talking about standards and conventions. Standards and conventions allow the products of research to be easily consumed, interpreted, adapted and re-used. For example, the metric system did wonders for standardizing how we measure distances and weight. The chaos that would ensue if every entomologist took specimen measurements with their own system - or determined wing span from different points of origin across the same species!

Standards and conventions allow us to also explore our data and outputs to greater extents than historically possible by allowing us to leverage computers; computers rely on standards to parse and merge data. Without standardization in what and how markers of climate change are recorded, we would be unable to pool the massive amounts of data collected globally to build the understanding that we are for how our climate is changing.

These qualities of standards and conventions - easy consumption, interpretation, adaptation, and re-use - are integral to robust, transparent, reproducible research; qualities that underpin the development of a strong evidence base on which to conduct further research and inform practices and policy.

Structure

This book is divided into 4 sections.

- File and Data Management
- Data Presentation
- Writing and Citing
- Glossary

File and data management covers content related to how to organize, name, and store research data. Broadly speaking, this covers all aspects of research data management.

Data presentation covers how data should be presented when summarized or analyzed.

Writing and citing touches on tools and approaches to support properly formatted and cited open publications.

Glossary is a set of standard definitions for concepts that one will encounter throughout the Biology undergraduate curriculum. When possible, course materials will link directly here.

File and Data Management

Chapter 1

File and Data Management

Well organized data is critical to transparency, reproducibility and generally maintaining one's sanity when conducting research. When we talk about file and data management we may be referring to one of many aspects of making our data understandable to others, to a computer, or to our future selves that have succumb to memory lapses. Making data comprehensible is really about well structured and communicated metadata that is, whenever possible, implemented according to conventions or standards.

So, when we talk about file and data management, broadly speaking, we're talking about

- File naming and file naming conventions
- Directory structures
- Organizing and formatting data at the variable level

Directory structures, being more complicated, bring with them the need to add additional documentation, such as a description of the directory structure and what we might expect to find where. It will also often include more detailed documentation about how to interpret what is inside specific files; one example of this is the data dictionary that describes each of the variables collected for the study.

Organizing and formatting data is a discussion about the best way to sort and parse our data into columns and rows so that we can effectively produce summaries, statistical calculations, and visualizations from these data; a core concept you'll be introduced to in this guidelines document is that of "tidy data".

Chapter 2

File Naming

File naming isn't exactly fun, but it is crucially important to being able to organize, describe, and manage any kind of work, especially research. So let's talk about file naming conventions, and those that you'll be expected to use while in Biology at UBCO.

We'll start with the rules, we'll then break these down and explain the process.

2.1 Quick Reference

- File names should only contain letters in the English alphabet, numbers from 1-9, dashes "-", and underscores "_".
- Do not use spaces or special characters, including # % & < > : " / | ? * { } $! ' @ + ` =
- File names should be broken down into components that are separated by underscores "_".
- If more than one word is needed in each component, these are separated by dashes "-".
- All file should start with your last name and all other components should be meaningful (read on for what it means for a file name to be meaningful!)

There are four variations on how these gudielines are implemented depending on what your file contatins.

2.1.1 Lab reports and manuscripts

Format LastName_Project_File-contents_Version.file-type

Example Pither_BIOL116RProject_Manuscript_V0.docx

2.1.2 Figures and plots

Format LastName_Project_Figure-title_Version.file-type

Example Pither_BIOL116RProject_Figure-freq-plot_V1.png

2.1.3 Analysis

Format LastName_Project_Analysis_Version.file-type

Example Pither_BIOL116RProject_Analysis_V0.xlsx

2.1.4 Data

Format LastName_Date_Project_Data-file-description.file-type

Example Pither_20210921_BIOL116RProject_Data.csv

2.2 What's in a name

File names need to achieve two primary goals, they need to make sense to a human reading them and they need to be constructed in a way that allows a computer to parse or process them. That is, file names should be **human interpretable** and **machine readable**. How do we achieve this?

Human interpretable

To be human interpretable, a file name needs to be meaningful. To do this, it needs to convey some basic information to a person reading it. We do this by integrating metadata into the file name. The metadata elements we include are:

- Who created the file
- The date on which it was created
- The project to which it is connected
- The nature of the contents of the file
- If it's been modified
- The type or format of the file

That is, we should be able to look at a file and tell, *who* created it, *when* it was created, *what* it is related to, *what* is inside of it, if it has been *updated*, and what *application* I should expect to be able to open it with. As we'll see shortly, we don't always include a date, and we don't always include information about modifying a file.

All said though, that's a fair bit of information to hold in a file name!

Machine readable

What does it mean for a file name to be machine readable or machine interpretable? This means building our file names in such a way that we can easily organize our files so that they can be sorted by an application and in a way that makes sense to us. It also means building our names according to set patterns, which can then be parsed along known delimiters. Lastly, it means building our names in such a way that if we move them from one computer to another, from one application to another, or from one operating system to another, the files remain interpretable in exactly the same way.

How do we this? We avoid special characters and we follow conventions.

Special characters

Special characters are any characters that are **not** part of the English alphabet, a number from 0 - 9, or one of either a dash "-" or underscore "_". This means that a space " " is a special character, which means that your file names should not have spaces.

When operating in a multi-lingual or non-English environment, this can prove problematic, but it is an unfortunate legacy of the development of computer standards that has yet to be fully resolved.

Conventions

Convention has file naming proceed in the following order, with each element separated by an underscore "_", and words within an element joined with a dash "-". The file type is generally added with a period "." and usually automatically generated when an application creates a file.

Element-1	_Element-2	_Element-3	_Element-4	_Element-5	.Element-5
Last-Name	_Date	_Project	_File-Contents	_Version	.File-type

Dates

Dates should be written in the format **yyyymmdd**. No spaces, no dashes, no words, just 8 numbers representing the year, month, and day, with months and days that are from 1 - 9 being led by a 0. So, January 23, 2020, would be written **20200123**. And September 5, 2021 would be written **20210905**. When written this way, they will always be sorted by your computer from earliest date to latest date.

Dates are very important for things like data, because the date of collection has direct relevance. Dates are less important for things like figures and manuscripts, as these are derived from the already dated data.

Versions

Version tracking is achieved in file naming by adding _Vn where n is the version number. With each major change, we increase n by 1. So version 1 would read _V1 and when updated, it would read _V2.

Versions are very important for things like manuscripts and interpretations of data, such as figures and other visualizations, where we will continue to change and modify these throughout a project. Our data, however, while it has a collection date, should not be modified, and should not then be versioned.

2.3 An example

So what does this look like?

Say you're in BIOL 116 and you're working on your research project. Your research project involves:

- Preparing the beginning of a manuscript that states your research question, hypothesis, and proposed methods.
- Conducting your experiment and recording your data. This process might span more than one day.
- Updating your manuscript, describing any changes that were made to your methods
- Organizing the results of your experiment and interpreting and visualizing your data
- Updating your manuscript to include your results and your interpretation of these results, including a visual interpretation
- Completing your manuscript by discussing the importance of and / or limitations of the experiment, and finally producing a conclusion.

In this scenario, we have **1 project**, **1 manuscript**, **1 dataset**, and at least **1 figure**. In addition, our dataset is constructed from data collected over several days, and our manuscript is revised 3 times before final submission.

So, first we will come up with a project name, and then we will **date our data** and **version our figures and manuscript**. And we'll see how this evolves over the course of several days.

Day 1

We create the following file:

Pither_BIOL116RProject_Lab-report_V0.docx

This is our manuscript, so it will get a version, but no date. Looking at it, we quickly see that this is a lab report (Lab-report), authored by someone with the last name Pither (Pither) associated with a BIOL 116 Research Project (BIOL116RProject), that it has only just been created (V0), and that I should expect it to open in Microsoft Word (docx).

Let's imagine that I will put my research question, hypothesis, and methods in this document and submit it.

Day 2

Today, I conducted the first part of our experiment and collected some data. Now we have the following files:

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Lab-report_V0.docx
```

We have not changed our manuscript, so there's no change to the name. We have collected some data though. We can easily see who collected this data (Pither), when it was collected (September 21, 2021), that it's connected to the BIOL 116 Research Project (BIOL116RProject), and that it's data related to PH exposure. Lastly, it is formatted as comma separated values (csv), which can be opened by any spreadsheet program or text editor.

Day 3-5

I continue to collect data over the next several days, and here is what my files now look like:

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Lab-report_V0.docx
```

Again, we have not changed our manuscript, so there's no change to the name. We have collected some more data though, all related to PH, and we have one file for each day, organized from earliest day of collection to most recent.

Day 6

Today, I did two things. I have no more data to collect, so I updated my manuscript to include any modifications made to my original methods section, I then submitted this. I also started to analyze my data; to do this, I merged all of the data that I have into a single file for analysis. Now my files look like this:

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Analysis_V0.xlsx
Pither_BIOL116RProject_Lab-report_V0.docx
Pither_BIOL116RProject_Lab-report_V1.docx
```

At this stage, I have my data collated into a document where I can work with it without impacting the original data. We can see that I have done this in Excel (xlsx), and that I should expect to be able to open this file in Excel. I also now have a V1 of my manuscript, as I have now added a new section to it; when submitting it, my TA knows that the file with V1 should have this updated section.

Day 7

Today, I built two visualizations using the data in my Analysis document, one linear regression and one bar plot of frequency counts; I save these as images to insert into my manuscript. I then updated my manuscript to include my results and these two figures and submitted V2 of my manuscript. Now my files look like this:

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Analysis_V0.xlsx
Pither_BIOL116RProject_Figure-freq-plot_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V0.png
Pither_BIOL116RProject_Lab-report_V0.docx
Pither_BIOL116RProject_Lab-report_V1.docx
Pither_BIOL116RProject_Lab-report_V2.docx
```

We can start to see the advantage here of naming conventions. I can easily see which files are which, what they contain, and what their timeline of development

is. Also, my computer easily sorts these into meaningful categories - my data is grouped together, sorted by date. My analyses, figures, and manuscripts are all respectively grouped together and sorted by version.

