

Passion Driven Statistics (Light Weight)

Alan T. Arnholt modified from materials compiled by Lisa Dierker

2015-08-31

Contents

Chapter 1	1
An Introduction	1
Chapter 2	7
Data Sets and Code Books	7
Chapter 3	9
Data Architecture	9
Chapter 4	12
Conducting a Literature Review	12

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Please report any mistakes/typos/errata, suggestions, or problems at <https://github.com/alanarnholt/PDS-Book/issues>.

Chapter 1

An Introduction

Overview

This statistics course is presented in the service of a project of your choosing and will offer you an intensive hands-on experience in the quantitative research process. You will develop skills in 1) generating testable hypotheses; 2) understanding large data sets; 3) formatting and managing data; 4) conducting descriptive and inferential statistical analyses; and 5) presenting results for expert and novice audiences. It is designed for students who are interested in developing skills that are useful for working with data and using statistical tools to analyze them. No prior experience with data or statistics is required.

Our approach is “statistics in the service of questions.” As such, the research question that you choose (from data sets made available to you) is of paramount importance to your learning experience. **It must interest you enough that you will be willing to spend many hours thinking about it and analyzing data having to do with it.**

Resources

This course is unlike any you have likely encountered in that you will be driving the content and direction of your own learning. In many ways we will be asking more from you than any other introductory course ever has. To support you in this challenge, there are a number of useful resources.

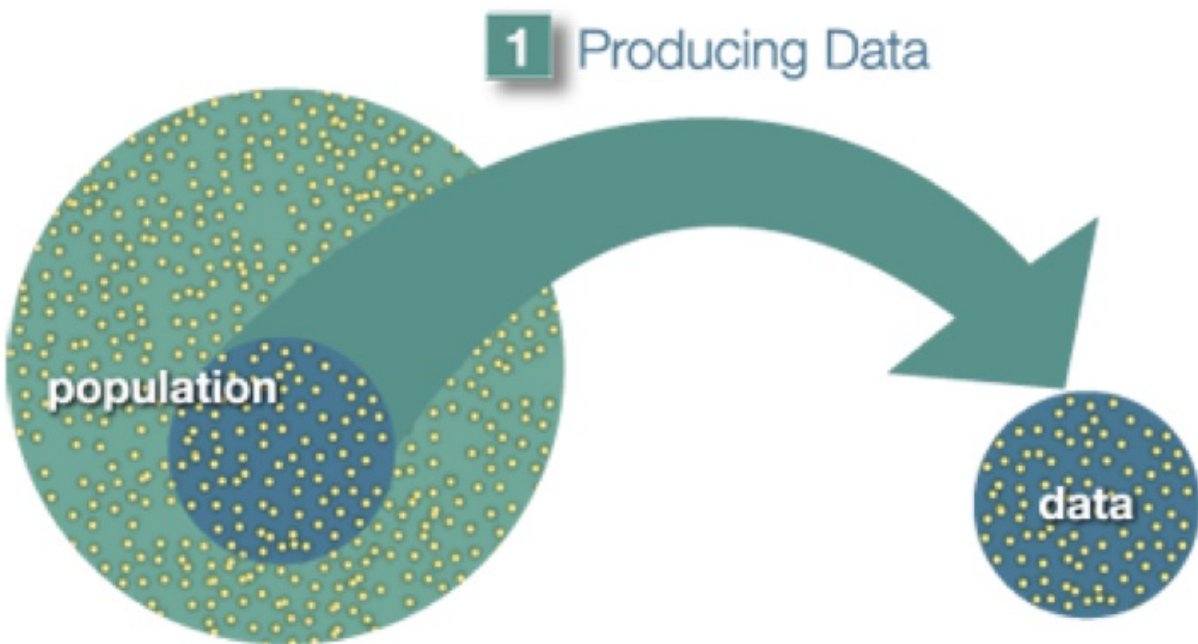
This Book: This book integrates the applied steps of a research project with the basic knowledge needed to meaningfully engage in quantitative research. Much of the background on descriptive and inferential statistics has been drawn from the [Open Learning Initiative](#), a not-for-profit educational project aimed at transforming instruction and improving learning outcomes for students.

