Data Set Basics

Kelsey Shoub
July 2021

Overview

Maybe you're tackling building your first data set and are looking for a quick refresher in construction and considerations, or maybe you're looking for a quick refresher on data sets and data documentation as you dive into your first analysis. Either way, this is for you.

This is not intended to be comprehensive how to document, but rather it should be used as a quick resource to either remind you of the basics or as a launch point for investigating topics related to working with and documenting data. This handout briefly touches on the following topics:

1. So, what is this data thing anyway?
   This is a brief (re)introduction to what a data set is and their typical construction.

2. How is data stored and shared?
   This is a brief (re)introduction to where you might find data sets, what formats you may need to work with, and what you might consider when choosing the file format for your eventual data set.

3. What documentation should be provided or created?
   Ideally, when a data set is created, an accompanying code book is also create. This is a brief overview of what must be in a codebook and what ideally is contained in a codebook.

A caveat on this document: this is meant as a quick, birds eye view – not as an all encompassing document covering each topic in detail. As such, some knowledge is assumed, and you'll likely need to access supplemental resources..

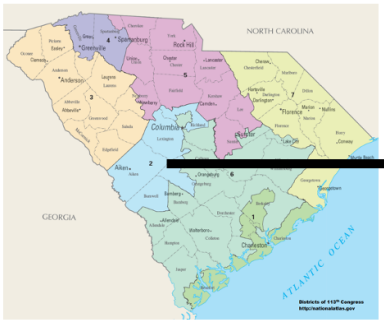
So, what is this data thing anyway?

To conduct quantitative analysis, you need a data set. In its most basic form, a data set is a table, where each row of the table provides information on the specified unit of analysis and where each column indicates a different piece of information. You, the researcher, then use this data set to tell us something interesting about the topic you're studying. Note that the ideal data frame, table, matrix, or spreadsheet – which is what this represents – only includes column names or titles but not additional information on what is contained within each cell, where the

data comes from, coding schemes, etc. All of that information should exist, but it should be housed elsewhere to facilitate cleanly working with the data set.

Let's walk through an example. Pretend that you want to understand whether Congressional candidates who raise more money are more likely to win their general election races for the US House of Representatives. To answer this question, or at least to begin to answer it, you would turn to a data set that records information by candidate on: the candidate's name, what race they are running in, the specific election (i.e., type and year), how much they raised, how much their opponent raised, the difference, whether or not they raised more than their opponent, and whether that candidate on.

Let's look at a snapshot of what this data set might look like for South Carolina's 2nd district.



Candidate Name	House District	Amount Raised	Amount Opponent Raised	Difference in Amounts Raised
Joe Wilson	2	$1.2	$0.1	$1.1
Sean Carrigan	2	$0.1	$1.2	-$1.1

Let's break this down a little more.

Variables in the columns. This may be represented by the letter *k*.

Observations in the columns.

This may be represented by the letter *n*, such that *n* indicates a specific row while *N* indicates the total number of observations.

Candidate Name	House District	Amount Raised	Amount Opponent Raised	Difference in Amounts Raised
Joe Wilson	2	$1.2	$0.1	$1.1
Sean Carrigan	2	$0.1	$1.2	-$1.1

In matrix notation, this would be summarized by saying it is an n by k (nxk or [n,k]) matrix. This means that there are N observations or rows and K variables or columns.

How is data stored and shared?

Data can come from many different sources. Sometimes, you create it, as is done to construct primary data sets. For example, you may run a survey, or you may scour records for information that you in turn input into a spreadsheet. Alternatively, you may access someone else's data, as is done with secondary data analysis. For this you, may be accessing data through a government's open data portal, accessing data via a dataverse, or using data provided by your work or supervisor.

Where is it stored?

There are two general ways to store data. When choosing where to store your data where you may look for already built data sets, there are two general places to run to: private files and resources (i.e., your local computer or cloud or your work) or public files and resources.

- Privately
    - When you're creating and working with a data set, you will likely locally save it (e.g., on your computer) or save it privately on a cloud (e.g., on Dropbox).
- Publicly
    - As the push for transparency in academia has increased, teams and individuals have made the data sets used in their research more available. While this is sometimes done via personal websites, many turn to Dataverses (https://dataverse.org/researchers), which have been built to store and catalog data sets. Three common Dataverses you may want to explore or house your data on upon completion of a project are: ICPSR's Dataverse (https://www.icpsr.umich.edu/web/pages/ICPSR/index.html), Harvard's Dataverse (https://dataverse.harvard.edu/), and Odum's Dataverse.
    - As the push for transparency in government has increased, bureaucratic agencies, government institutions, and governments generally have begun releasing troves of information and data sets. While this data may be housed on the same Dataverses that academics turn to, many are housed on open data platforms built specifically for them or housed on their specific website.
    - Note that you should not post a data set that has not been completed (i.e., constructed and cleaned) on the internet. Additionally, many wait to post new datasets online until they have published with it. Further some data will never be publicly posted, as it contains sensitive information.