Day 8

I got feedback that my linear regression model had an error in it. So I fixed this today, added the new figure into my manuscript, and wrote the discussion and conclusion sections. I'm now ready to submit. Here is what my files look like now (I will be submitting V3 of my manuscript):

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Analysis_V0.xlsx
Pither_BIOL116RProject_Figure-freq-plot_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V1.png
Pither_BIOL116RProject_Lab-report_V0.docx
Pither_BIOL116RProject_Lab-report_V1.docx
Pither_BIOL116RProject_Lab-report_V2.docx
Pither_BIOL116RProject_Lab-report_V3.docx
```

Chapter 3

Directory Structures

Now that we've covered naming conventions, let's pretend that you store everything on your computer in a single folder. Imagine how long it would take you to find data you collected on a specific day a few years ago. Instead of keeping every document in a single place, we often organize our files using directory (aka folder) structures. This helps us save precious time and improve our productivity.

One major aspect of Open Science is ensuring transparency in the research process. This includes sharing files from all the steps of the research lifecycle (ie. a priori hypothesis, study design, data, analysis etc.) with others so that our research can be understood and replicated more easily.

The way we currently organize folders and files on our computer may make sense to us, but a problem arises when we need to share those folders and files with other people. A folder name that is meaningful for us may make no sense to another person. So let's cover some conventions to help you organize the files on your computer in a way that is meaningful to both you and others.

3.1 Directory Hierarchies

First, let's talk about how to properly structure a folder hierarchy.

The highest ranking folder is generally called the **root directory**, or sometimes the top-level folder. We'll call it the root directory here. This folder will contain all of the subfolders and files related to a particular project, including its data, analysis, lab reports etc. It will also contain what is called a readme file.

The structure should look something like the following:

```
Project-Folder/Experiment-Data/File-1
```

```
Project-Folder/Experiment-Data/File-2
Project-Folder/Experiment-Analysis/File-1
Project-Folder/Experiment-Report/File-1
Project-Folder/Experiment-Report/File-2
```

Here, our root folder is called Project-Folder and it contains three subdirectories, one for data, Experiment-Data, one for analyses, Experiment-Analysis, and one for a report Experiment-Report. Each subdirectory then contains one or many files.

3.2 Directory Naming

Key file naming conventions, such as avoiding special characters, are equally as important for directories as they are for naming files. Remember, we consider special characters to be anything other than letters in the English alphabet, numbers from 1-9, dashes "-", and underscores "_".

In case there is any doubt, here are some examples of what are considered special characters - characters you want to avoid!

% & < > : " / | ? * { } $! ' @ + ` =

Remember spaces, " ", qualify as special characters when it comes to file naming.

Remember when we name the root folder we want to clearly communicate what the project is. And similarly, within this root folder, we want clearly labeled subfolders for each relevant aspect of the project. Common discrete subdirectories include ones for figures, data, manuscript etc.

3.3 readme files and data dictionaries

When naming files we embed metadata into our file naming conventions to encode relevant information for the reader. But we can only store so much information in a file name. So we also include three additional files

- one called _README that resides in our root folder and elaborates on the contents of our folder structure;
- a second, also called _README, but that resides in our data directory and discusses some of the particulars of the how, where, and who of the actual data collection; and
- one called _DATA-DICTIONARY that also resides in our data directory and elaborates on how our data is stored and organized.

These files - containing a brief description of the major folder contents, naming conventions that were followed, and data structure - are critical for transparency and reproducibility, because they allow others to easily understand the contents of your directory and data without needing to ask you. This is especially helpful when working with a group or sharing directories with others.

General rules

A readme file and data dictionary should

- exist in at least two locations, the root directory and the data directory.
- be prepended with an underscore "_". This will push these files to the top of the directory for easy access.
- _README and _DATA-DICTIONARY files should be in all caps, so they really stand out; this should be the first thing you look at when looking at any directory or folder, as this is your guide to its contents.

File formats

readme files and data dictionaries should be written in plain text, this will ensure that the file describing your project and all of its files can be opened on any computer. You will often see readme files called _README.txt or _DATA-DICTIONARY.txt.

Our example readme and data dictionaries use a plain text format called markdown.

markdown

We recommend that you create your readme files as markdown, a way of formatting plain text files, allowing us to provide additional meaning to our content. For example, in plain text, if we want to emphasize content, we don't really have a way of doing this. In markdown, we can use italics and bolding if needed. We can also create lists and tables.

Learn the basic syntax of markdown here.

3.4 Root folder readme

To create a root folder readme file, use any markdown or text editor (ie. Typora, notepad etc), open a new file and save it to the root folder for your project ensuring the file type is .md.

Name your readme file _README.md.

Next we want to add some content to our _README.md. The purpose of this document is to describe the directory structure of our project. To adequately describe our directory structure we should include:

- A brief description of the project or purpose of the root folder
- Date when the root folder was created and who created it
- Date when the readme file was last updated and who updated it
- A brief description of the contents of each major folder within the root folder
- A brief description of file naming conventions used within the directory

To see an example root directory readme file click here.

3.5 Data directory readme

Next, we want to create another readme file but this file will be placed within the subfolder that contains our project's data. To do this, open any markdown or text editor (ie. Typora, notepad etc), open a new file and save it to the data subfolder for your project ensuring the file type is .md. Name your readme file _README.md.

The purpose of this readme file is to provide a description of data collection methods. We will include:

- Date when the data directory was created and who created it
- Date when the readme file was updated and who updated it
- A brief description of each data that was collected, the methods used for collection, and the date range for when each dataset was collected
- A description of who was involved in data collection
- A brief description of where the data was collected

To see an example data directory readme file click here.

3.6 Data dictionary

Lastly, we need to create a data dictionary which elaborates on how our data is stored and organized. To do this, open any markdown or text editor (ie. Typora, notepad etc), open a new file and save it to the data subfolder for your project. This time we will save the file as _DATA-DICTIONARY.md.

A data dictionary helps others understand the meaning of each element in your datasets within the broader context of the project. Typically you will have an individual readme file for each dataset. This file should include:

- Date when the data dictionary was created and who created it
- Date when the data dictionary file was updated and who updated it
- A description of the raw data file
- A description of each variable for all datasets including data type, units, number of levels if categorical, and a description of variable levels where relevant
- When describing variables you need to provide the full names and definitions of each variable because often variables are abbreviated in datasets

To see an example data dictionary click here.

3.7 Example BIOL 116

In BIOL 116, we used a flat folder structure to hold all of our files. In the example used, no readme files were created, as we made the assumption that the project was "simple" enough in its structure to not warrant a readme file. On reflection, this was an oversight and we probably should have created a readme file describing what was in each file. Neither did we create a data dictionary. We'll do better on future assignments now that we know about the value of both forms of documentation. Anyway, in that example, we ended up with one folder of files that looked like the following before submitting our final assignment:

```
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
Pither_BIOL116RProject_Analysis_V0.xlsx
Pither_BIOL116RProject_Figure-freq-plot_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V1.png
Pither_BIOL116RProject_Lab-report_V0.docx
Pither_BIOL116RProject_Lab-report_V1.docx
Pither_BIOL116RProject_Lab-report_V2.docx
Pither_BIOL116RProject_Lab-report_V3.docx
```

We can see that this might start to get unwieldy if we have a few more files joining the party. So let's break this apart into folders…

Top Level folder

```
BIOL116RProject/
```

Inside of BIOL116RProject we have one file and four subdirectories:

```
 _README.md
Data/
Analysis/
Figures/
Report/
```

Note that we created a _README.md file to describe our directory structure. We'll now distribute our files across these folders…

Data Folder

Creating a _README.md and a _DATA-DICTIONARY.md to describe our data files and their contents…

```
_DATA-DICTIONARY.md
_README.md
Pither_20210921_BIOL116RProject_ph-data.csv
Pither_20210922_BIOL116RProject_ph-data.csv
Pither_20210923_BIOL116RProject_ph-data.csv
Pither_20210924_BIOL116RProject_ph-data.csv
```

Analysis Folder