Empowerment Through Statistical Computing: While there is widespread argument that introductory students need to learn statistical programming, opinions differ widely both within and across disciplines about the specific statistical software program that should be used. While many introductory statistics courses cover the practical aspects of using a single software package, our focus will be more generally on computing as a skill that will expand your capacity for statistical application and for engaging in deeper levels of quantitative reasoning. Instead of providing “canned” exercises for you to repeat, you will be provided with flexible syntax for achieving a host of data management and analytic tasks in the pursuit of answers to questions of greatest interest to you. Most importantly, syntax for R will be presented in the context of each step of the research process.

Loads of Support: Through the in-class workshop sessions and peer group exchanges, a great deal of individualized support will be available to you. Taking advantage of this large amount of support means that you are succeeding in making the most of your experience in this course.

GitHub Repository: To provide reliable backup of your work, you will use a private GitHub repository. While you will have read/write access to your own repository, you will also have read access to all public repositories in the organization (Course). Aside from providing a centralized way to share files, GitHub is meant to function as a resource in support of collaboration. Put simply, our hope is that you work together!

An Introduction to Statistics¹



¹<https://oli.cmu.edu/jcourse/workbook/activity/page?context=434b846480020ca6018dda7ea62a2528>

Statistics plays a significant role across the physical and social sciences and is arguably the most salient point of intersection between diverse disciplines given that researchers constantly communicate information on varied topics through the common language of statistics. In a nutshell, what statistics is all about is converting data into useful information. Statistics is therefore a process where we are:

- Collecting Data
- Summarizing Data, and
- Interpreting Data

The process of statistics starts when we identify what group we want to study or learn something about. We call this group the **population**. Note the word “population” here (and in the entire course) is not just used to refer to people; it is used in the more broad statistical sense, where population can refer not only to people, but also to animals, things, etc. For example, we might be interested in:

- The opinions of the population of U.S. adults about the death penalty
- How the population of mice react to a certain chemical
- The average price of the population of all one-bedroom apartments in a certain city

Population, then, is the entire group that is the target of our interest. In most cases, the population is so large that as much as we want to, there is absolutely no way that we can study all of it (imagine trying to get opinions of all U.S. adults about the death penalty. . .).

A more practical approach would be to examine and collect data only from a sub-group of the population, which we call a sample. We call this first step, which involves choosing a sample and collecting data from it, *Producing Data*.

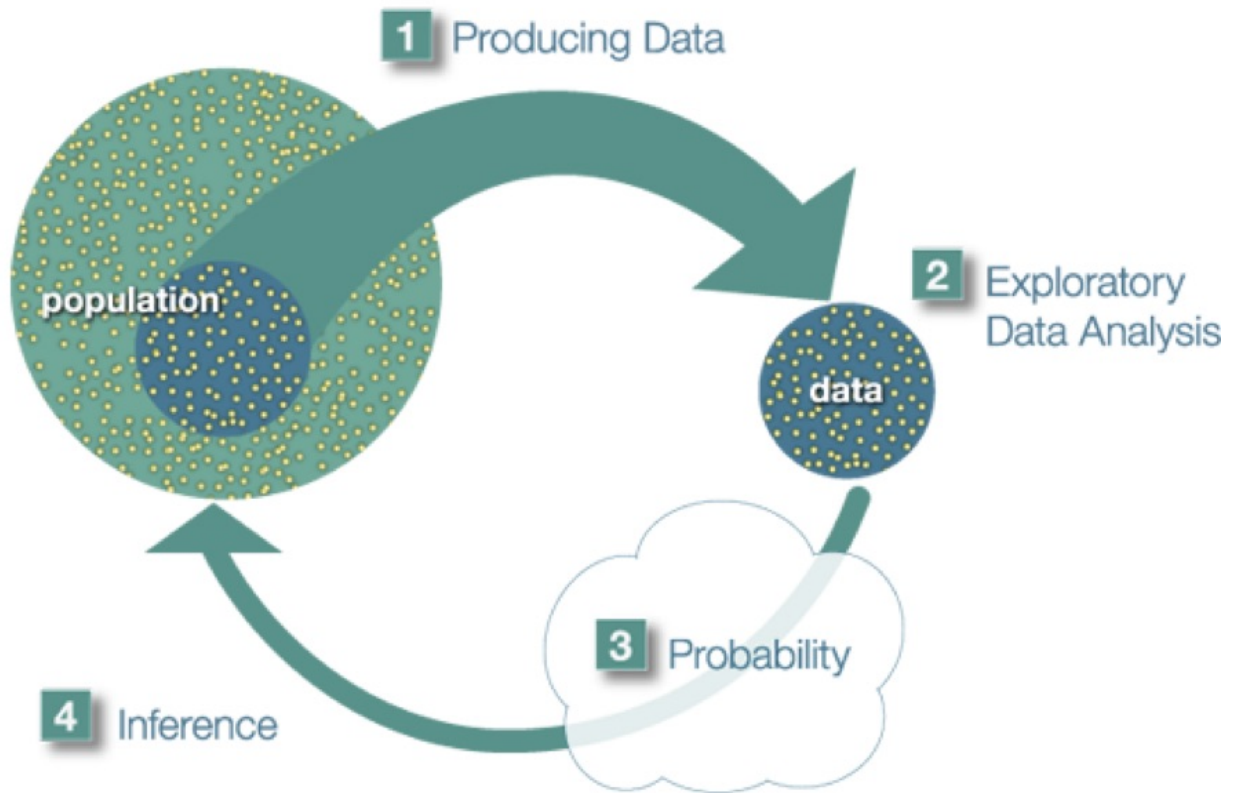
Since, for practical reasons, we need to compromise and examine only a sub-group of the population rather than the whole population, we should make an effort to choose a sample in such a way that it will represent the population as well.