How is it stored?

Data is stored in a variety of formats. This table, shown below, presents a quick summary of formats that one may encounter or chose between when storing the final version of your own data set. Two considerations when considering file formats is: how can the file be opened and may the format or platform affect the data stored within it.

File Extension	"Native" Program	What can (freely) open it?
Text Files		*Each is saved as a flat file (i.e., one sheet*)
.csv		Any. Text file with a comma delimiter.
.tab		Any. Text file with a tab delimiter.
.txt		Any. Text file delimited by another means.
Formatted Files		*Each can be opened by it's native program.*
.xlsx; .xls	Excel	Most statistical software. You will have to choose which "sheet" to open.
.rda; .RData	R	--
.sav	SPSS	R via a package.
.dta	Stata	R via a package.

Two additional things to note on file formats.

1. Regardless of the software you use (e.g., excel, R, etc.), you typically need to know what type of file it is in order to open it appropriately.
2. While Excel offers a lot of great functionality, the built-in short cuts can negatively affect studies. For examples, see: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7; https://www.bloomberg.com/news/articles/2013-04-18/faq-reinhart-rogoff-and-the-excel-error-that-changed-history; https://retractionwatch.com/2016/08/17/doing-the-right-thing-authors-pull-psych-review-after-finding-inaccuracies/. If you do use Excel for your primary data set construction, management, and analysis, see this article for practical advice on best practices: https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989.

What documentation should be provided or created?

Every data set should have accompanying documentation that provides information on where the data came, what variables are included in the data set, and basic information about those variables. Additionally, your documentation may have a short overview section that gives background information on the data set. The goal of a documentation and a codebook is to communicate the contents of a data set, such that the data set's creator and researchers more generally can understand what's in the data set at a later date.

Let's dig into each of these pieces a bit more.

- Overview
    - Depending on how long your codebook is and whether the data set being described fits into a broader project with multiple data sets, you may want to include an overview section.
    - This section should contextualize where this data set sits in a broader project (if applicable), provide basic information on what information the data set contains and how much information the data set contains, and a quick explanation of where the data came from.
    - Think of this like an abstract for your data set and codebook.

- Data Set Construction
    - Next you should explain where the information contained within the data set came from and what if any cleaning of the data set occurred.
    - To the first point, you should provide enough detail, such that if someone wanted to replicate the process they could. For example, for surveys, this might

be explaining when and how the survey was fielded and providing basic information about the survey (e.g., number of respondents). For data sets built from searches on private platforms, such as LexisNexis, CrowdTangle, or BrandWatch, you should provide information on each of the searches constructed, how those searches were constructed (when appropriate), what data is actually being made public (if any) from those sources, and how the final data set was constructed from the downloaded output. If you accessed a publicly available data set, you should explain the process by which one gains access. If you compiled data from multiple sources, you should list and explain those sources.
    - To the second point, regardless of where the information initially came from, you need to detail either how information was compiled from multiple sources and/or how the data was cleaned. This should be done in enough detail that if someone else took up the raw data they would produce the same data set as you. This may also take the form of discussing the process in broad strokes and then pointing to code that will transform the raw data into the cleaned data set.

- Variable Summary (Optional)

    - Some codebooks provide a table that researchers can quickly reference that provides the variable name as it appears in the data set next to its real name, which is understandable to yourself and other people (i.e., it's long name or description).

- Variable Definitions (Mandatory)

    - List each variable name as it appears in the data set.
    - For each variable, include a description of the variable and indicate the values that the variable can take one. If the variable is categorical, provide a list of the possible categories and labels for those categories (if applicable).

There is a plethora of information on what should be included in codebooks and how to construct them on the internet. To get you started on sorting through and finding these resources, check out the following:

- UCLA's Social Science Data Archive Dataverse document "All About Codebooks:" https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/U5FWYL
- ICPSR developed a page on what a codebook is: https://www.icpsr.umich.edu/icpsrweb/content/shared/ICPSR/faqs/what-is-a-codebook.html

For examples of thorough codebooks, explore the NAJCD's dataverse housed by ICPSR (https://www.icpsr.umich.edu/web/pages/NACJD/index.html) or explore the Congressional Election Studies documentation (https://cces.gov.harvard.edu/).