```
Pither_BIOL116RProject_Analysis_V0.xlsx
```

Figures Folder

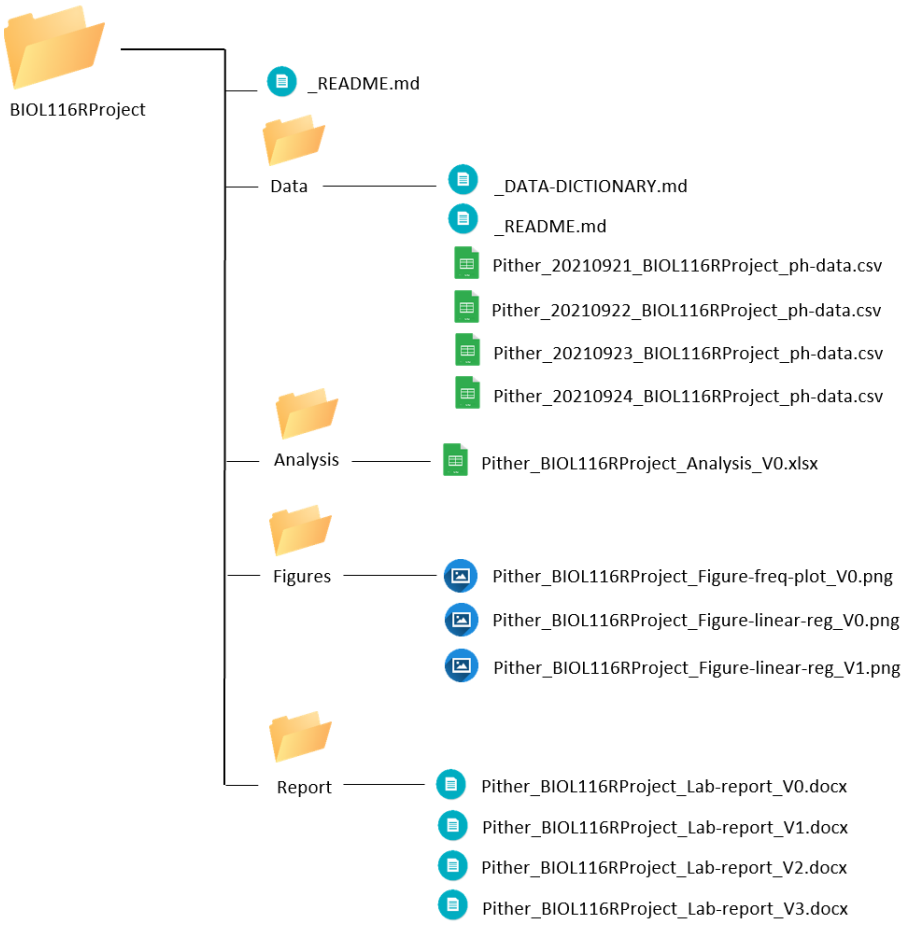
```
Pither_BIOL116RProject_Figure-freq-plot_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V0.png
Pither_BIOL116RProject_Figure-linear-reg_V1.png
```

Report Folder

```
Pither_BIOL116RProject_Lab-report_V0.docx
Pither_BIOL116RProject_Lab-report_V1.docx
Pither_BIOL116RProject_Lab-report_V2.docx
Pither_BIOL116RProject_Lab-report_V3.docx
```

Screenshot

And finally a screenshot from our desktop file manager...



3.8 Example BIOL 125

Let's work through another example where we start our project off using both appropriate directory structure and file naming conventions. Say you're a student in BIOL 125 working on a research project testing mealworm food preferences...

Day 1

On day one of our research project, we are asked to prepare the beginning of a lab report that states our research question, hypothesis, and proposed methods. First, we need to create the root folder for our project:

```
BIOL125MealwormProject/
```

Within our root folder we create a _README.md file to describe our directory structure. Let's add our project name, date the folder was created and who created it, a short description of the project, group member names, file structure (major folders and their proposed content), and naming conventions to this readme file. Your file should look something like this:

_README.md

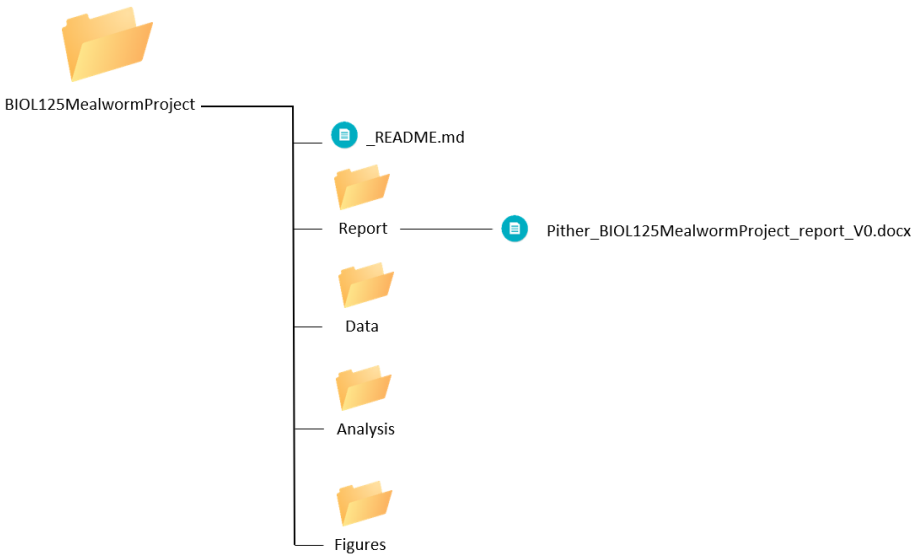
Since we have just started our project, there won't be files in most subfolders we create. However, it's good to have the skeleton of what we want our directory to look like so everyone in the group places new files in the correct location. Later, if needed, we can modify our directory structure and update our readme file to reflect those changes.

Since we will be using the naming conventions outlined in Chapter 1, we can list those naming conventions here. It may seem strange to outline the naming conventions for documents that haven't been created yet, but having a strategy for naming files from the beginning of the project is very important. It ensures everyone is following the same set of rules when they add or edit files in the project, which helps everyone stay on the same page when working in a shared directory.

Now that our root folder and root readme files are set up, we need to create the subfolders within BIOL125MealwormProject/. Since we outlined the major subfolders in our readme file as Report, Data, Analysis, and Figures, we'll use these same names when we create the subfolders. Finally, we can open up a new Word document for our lab report and save it to the Report folder using the appropriate naming conventions.

```
Pither_BIOL125MealwormProject_report_V0.docx
```

Here is what our project directory looks like so far:



Day 2

Today, we completed a pilot experiment and collected some data. We saved this data file into our project's corresponding Data folder using appropriate naming conventions.

New files:

```
Pither_20210915_BIOL125MealwormProject_food-preference-data.csv
```

Since we have added our first dataset into our project folder, we need to create a corresponding data directory _README.md and _DATA-DICTIONMARY.md.

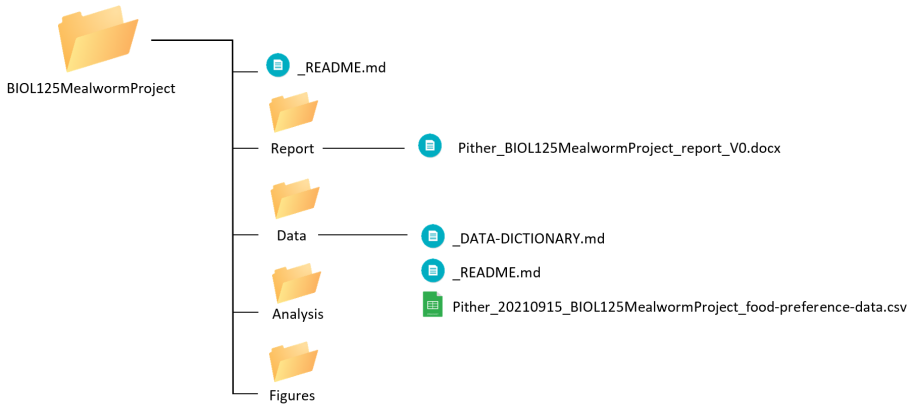
Let's start by creating the data directory readme and provide a description of our data set, collection methods, who collected the data, and where it was collected. It should look something like this:

_README.md

Next, let's create a data dictionary for our new dataset. It should look something like this:

_DATA-DICTIONARY.md

Now our project directory looks like this:



Day 3

Now that our group has completed its pilot project, we decided to expand our data collection and start recording how far mealworms are willing to travel to get each food. In addition to this new distance data we continued to collect data on food preferences.

New files:

```
Pither_20210916_BIOL125MealwormProject_food-preference-data.csv
Pither_20210916_BIOL125MealwormProject_distance-data.csv
```

Since we have a new dataset we'll have to update our data directory readme file with a description of the new dataset. Remember to note down the date it was updated and who it was updated by. Our updated data directory _README.md file should look something like this:

_README.md

Now our updated data directory readme file includes descriptions for both datasets.

In the interest of organization, let's keep our food preferences and distance data in their own subfolders. So, we'll create two new subfolders within the Data folder. We'll call one Food-preferences/ and the other Distance/. This way we can organize csv files into the folder for the corresponding dataset. After doing this, we need to update the _README.md in our root folder, since we've modified our directory structure. This readme should now look something like this:

_README.md

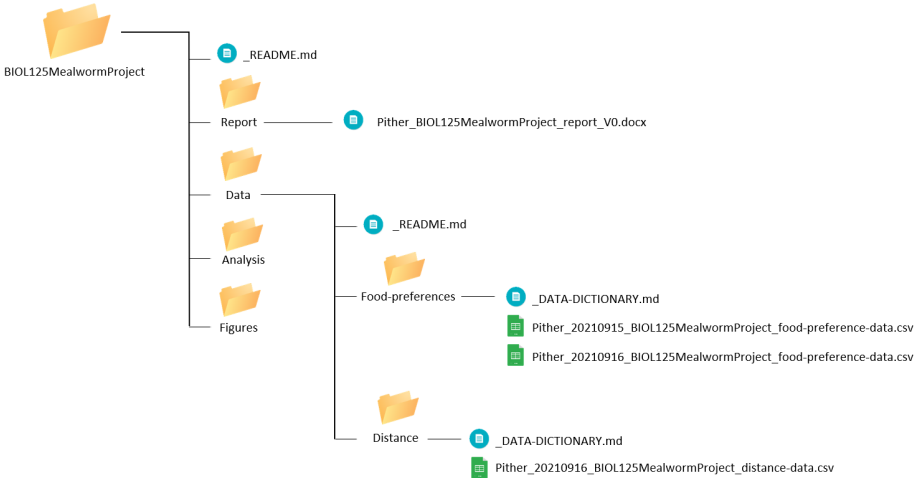
Next, let's also create a data dictionary for our new dataset within our distance subfolder. Remember to note down the date it was updated and who it was updated by. It should look something like this:

_DATA-DICTIONARY.md

Since we made some updates to our project design and methods, I'll also go ahead and updated our lab report to reflect those changes alongside justification for the changes. Then, I'll be sure to save my updated lab report using the appropriate naming conventions.