For example, if we choose a sample from the population of U.S. adults, and ask their opinions about the death penalty, we do not want our sample to consist of only Republicans or only Democrats.

Once the data have been collected, what we have is a long list of answers to questions, or numbers, and in order to explore and make sense of the data, we need to summarize that list in a meaningful way. This second step, which consists of summarizing the collected data, is called *Exploratory Data Analysis*.

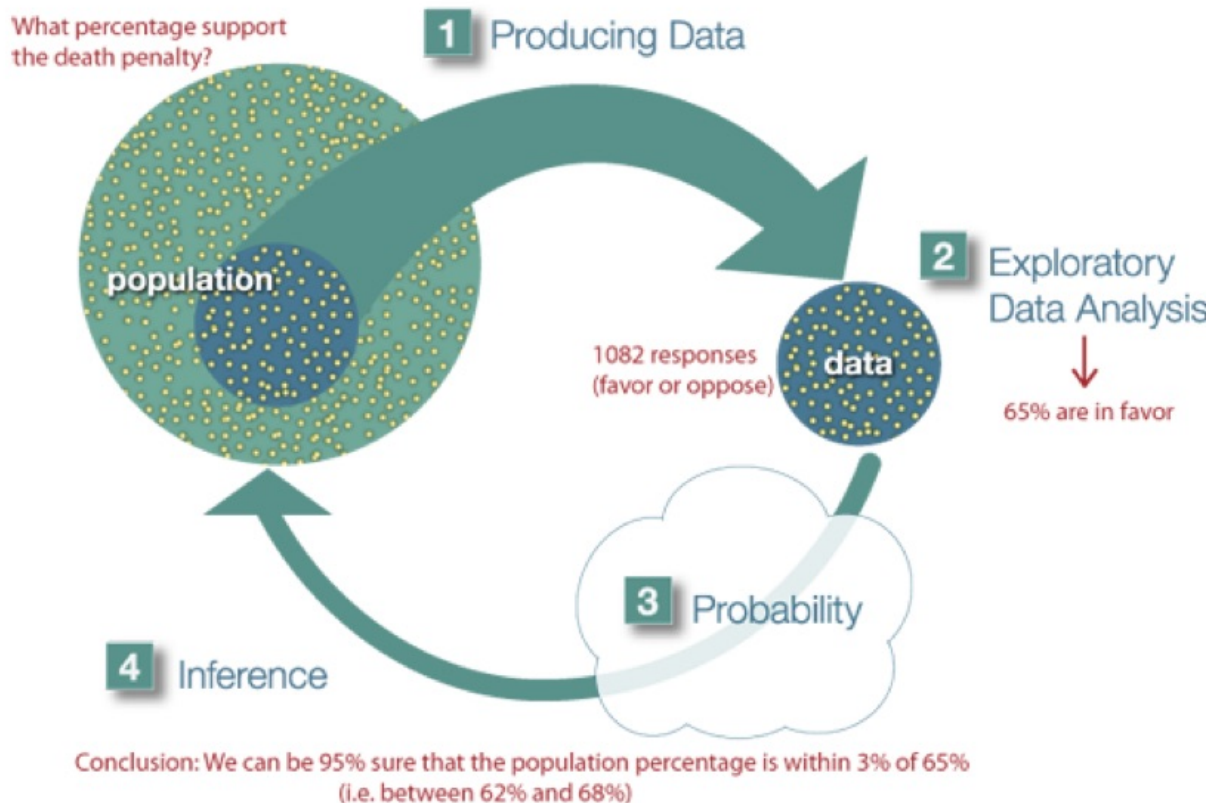
Now we’ve obtained the sample results and summarized them, but we are not done. Remember that our goal is to study the population, so what we want is to be able to draw conclusions about the population based on the sample results. Before we can do so, we need to look at how the sample we’re using may differ from the population as a whole, so that we can factor that into our analysis. Finally, we can use what we’ve discovered about our sample to draw conclusions about our population. We call this final step Inference. This is the *Big Picture of Statistics*.