```
Pither_BIOL125MealwormProject_report_V1.docx
```

Now our project directory looks like:



Day 4-5

Over these days, we collected our last rounds of data, created some figures, and analyzed the data. So we have a bunch of new files that we need to make sure are placed correctly within our project directory.

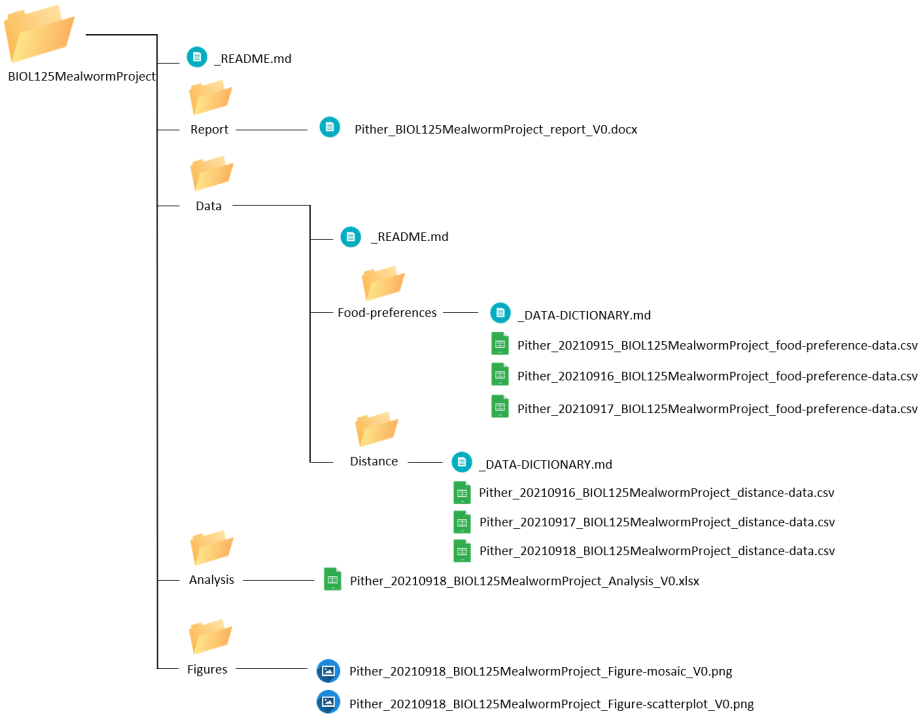
New files:

```
Pither_20210917_BIOL125MealwormProject_food-preference-data.csv
Pither_20210917_BIOL125MealwormProject_distance-data.csv
Pither_20210918_BIOL125MealwormProject_distance-data.csv
Pither_20210918_BIOL125MealwormProject_Analysis_V0.xlsx
Pither_20210918_BIOL125MealwormProject_Figure-mosaic_V0.png
Pither_20210918_BIOL125MealwormProject_Figure-scatterplot_V0.png
```

We'll save all 3 new data files into the Data/ subfolder of our directory. Since we've already described these datasets in the data directory _README.md file and have a corresponding _DATA-DICTIONARY.md for both, there are no more updates needed.

Next, we'll save our analysis into the Analysis/ subfolder and the figures into the Figure/ subfolder.

Now our project directory looks like:



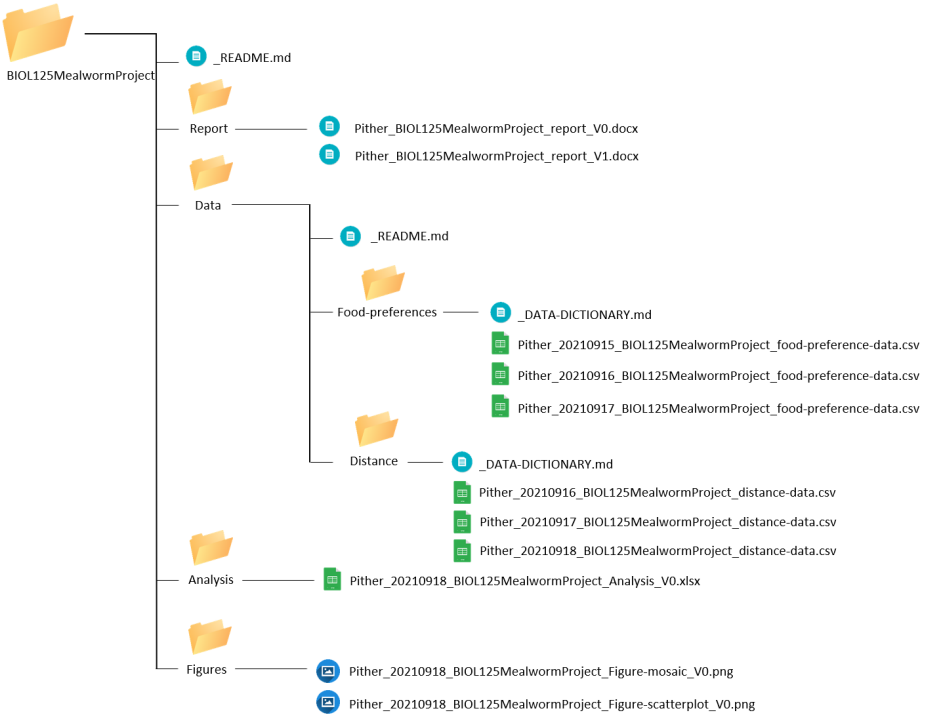
Day 6

Today is the last day of our project and we completed the final copy of our report. We will make sure this is saved into the Report folder using the appropriate naming conventions.

New files:

```
Pither_BIOL125MealwormProject_report_V2.docx
```

Our final directory looks like this:



We can start to see that if you were to share your entire project directory with another person it would be relatively easy for them to locate files and understand the meaning behind each document in our project. They would also know when changes were made and who made these changes, so if they had any questions, they'd know exactly who to ask!

Chapter 4

Tidy data

We've talked about file naming, directory structures, and documentation to ensure accessible, interpretable, and transparent data. Now it's time to talk about organizing individual variables within a given file. When well organized, data values can be effectively analyzed, summarized, and visualized. When not, they can be onerous to work with and risk misinterpretation.

In general, your data files should adhere to the principles of "tidy data". Tidy data is governed by the following 3 rules[1]:

- Each variable must have its own column.
- Each observation must have its own row.
- Each value must have its own cell.

It's easy to veer from these rules, as it's often easier to collect data using data collection tools that violate these rules. When this is the case, we need to know how to re-organize our data to make it "tidy".

4.1 Wide Data

Non-tidy data - sometimes called "wide" data as it tends to use more columns, but fewer rows - tends to lump observations together in cells. It is often easier to collect data in this way.

So, say we collected data about the number of trout caught at local lakes across several days. We might end up with the following data tables if we used a separate table, sheet of paper etc. to record our findings.

[1]See: Wickham, H. & Grolemund, G. (2017). Tidy Data. *In R for Data Science*.

Site	Trout_Caught_Day_1
Mabel-lake	1
Postill-lake	3
Ellison-lake	0

Site	Trout_Caught_Day_2
Mabel-lake	1
Postill-lake	3
Ellison-lake	0

Site	Trout_Caught_Day_3
Mabel-lake	1
Postill-lake	3
Ellison-lake	0

It is also very likely though that we set up an Excel sheet where we recorded the site as the first column and our days and fish caught combined in subsequent columns, one column for each day. Even if we hadn't collected our data this way, we might be tempted to group our above data together for analysis or assignment submission in this way.

Doing this, we'd end up with a table something like the following:

Site	Trout_Caught_Day_1	Trout_Caught_Day_2	Trout_Caught_Day_3
Mabel-lake	1	3	3
Postill-lake	3	4	5
Ellison-Lake	0	5	1

But for analysis - for "tidy" data - we want one column per variable. In this case, we have three variables:

- site
- day
- quantity caught

So let's get this cleaned up…

4.2 Tidy Data

In the previous example, our data were organized where day and quantity caught shared common columns. That is, not every variable had a dedicated column and consequently, not every variable had a value in every given cell - day did not have any cell values.

Tidy data breaks this down and reserves on column per variable and one row per observation. Remember, we have three variables: site, day, and quantity caught. So let's transform this…

First, working with a collection tool where we have one table per day:

Site	Day	Trout_Caught
Mabel-lake	1	1
Postill-lake	1	3
Ellison-lake	1	0

Site	Day	Trout_Caught
Mabel-lake	2	3
Postill-lake	2	4
Ellison-lake	2	5

Site	Day	Trout_Caught
Mabel-lake	3	3
Postill-lake	3	5
Ellison-lake	3	1

And second, gathering this data into a single dataset, sorted by site:

Site	Day	Trout_Caught
Mabel-lake	1	1
Mabel-lake	2	3
Mabel-lake	3	3
Postill-lake	1	3
Postill-lake	2	4
Postill-lake	3	5
Ellison-lake	1	0
Ellison-lake	2	5
Ellison-lake	3	1

Now that's tidy data!

4.3 Side by Side Comparison

Wide Data

Site	Trout_Caught_Day_1	Trout_Caught_Day_2	Trout_Caught_Day_3
Mabel-lake	1	3	3
Postill-lake	3	4	5
Ellison-Lake	0	5	1

Tidy Data

Site	Day	Trout_Caught
Mabel-lake	1	1
Mabel-lake	2	3
Mabel-lake	3	3
Postill-lake	1	3
Postill-lake	2	4
Postill-lake	3	5
Ellison-lake	1	0
Ellison-lake	2	5
Ellison-lake	3	1

Data Presentation

Chapter 5

Figures & Tables

These guidelines are based on current "best practices" in Biology. You may encounter small differences when working with data or reading the results of research from other disciplines. They aim to achieve consistency among faculty, instructors, and students in how data are summarized and presented within lab reports and research papers.

5.1 Tables

When presenting data in a table keep in mind the following:

- The heading is placed above the table.
- The table should be interpretable as a stand alone object using an informative heading and judicious footnotes
- Sample sizes and units are always included
- Use horizontal lines only; these are often placed above and below headings, and at bottom of table

Example

The following is an example of a properly formatted table

Table 1. Summary of trait measurements made on individuals of *Solidago* ssp. collected within shaded and open habitats in the vicinity of Portland, Oregon.

Trait

Habitat: Shaded (n = 20)

Habitat: Open (n+ = 18)

Mean (sd)

95% confidence limit

Mean (sd)

95% confidence limit

Leaf area (cm2)

4.59 (0.974)

4.14, 5.05

4.54 (0.972)

4.24, 5.15

Leaf mass (mg)

2.52 (0.765)

2.15, 2.89

w.62 (0.705)

2.25, 2.99

Root mass (mg)

9.97 (2.754)

8.67, 11.26

9.90 (2.454)

8.37, 11.16

+ data for two individuals misplaced

5.2 Descriptive & Summary Statistics

Here are some general guidelines to follow when displaying descriptive or summary statistics:

- Round numbers to one more digit for measures of centre (e.g. mean), and 2 more digits for measures of spread (e.g. sd) than was used in measuring the data
  - For detailed guidelines about significant digits, consult the following webpage: https://www.physics.uoguelph.ca/significant-digits-tutorial
- Units are preceded by a space within text passages:

  - e.g. "Average height was 34.2 cm (± 3.43 SEM)."

Describing Numerical Variables

- Report mean with standard deviation, and additionally median with interquartile range for variables that exhibit a non-normal frequency distribution (e.g. is skewed) or that includes outliers
- Parameter estimates (i.e. mean) should be accompanied by measures of uncertainty, i.e. the *standard error of the mean* (SEM) or confidence *interval* (notation: lower limit – upper limit); confidence *limits*: (lower limit, upper limit)
- Confidence intervals are strongly encouraged because they inform about *effect size*
- Measures of uncertainty for an estimate, such as SEM, can be preceded by a ± sign; do not make the common mistake of reporting a ± sign with a standard deviation, as it is not a measure of uncertainty in an estimate

Describing Categorical Variables

- Report a frequency table (raw data) or a summary table with proportions for categories (the main descriptive statistic of interest), along with the confidence interval for the proportion if appropriate

5.3 Results of Statistical Tests

Here are some guidelines for reporting the results of statistical tests:

- Your *Methods* section should clearly state the significance level ( ), and this should be decided prior to the study