Since we will be relying on data that has already been produced, the focus of your individual project will be *exploratory* and *inferential* data analysis.



Example At the end of April 2005, a poll was conducted (by ABC News and the Washington Post), for the purpose of learning the opinions of U.S. adults about the death penalty.

1. **Producing Data:** A (representative) sample of 1,082 U.S. adults was chosen, and each adult was asked whether he or she favored or opposed the death penalty.
2. **Exploratory Data Analysis (EDA):** The collected data was summarized, and it was found that 65% of the sample's adults favor the death penalty for persons convicted of murder.
3. **Inference:** Based on the sample result (of 65% favoring the death penalty), it was concluded (within 95% confidence) that the percentage of those who favor the death penalty in the population is within 3% of what was obtained in the sample (i.e., between 62% and 68%). The following figure summarizes the example:

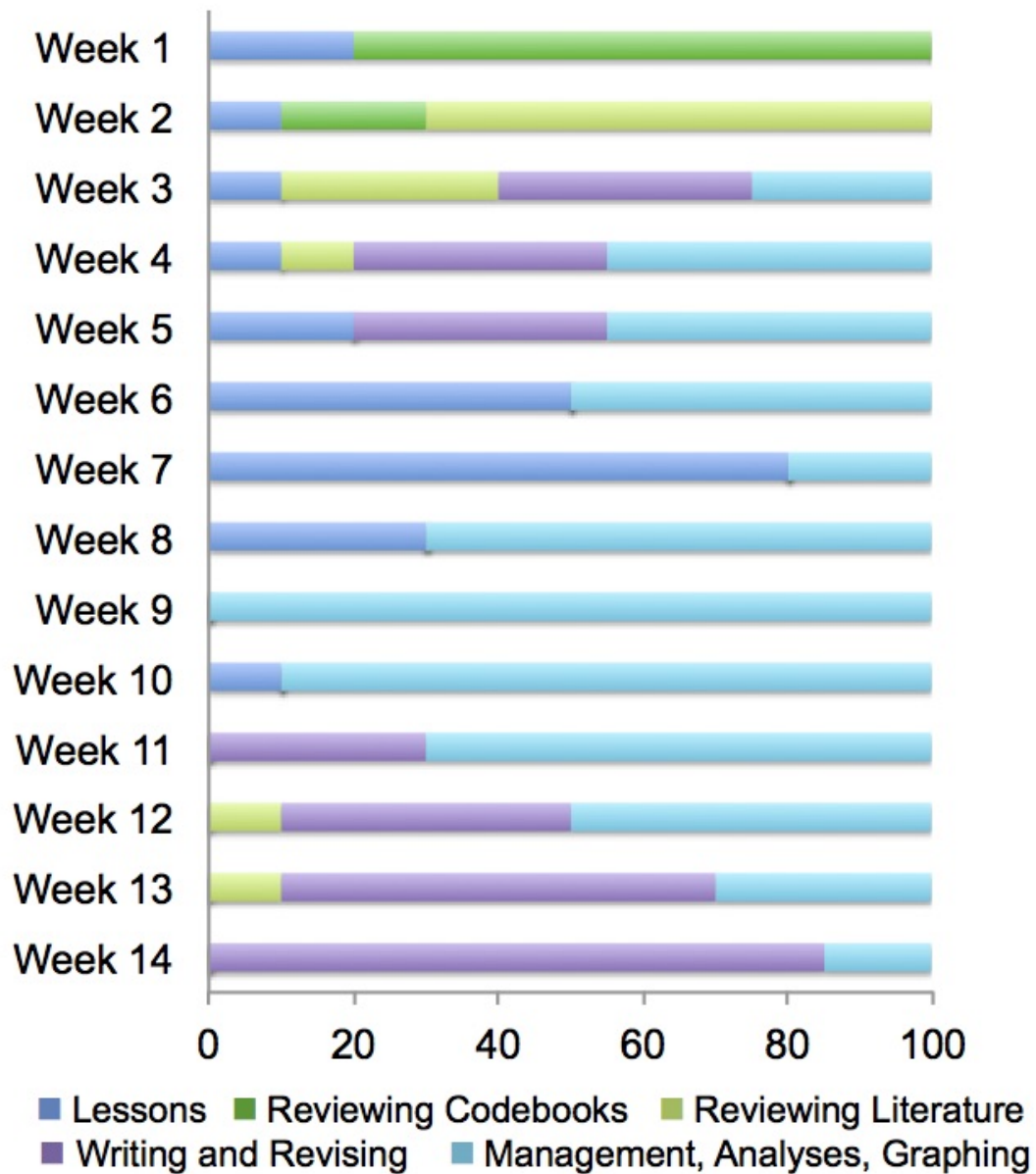


Final Notes:

Statistics education is often conducted within a discipline specific context or as generic mathematical training. Our goal is instead to create meaningful dialogue across disciplines. Ultimately, this experience is aimed at helping you on your way to engaging in interdisciplinary scholarship at the highest levels.

MANAGING YOUR TIME

You will spend your time in this course working in a variety of ways. You will need to review codebooks and literature, perform data management, do analyses, make graphs, write and revise your findings and read and/or watch IBook lessons.



Chapter 2

Data Sets and Code Books

Since we will not be producing data for this course, the first step of your project will be to choose a data set (from those made available) that offers the opportunity to conduct research on a general topic that will be of significant interest to you.

A full list will be presented in class. Here are a few examples:

The U.S. National Longitudinal Survey of Adolescent Health (AddHealth) is representative school-based survey of adolescents in grades 7-12 in the United States. (Wave I and Wave IV)

The U.S. National Epidemiological Survey on Alcohol and Related Conditions (NESARC) is a survey designed to determine the magnitude of alcohol use and psychiatric disorders in the U.S. population. It is a representative sample of the non-institutionalized population 18 years and older.

The Mars Craters Study (<http://craters.sjrdesign.net>) created by Stuart Robbins, presents a global database that includes over 300,000 Mars craters 1 km or larger. Heavily cratered terrain on Mars was created between 4.2 and 3.8 billion years ago during a period of heavy bombardment (i.e. impacts of asteroids, proto-planets, and comets). Mars craters allow inferences into the ancient climate of Mars, and they add a key data point for the understanding of impact physics.

Integrated Post-secondary Education Data System (IPEDS) is the primary source for data on colleges, universities, and technical and vocational postsecondary institutions in the United States.

Outlook On Life Surveys (OOL)

Code Books

Before accessing any data, you will be reviewing the available codebooks (sometimes called “data dictionaries”). Codebooks commonly offer complete information regarding the data set (e.g. general topics addressed, questions and/or measurements used, and in some cases the frequency of responses or values). Reviewing a code book is always the first step in research based on existing data since 1) code books can be used to generate research questions; and 2) data is generally useless and uninterpretable without it.

The code book describes how the data are arranged in the computer file or files, what the various numbers and letters mean, and any special instructions on how to use the data properly. Like any other kind of book, some codebooks are better than others.

INTERACTIVE 2.1 Example of AddHealth Code Book

Section 22: Romantic Relationship Roster

In Section 22 the respondent identifies as many as three recent romantic relationships.