- Test statistics (e.g. Student's t or an "F" from ANOVA) should be rounded to 2 decimal places, and associated P-values should report 3 decimal places, or if smaller than 0.001, then <0.001. P-values do not indicate effect size, so reporting $P = 10\hat{}$-6 is not more impressive than P <0.001
- Concluding statements should, in the absence of a table, include the test, test statistic value, degrees of freedom (*df*) or sample sizes, P-value, and confidence interval (if appropriate) in parentheses.
  - For example: "On average, hair loss was significantly greater among fathers compared to childless men (Welch's 2-sample t-test; $t = 4.23$; n F = 18, n C = 20; $P = 0.018$; 95% CI for difference: 9.34 – 18.22%)."

- Regression and ANOVA results should be shown in a standard ANOVA table format

Example of ANOVA table format

Table 3.

	SS	df	MS	F	P
Treatment	7.224	2	3.6122	7.29	0.004
Error	9.415	19	0.4955		
Total	16.640	21	4.1078		

5.4 Figures

When displaying data using a figure, follow these guidelines:

- The figure heading should be placed below the graph and should provide sufficient information so the figure can be interpreted on its own
- The heading can include statistical statements (Fig. 1) or simply describe what is being shown (Fig. 2)
- Sample size(s) must be reported
- The first time a particular type of graph is shown (e.g. boxplot), details of graph features must be provided. Subsequent figures of the same type can refer to the first for details. See Fig. 2 for an example
- Use hollow symbols so that overlapping points can be seen (Fig. 3)
- Orient all text horizontal (except y-axis label), including all tick labels
- Place axis tick marks outside of figure border to avoid overlapping with observations
- Data points should not touch the axes
- Fitted lines (e.g. least-squares regression) included in figures should be fully explained in the heading,
  - e.g. "Line represents a least-squares linear regression line, $y = 0.3 + 4.5x$ ($F = 5.65$, df = 36; P = 0.021)".
- For more complex statistics (e.g. lines associated with mixed effects models) refer the reader to the text for details
- Bar plots should **only** be used to visualize categorical data (e.g. proportion of students with brown or blue eyes) or counts (number of flies on scat)
- When comparing numerical data among categories or groups (Figs. 1,2) use stripcharts (Fig. 1) when sample sizes are small (i.e. <20) and boxplots otherwise (Fig. 2).
- Note that the stripchart has the advantage of showing all the data (i.e. each observation is represented by a point), whereas the boxplot summarizes the data visually. When sample sizes per group are very large, it would get very messy if you tried to show all the data. However, there are graphs like "violin plots" that offer a nice compromise.

Examples

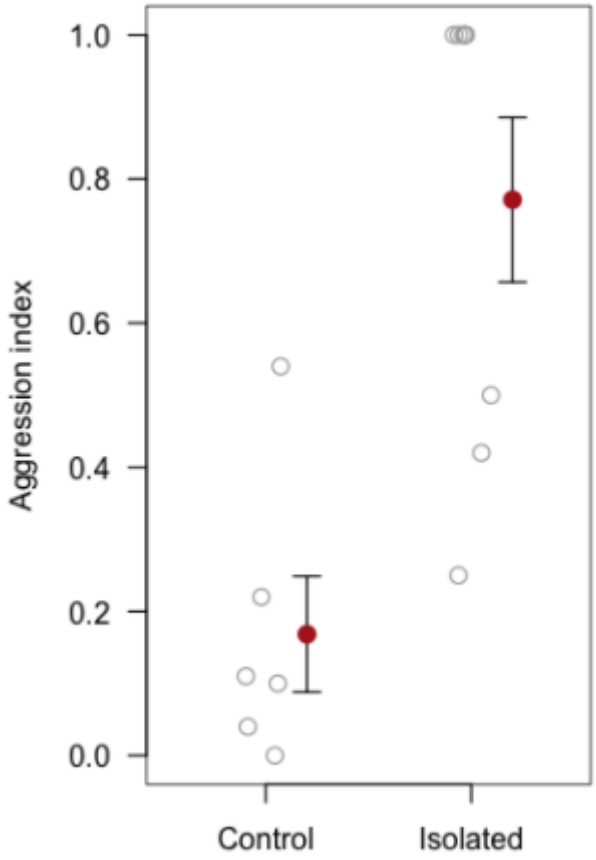


Figure 1. Aggression was significantly higher among isolated ants ($n = 8$) compared to the control group ($n = 6$) (see text for details). Shown are individual observations (grey circles), group means (solid circles) with +/- 1 SEM.

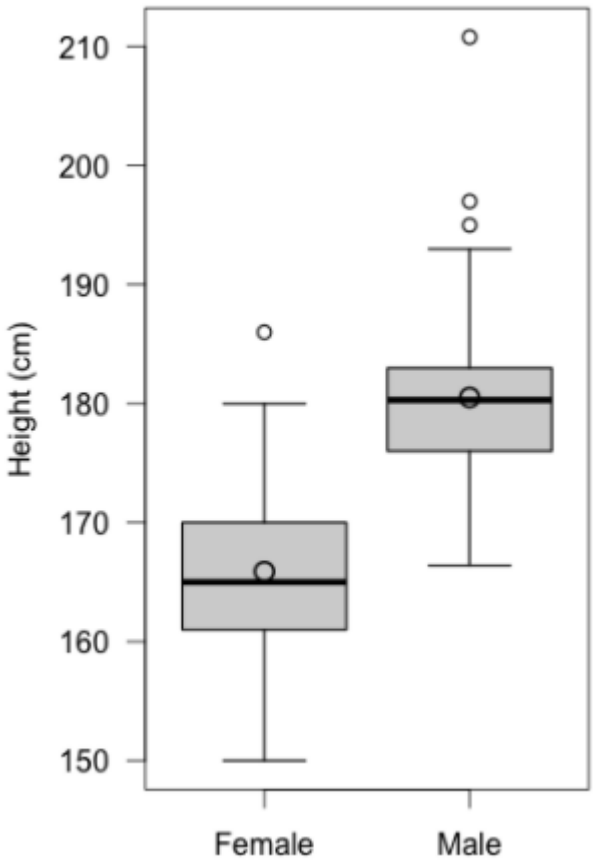


Figure 2. Height of male (n = 64) and female (n = 90) students within BIOL202. Thick horizontal lines represent group medians, large circles represent group means, boxes delimit 1st to 3rd quartiles, whiskers extend to 1.5 x IQR, and small circles represent extreme observations.

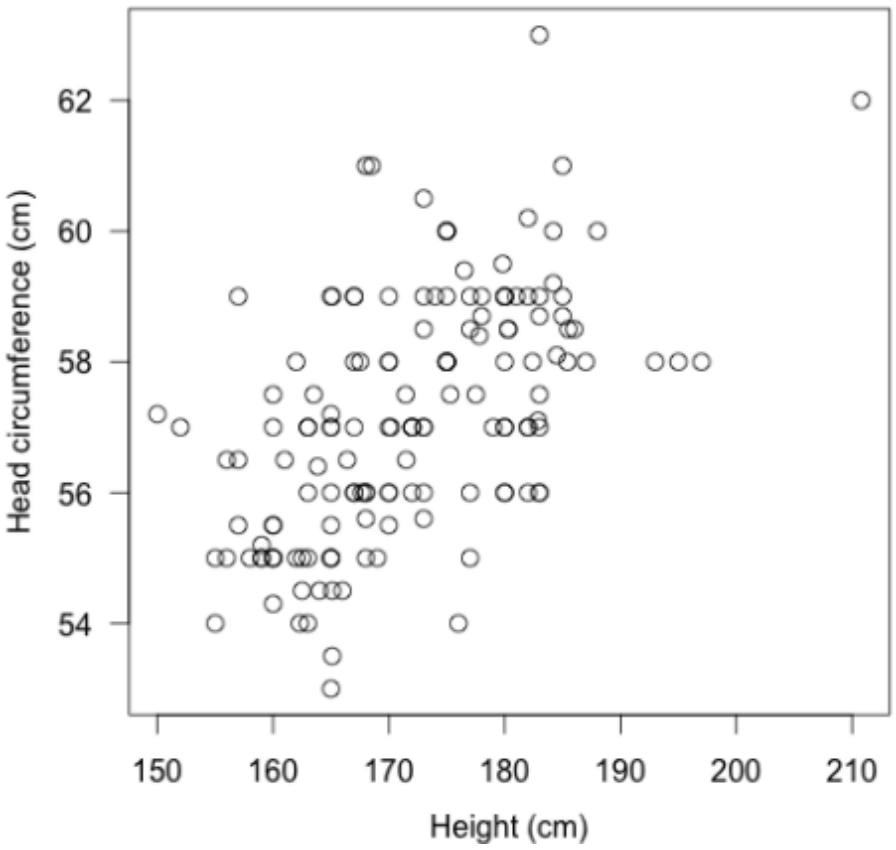


Figure 3. Head circumference versus height for $n = 150$ BIOL202 students. The positive association is highly significant (Pearson $r = 0.82$; $P < 0.001$).

Writing and Citing

Chapter 6

Markdown

Markdown is a markup language that is used to format plain text files to help us provide additional meaning to our content. Markup languages are ideal authoring tools because they work on a principle of separating out content from formatting. Markup languages are ideal authoring tools for the sciences because they rely on plain text so any computer anywhere at anytime will be able to open them and consequently, we will be able to read them.

Benefits of Markdown:

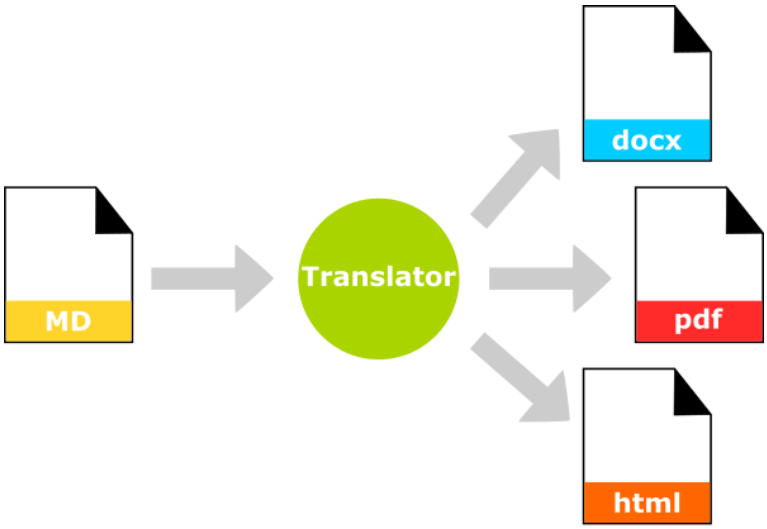
- Simple and easy to learn
- Can be used to geerate many different outputs formats (ie. pdf, html, docx etc.)
- Can be read on any device with any operating system
- Many different web-applications and websites (ie. Reddit, GitHub) use it

In fact, when we think about best practices in science leveraging tools that allow for reproducibility and interoperability (they are independent of specific operating systems or programs for example), Markdown is an essential authoring tool.

To learn more about Markdown see Matt Cone's Markdown Guide here.

6.1 How Markdown Works

Markdown is a two step process. You write, with markup, in a plain text file. Another application then formats your document based on your markup.



This makes markdown very useful for producing many different types of documents from the same piece of prose, whether that be pdf, html, docx you name it. It also means that you have one working, editable document, which is plain text, and a series of distributable copies in other formats. More importantly, formatted in a way to address the needs of that particular audience.

All of the content that you're reading right now was authored in Markdown. Note at the top of the page, you can download this content as a pdf or as an epub. But we only had to write it once.

6.2 What You Need to Get Started

You'll need a text editor or a dedicated Markdown editor! What's the difference?

A text editor will not format your markdown, it'll just display the plain text. When you open the _README.md files that are dotted througout this book in **textedit** on a Mac or **notepad** on a Windows machine, you're seeing this raw, plain text output.

A dedicated Markdown editor will format your Markdown as you type, so you get an idea as to how it will render if you were to save it as a pdf, html etc. There are a lot of Markdown editors available, many of them free and open source.

Already have a favourite text editor or Markdown editor? Great. If not, we'd like to suggest the following, which provide an easy to use interface and work across two panes, one where you see your plain text output and one where you see how it will be rendered when formatted. Both are free and open source.

Mac OS

MacDown available at https://macdown.uranusjr.com/

Windows

Markdown Edit available at https://mike-ward.net/Markdown-Edit/

6.3 Prose

You can just start writing if you've opened up a new document in MacDown, Markdown Edit, or whatever text or Markdown editor you happen to be using.

There's a couple of things you'll note right away.

- Whether you use one space, " ", or many spaces in between your words, it will only render (display with formatting) as if there was one space.
- Just hitting 'Enter' once doesn't put you on a new line. You need to hit 'Enter' twice.

6.4 Structure

Headings

You can add structure to your text document by adding headers with different hierarchies. To do this you add a number sign "#" before the text. The number of "#" symbols indicates the hierarchy of the header.

Make sure you include a space " " between the "#" and your header

For example, the following…