1. In the last 18 months—since {MONTH, YEAR}—have you had a special romantic relationship with any one?			H1RR1 3	num 1
1 2878	2 0	no [skip to the next section]		
3582	1	yes		
22	6	refused [skip to the next section]		
20	8	don't know [skip to the next section]		
2	9	not applicable [skip to the next section]		
A flag indicating respondents who answered “yes” to Q.1 below but who do NOT have data for any romantic relationships in Section 25 due to a programming error.			RR_FLAG	num 1
6481	0	skips followed correctly		
23	1	skips NOT followed correctly		
The next part of the interview is about your romantic relationships.				
2. Please tell me the first and last initials of each person with whom you have had a special romantic relationship in the last 18 months. When you have finished this part of the interview, all the initials will be erased from the computer. You can list boys and girls.				
		[list of initials of up to 3 romantic partners]		
INTERVIEWER: Record any comments R may have about additional relationships below. Do not record additional initials of romantic partners.				
If no initials recorded, go to Section 23: Liked Relationship Roster. Otherwise, skip to Section 24: Contraception.				

1. Number of Observations - This is the number of participants who answered the question “no”

2. Numerical Value for Answer - “0” is the numerical value for the answer “no”

3. Variable Name - “H1RR1” is the name of the variable

Selecting A Data Set At this point, you should review the available codebooks for the data set that most interests you. The **PDS²** R package has the codebooks and data sets for this course.

²<https://github.com/alanarnholt/PDS>

Selecting A Data Set Assignment Select a data set that you will work with. Add the abbreviated title of that data set (i.e. AddHealth, NESARC, Mars Crater, IPEDS, or OOL) to the README file of your GitHub repository.

Chapter 3

Data Architecture³

What do we really mean by data?

Data are pieces of information about individuals or observations organized into variables. By an *individual* or *observation*, we mean a particular person or object. By a *variable*, we mean a particular characteristic of the individual or observation.

A dataset is a collection of information, usually presented in tabular form. Each column represents a particular *variable*. Each row corresponds to a given individual (or *observation*) within the dataset.

Relying on datasets, statistics pulls all of the behavioral, physical and social sciences together. It's arguably the one language that we all have in common. While you may think that data is very, very different from discipline to discipline, it is not. What you measure is different, and your research question is obviously dramatically different; whom you observe and whom you collect data from – or what you collect data from – can be very different, but once you have the data, approaches to analyzing it statistically are quite similar regardless of individual discipline.

Example: Medical Recordings

The following dataset shows medical records from a particular survey:

Variables							
	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (0=No, 1=Yes)	Race	
Individuals	Patient #1	M	59	175	69	0	White
	Patient #2	F	67	140	62	1	Black
	Patient #3	F	73	155	59	0	Asian

	Patient #75	M	48	190	72	0	White

³<https://oli.cmu.edu/jcourse/workbook/activity/page?context=434b846d80020ca60084af88b54dce2b>

In this example, the individuals are patients, and the variables are Gender, Age, Weight, Height, Smoking, and Race. Each row, then, gives us all the information about a particular individual or observation (in this case, patient), and each column gives us the information about a particular characteristic of all the patients.

Variables can be classified into one of two types: quantitative or categorical.