```
# This is a first-tier header
## This is a second-tier header
### This is a third-tier header
#### This is a fourth-tier header
```

…would render as

This is a first-tier header

This is a second-tier header

This is a third-tier header

This is a fourth-tier header

6.5 Emphasis and Style

There are a number of different Markdown syntaxes that we can use to add style and emphasis to specific parts of our document. Below are a few examples.

Italics

You can make text italicized by encasing the text in a single asterisk "*" or underscore "_".

Input

```
*This is italicized*

_This is also italicized_
```

Output

This is italicized

This is also italicized

Bold

You can make text bold by encasing the text in two asterisks "**" or underscores "__".

Input

```
**This is bold**

__This is also bold__
```

Output

This is bold

This is also bold

Bold and Italics

If you want to make text both bold and italicized you can encase the text with three asterisks "***" or underscores "___".

Input

```
***This is bold and italicized***

___This is also bold and italicized___
```

Output

This is bold and italicized

This is also bold and italicized

NOTE When italicizing or bolding characters within a string of text it's better to use an asterisk "*" rather than an underscore "_". For example: Biology *is* awesome NOT Biology _is_ awesome

Strikethrough

You can strike through text by encasing it with two tildes, "~~".

Input

```
~~This text has a strike through it~~
```

Output

~~This text has a strike through it~~

6.6 Code

You can show text as code by encasing it with three back ticks, "`.

Input

"`

This text looks like code

"`

Output

```
This text looks like code
```

6.7 Blockquotes

To create a blockquote using markdown, you need to place a greater-than, ">" sign in front of the text.

Input

```
> This text is placed within a block quote
```

Output

> This text is placed within a block quote

NOTE Leave a blank line both before the blockquote and after it.

6.8 Lists

Ordered lists

To create an ordered list in Markdown, you'll need to place the item number and a period in front of the text.

Input

```
1. First item
2. Second item
3. Third item
```

Output:

1. First item
2. Second item
3. Third item

Unordered lists

You can create an unordered list by placing an asterisk "*", dash "-", or plus sign "+" infront of the text.

Input

```
* first item
* second item
* third item

- first item
- second item
- third item

+ first item
+ second item
+ third item
```

Output

- first item
- second item
- third item

6.9 Tables

When creating tables using Markdown, pipes "|" separate columns while line breaks are used to separate rows. The column header is seperated by three or more hyphens "---" between each column's pipe "|".

Input

```
| Column 1 Header | Column 2 Header | Column 3 Header |
| --- | --- | --- |
| Column 1 item | Column 2 item | Column 3 item |
| Column 1 item | Column 2 item | Column 3 item |
```

Output

Column 1 Header	Column 2 Header	Column 3 Header
Column 1 item	Column 2 item	Column 3 item
Column 1 item	Column 2 item	Column 3 item

NOTE

- The number of hyphens, "-", used can make the cell width look incorrect. However, as long as there are three or more hyphens the rendered output will be the same.
- Put a space, " ", between each pipe "|" and the following word or dash "-"

6.10 Links

To create a link to a url or another document, encase the text with square brackets, "[]", and follow the text immediately with the link encased in parentheses, "()".

Input

```
To visit the UBC Okanagan Faculty of Biology website click [here](https://biology.ok.ul
```

Output

To visit the UBC Okanagan Faculty of Biology website click here.

6.11 Images

Creating a link to an image follows a similar format to that of links, but the square brackets encasing the text are preceded with an exclamation mark "!". You can place either the url link or path to the image on your computer in the parentheses. If using a path on your computer, use a relative path.

Input

```
![A magnificent caterpillar. Photo by Erik Karits on Unsplash](images/caterpillar)
![A magnificent caterpillar. Photo by Erik Karits on Unsplash](https://unsplash.com/pho
```

Output

6.12 Markdown Flavours

Everything here is basic, or core, Markdown and will be supported by any markdown editor. Since Markdown is simply encoding document structure through markup, several different implementations have expanded on this core set and allow you to do other things, including footnotes, references etc.

We'll just keep it simple for the moment; this is all you really need for your _README.md files! Later, you'll be introduced to RMarkdown which, when used in conjunction with R, will allow you to render statistical analyses within your Markdown document and build reference lists among other things.

Figure 6.1: A magnificent caterpillar. Photo by Erik Karits on Unsplash

Chapter 7

APA Citations

This is a brief overview of the 7th edition of APA for quick reference. For a more in depth review, please consult the library's APA Citation Guide. For the full APA manual, please consul the Publication Manual of the American Psychological Association, available through the library.

Additional resources you may wish to consult include:

- The APA Style Blog - great for searching for examples not listed in the 7th edition
- Purdue OWL's website for still more examples

7.1 In-text Citations

Always appear right after the content you are summarizing, paraphrasing, or quoting. The format is as follows:

Summarizing or Paraphrasing

- (Author, YYYY)

Quoting

- (Author, YYYY, p. #) 1 page
- (Author, YYYY, pp. ##-##) Multiple pages

Narrative vs Parenthetical

If you mention the author or authors in text, you do not need to include this information in the brackets, "()". This is called a narrative citation.

Narrative in-text citation: Raimi (2018) outlines the risks and benefits of fracking through an economic analysis and energy security benefits.

The alternative is called a parenthetical citation.

Parenthetical in-text citation: Several benefits and risks can be identified in the implementation of fracking for oil extraction. Considerations include regulation, water pollution, tremors etc. (Raimi, 2018).

Quick Format Guide

# of Authors	Narrative Example	Parenthetical Example
1	Bradley (2017)	(Bradley, 2017)
2	Janmaat and Rahimova (2018)	(Janmaat & Rahimova, 2018)
3 or more	Mei et al. (2018)	(Mei et al., 2018)

7.2 Reference List

NOTE

- Every source used in your in-text citations needs to be listed as part of your reference list, in alphabetical order by author(s)' last names.
- The word **References** should appear at the top of your reference list, and it should be centred and bolded on the page
- Titles should be written in sentence case, that is, capitalize the first word and only subsequent proper nouns. If the title is broken up by a colon (:), capitalize the first word after the colon.
- List all authors in the order that they appear in the source.

Journal article with a DOI (1-2 authors)

List all authors in reference list and in-text citations.

Janmaat, J., & Rahimova, N. (2018). Managing drought risk in the Okanagan: A roll for dry-year option contracts? *Canadian Public Policy*, *44*(2), 112-125. https://doi.org/10.3138/cpp.2017-003

Journal article with a DOI (3-20 authors)

List all authors in the reference list and only first author in the in-text citations.

Mei, Y., Yu, K., Lo, J. C. Y., Takeuchi, L. E., Hadjesfandiari, N., Yazdani-Ahmadabadi, H., Brooks, D. E., Lange, D., & Kizhakkedathu, J. N. (2018). Polymer-nanoparticle interaction as a design principle in the development of a durable ultrathin universal binary antibiofilm coating with long-term activity. *ACS Nano*, *12*(12), 11881-11891. https://doi.org/10.1021/acsnano.8b05512

Chapter 8

Academic Integrity

Academic integrity is the act of performing honest, responsible scholarship, much like scientific integrity is honest, responsible science. Academic integrity is very much about how we conduct ourselves in the pursuit of our studies, respecting those we learn from and work with and the contributions they make to our scholarly and scientific endeavours.

Learn more about academic integrity from the UBC Learning Commons.

Chapter 9

Copyright

Copyright is the legal ownership of a work, like a book, image, graph, journal article. Copyright law governs how and when we are allowed to redistribute what someone else has produced and how others can redistribute what we've produced.

Learn more about copyright at UBC from the UBC Learning Commons.

Chapter 10

Glossary

A priori hypothesis: Hypothesis that is generated before the research study takes place. Presenting the hypothesis before the study takes place helps in avoiding replacing the hypothesis later with one that fits the data better aka hypothesizing after the fact (HARKing).

Alternative hypothesis: In contrast to the null hypothesis, the alternative hypothesis suggests there is a relationship between phenomena, variables, or populations. In other words, any differences are not the result of random chance.

Analysis of variance (ANOVA): A statistical test used to compare the mean of a numeric variable in relation to a single categorical variable that has more than two groups.

Assumptions: There are often assumptions associated with statistical tests. This means that for the test to provide reliable results the data must meet specific criteria or conditions. These assumptions need to be checked prior to conducting any analyses.

Bias: Error is introduced and false conclusions might be drawn because our sample doesn't meet established standards for faithful representation of our population of interest.

Binomial distribution: A discrete probability distribution of the number of successes where there are exactly two possible outcomes (success and failure).

Binomial test: A statistical test that determines the probability of getting a particular proportion when there are exactly two possible outcomes (success and failure).

Burden of proof: The obligation that when a causal link is suggested that evidence to support this link must be presented. This can be accomplished through independent replication of studies where if they demonstrate the same conclusions it reinforces the validity of the causal link between those variables.

Chi-square (χ^2) contingency test: A statistical test used to assess whether there is an association between categorical variables. This test is used on contingency tables that are larger than 2 x 2.

Chi-square (χ^2) goodness of fit test: A statistical test used to test how well an observed discrete frequency (or probability) distribution fits some specified expectation.

Clinical trials: Experiments that involve i) human participants who are assigned in advance to a group that receives a particular treatment designed to produce a biomedical or behavioural result and ii) evaluation of the effect of the treatments (NIH, 2017)

Citizen science: When members of the public engage in the research process. Often involving collaboration with researchers.

Coefficient of determination (R2): The proportion of the variance in the response variable that can be explained or predicted by the independent variable. It describes the strength of the relationship between two variables.

Coefficient of variation (CV): A relative measure of variability that indicates the size of a standard deviation in relation to its mean." [Frost, Jim n.d.](Coefficient of Variation in Statistics - Statistics By Jim

Comma-separated values (CSV) file: A plain text file where each line of the file represents a record and each field (column) entry for that record is separated by a comma. This file format is frequently used by researchers to store data.

Confidence interval: An estimated range of values that has an associated probability. The probability describes the likelihood that this range of values will contain the true value of a parameter (ie. mean). For example, a 95% confidence interval suggests we can be 95% confident that the true parameter lies within that range of values. Or in other words, on average we can expect the true parameter to lie in this range, 95% of the time.

Confirmatory research: Researchers used a well designed experiment to test the validity of predetermined hypotheses that can be disproved.

Critical analysis: Careful examination and evaluation of all parts of a research article including consideration of the study's strengths and weaknesses as they relate to study design, implementation, data collection, data analysis, and interpretation.

Data transformation: A process where the format of the values within a dataset are changed. For example, all of the values within a dataset might be log transformed. This is often done when the original data does not meet the assumptions of a particular statistical test. After the data transformation researchers will re-assess those assumptions to see if they can perform the test on the newly transformed data.

Descriptive statistics: A number used to summarize or describe a given data

set or sample. Examples include mean, median, mode, standard deviation, and interquartile range.

Diversity: The practice or quality of having individuals who vary in terms of social class, ethnic background, sexual orientation, gender, religion, ability, etc.

Effect size: A measure of the degree of association between one variable and another, or in experimental contexts, of the impact of one variable on another.

Equity: The practice of treating all segments of society in such a manner that everyone has a similar chance of achieving a given outcome. Some individuals and groups may need more or different support than others to achieve that outcome. "Equality" , in contrast, refers to the practice of providing identical support and opportunities to all.

Exploratory research: Research that is performed to gain a better understanding of an existing problem. For example, it might give rise to hypotheses that can then be tested through confirmatory research.

File and data management: Refers to practices used to collect, generate, and store data and files throughout the research process. Researchers should document what type of data they have collected, the methods used, and any relevant context. Files and data should be stored such that they are organized, accessible, and interpretable by both the researcher and others.

Fisher's exact test: A statistical test used to assess whether there is an association between categorical variables. This test is used on contingency tables that have exactly 2 x 2 dimensions.

HARKing: A form of questionable research practices where the researcher changes their hypothesis after the study is conducted so that the hypothesis better fits the data. In other words, the researcher suggests that this post hoc (after the fact) hypothesis was formed a priori. This has a number of implications including: "harming the progress of science by preventing the research community from identifying already falsified hypotheses, contributes to the replication crisis, and it increases the probability that the findings are not reproducible or generalizable in the population of interest". (UCDavis Health, n.d.)

Hypothesis: A proposed explanation for an observed phenomenon. Often structured in an "If… then… because…" format. Hypotheses must be present a priori, be falsifiable, and measurable.

Hypothesis testing: Typically involves setting a null and alternative hypothesis and performing an appropriate statistical analysis to test those hypotheses. Often used in confirmatory research.

Inclusion: The philosophy or practice of considering individuals from diverse backgrounds in relation to the community, organization, or society, and ensuring that they feel that they belong, supporting them in giving their best efforts, and giving them equal opportunities to advance and participate in decision-making.

Interquartile range: A descriptive statistic that measures the variation within the middle section of a set of values. Specifically, it describes the range between the first and third quartiles of a set of values.

Linear regression: A statistical method used to model the linear relationship between independent and dependent variables.

Literate programming: A coding paradigm where natural language is written alongside or between lines of code to provide an explanation for the code's logic. This practice helps enhance reproducibility and understanding by guiding readers through the programmers thought process.

Literature review: A review of scholarly sources related to a specific research question or topic. Involves recording a list of research studies consulted, how they were found, and the strengths, limitations, and weaknesses of each.

Long format data: A method for organizing data where all of one subject's observations are represented by distinct rows.

Markdown: Markdown is a markup language that is used to format plain text files to help us provide additional meaning to our content. For example, using Markdown you can use bold, italics, and create tables. Markup languages are ideal authoring tools because they work on a principle of separating out content from formatting.

Mean: A commonly used descriptive statistic that measures the central tendency of a numeric variable. Specifically, the mean is the arithmetic average of a group of values.

Median: A descriptive statistic measuring the central tendency of a numeric variable. The median is the value separating the upper and lower halves of the variable. In other words the middle value of a group of numbers.

Metadata: Data that provides information about other data.

Mode: A descriptive statistic for a either a numeric or categorical variable. The mode is the value that appears most frequently.

Null hypothesis (Ho): Used alongside the alternative hypothesis in hypothesis testing. The null hypothesis states that there is no significant effect or relationship between phenomena, variables, or populations. Rather any differences observed are the result of random chance.

Odds ratio: Measure of the relative odds of the occurrence of a specific event (ie. cancer) given the exposure to a variable of interest (ie. smoking). This ratio is often used to determine the odds of health related outcomes.

One-sample t-test: A statistical test used to compare a numeric response variable (ie. mean) to an expected value.

Open notebooks: Involves i) publishing or linking to data on an online platform before results are published in a peer-reviewed journal; ii) making information about the methodology and equipment used in a study publicly available;

iii) openly discussing both positive and negative results in real time, as they are obtained.

Open science: A movement and set of practices intended to combat the replication crisis, QRPs, and style trumping substance by making all parts of the scientific research process transparent and accessible, allowing for a critical review of how a study was conducted, ultimately enabling that study to be independently replicated. It also involves changing scientific culture to reward not just novel findings, but also the many other aspects of conducting good scientific research.

P value (p): The probability of getting a result that is the same or more extreme than what was observed. If the probability of getting that result due to random chance is sufficiently low, then it could be interpreted that there is a significant relationship. In contrast, a high p value indicates a larger likelihood that the result was due to random chance and therefore there may be no significant relationship. The p value required to establish significance is set by the researchers in advance of the study and is known as the significance level (α).

Paired t-test: A statistical test used to compare the means of two samples where an observation in one sample can be paired with an observation in the other sample. For example, observations might be linked because they were before and after observations on the same subject or in the same place,.

Participatory research: Turns the relationship between researcher and subject into a partnership, where both contribute to the research question, methods, and outcomes.

Pearson correlation: A statistical test that measures the linear correlation between two numeric variables.

Peer review: Peers of the author critically review the author's study. Traditionally peer review focused on the evaluation of studies prior to publication, however open science practices suggest additional peer review at the study design stage prior to implementation. This ensures the study design meets accepted quality standards before it is conducted.

Plain text: Simple text that is human readable. It can includes letters, numbers, symbols, and spaces but does not have any special formatting and is not computationally tagged.

Post-hoc test: If a significant result is found when performing a statistical test, post hoc tests can be done to provide more details about where those significant differences are arising from. They are another form of statistical tests.

Probability: Describes the likelihood of an event occurring. For example, when a fair coin is flipped the probability of getting tails is 0.5.

Proportion: A number between 0 and 1 that represents the fraction of the total population with a certain attribute. For example, if 10 students have red

hair in a school with 100 students, then the proportion of students with red hair is 10/100 or 0.1.

Questionable research practices: A grey area of scientific practice in which researchers do not engage in outright misconduct such as fraud or plagiarism, but may unwittingly break rules of acceptable scientific practice in the pursuit of novel and promising results.

R: a programming language and free software for statistical computing. When used throughout the research process, it allows for openness in the research workflow and computational reproducibility.

Relative path: In contrast to an absolute path, a relative path is a URL that only contains a portion of the path. It is relative to the root of the document and thus should start the path with the directory name that contains the document. For example, if you are writing a Markdown document and would like to include an image of a mealworm, place the mealworm image into the same directory (folder) as the Markdown document and the relative path may look like /BIOL116/project/mealworm.png. Whereas the absolute path might look like C::/Documents/BIOL116/project/mealworm.png.

Random assignment: Assigning participants of a research study to each condition using a method of randomization. This ensures that each participant has an equal chance of being placed in each condition and helps to minimize bias.

Range: A descriptive statistic or measure of variation for a set of values. Specifically, it measures the difference between the highest and lowest values.

Replication: Thorough repetition of a study, using the same methods but different data.

Replication crisis: Many studies cannot be competently analyzed or replicated. This is because critical information about them—design, data, methods, lab notes, analyses and code—may not be made available, or may be poorly communicated. This problem is escalated further because new and original findings are considered more exciting than re-testing or replicating previously conducted studies.

Reproducibility: "Obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis".

Research transparency: The quality or practice of revealing all inputs and outputs of the research process clearly, as well as making evident the exact reasoning and process used in coming to a decision or taking actions in research, in such a way that the study can be replicated. As well, transparency means taking care to disclose important information in a respectful and responsible fashion.

Research lifecycle: The traditional research cycle involves five stages, 1) develop idea, 2) design study, 3) collect and analyze data, 4) write report, and 5) publish report. Traditionally, peer review has been conducted after writing

the report and prior to publication. However, open science proposes revising the research life cycle by introducing an additional peer review after the study design stage.

Scientific method: An empirical method for acquiring knowledge which includes making an observation, asking a question, forming a hypothesis, making a prediction based on the hypothesis, and testing the prediction.

Scientific integrity: "The condition resulting from adherence to professional values and practices when conducting, reporting, and applying the results of scientific activities that ensures objectivity, clarity, and reproducibility, and that provides insulation from bias, fabrication, falsification, plagiarism, inappropriate influence, political interference, censorship, and inadequate procedural and information security." (USDA, n.d.)

Significance level (α): The probability of rejecting the null hypothesis when it is true. For example, a significance level of 0.05 means there is a 5% chance that researchers might conclude a significant relationship or difference exists when there is no true relationship or difference. This is also known as the Type I error rate. The significance level is set by researchers before conducting a study and the p value result is compared to the α to determine if there is a significant relationship or difference.

Spearman rank correlation: A nonparametric statistical test that measures the statistical dependence of ranking between two numeric variables.

Standard deviation: A type of descriptive statistic that is used to quantify the amount of variation within a set of values. A set of values with a large standard deviation exhibits high variability whereas a low standard deviation indicates the values are close together.

Standard error: The standard deviation of the sampling distribution for a specific parameter. For example, if the parameter of interest is the mean, the standard error of the mean would be the standard deviation of the sampling distribution of the mean.

Statistical analysis: Involves collecting, organizing, exploring, interpreting, and presenting data to uncover patterns or trends in the data. Involves using statistics to describe the study sample and use that sample to make inferences about the population of interest.

Statistical significance: If a result is determined to have statistical significance it means that the result from the study is not likely to have occurred randomly or by chance. In other words, the result is likely to be caused by something other than chance. The significance level ( ) is set by the researcher in advance of the study being performed. Often is set to 0.05, which indicates a 5% chance of making the wrong decision and determining that the null hypothesis is false when it is in fact true.

Study power (aka statistical power): The probability that a random sample taken from a population will lead to rejection of the study's null hypothesis if

that null hypothesis is in fact false. That is, power is a measure of how reliable a study is as a test for its hypothesis; power is positively influenced by things like large sample sizes and relationships characterized by large effect sizes.

Two-sample t-test: A statistical test used to compare the means of two independent samples.

Variance: A descriptive statistic that "measures how far a set of numbers is spread out from their average value." Wikipedia, n.d. In other words, a high variance indicates the values are spread further from the mean whereas a low variance indicates they are close to the mean.

Version control: Saving changes to files while retaining the changes on all previous versions of the file. This practice contributes to transparency and openness in science.

Wide format data: A method for organizing data where each row represents an individual subject and each column represents an observation for that subject.