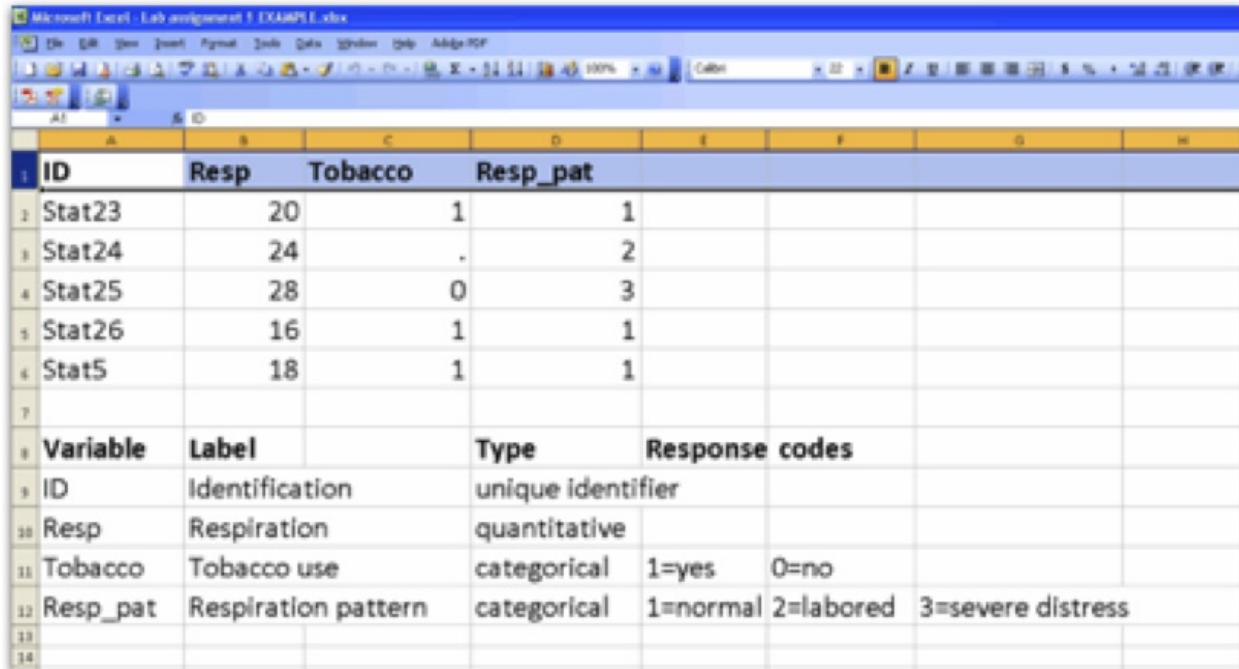
- *Quantitative Variables* take numerical values and represent some kind of measurement
- *Categorical Variables* take category or label values and place an individual into one of several groups. In our example of medical records, there are several variables of each type:
 - Age, Weight, and Height are **quantitative** variables
 - Race, Gender, and Smoking are **categorical** variables

Notice that the values of the categorical variable, Smoking, have been coded as the numbers 0 or 1. It is quite common to code the values of a categorical variable as numbers, but you should remember that these are just codes (often called dummy codes). They have no arithmetic meaning (i.e., it does not make sense to add, subtract, multiply, divide, or compare the magnitude of such values.) A unique identifier is a variable that is meant to distinctively define each of the individuals or observations in your data set. Examples might include serial numbers (for data on a particular product), social security numbers (for data on individual persons), or random numbers (generated for any type of observations). Every data set should have a variable that uniquely identifies the observations. In this example, the patient number (1 through 75) is a unique identifier.

Medical Records Assignment Although you will be working with previously collected data, it is important to understand what data looks like as well as how it is coded and entered into a spreadsheet or dataset for analysis. Using medical records for 5 patients seeking treatment in a hospital emergency room.

1. Select 4 variables recorded on the medical forms (one should be a unique identifier, at least one should be a quantitative variable and at least one should be a categorical variable)
2. Select a brief name (ideally 8 characters or less) for each variable
3. Determine what range of values is needed for recording each variable (create dummy codes as needed)
4. Label variables within an Excel spreadsheet
5. Enter data for each patient in the Excel spreadsheet
6. List the variable names, labels, types, and, response codes below the data set (i.e. the code book).
7. Push the Excel spreadsheet to your private GitHub repository.

Model:



ID	Resp	Tobacco	Resp_pat
Stat23	20	1	1
Stat24	24	.	2
Stat25	28	0	3
Stat26	16	1	1
Stat5	18	1	1

Variable	Label	Type	Response codes
ID	Identification	unique identifier	
Resp	Respiration	quantitative	
Tobacco	Tobacco use	categorical	1=yes 0=no
Resp_pat	Respiration pattern	categorical	1=normal 2=labored 3=severe distress

Personal Code Book (AKA Research Question) Assignment One of the simplest research questions that can be asked is whether two constructs are associated. For example: a) Is medical treatment seeking associated with socio-economic status?

- b) Is water fluorination associated with number of cavities during dentist visits?
- c) Is humidity associated with caterpillar reproduction?

Example:

After looking through the codebook for the NESARC study, I have decided that I am particularly interested in nicotine dependence. I am not sure which variables I will use regarding nicotine dependence (e.g. symptoms or diagnosis) so for now I will include all of the relevant variables in my personal codebook.

At this point, you should continue to explore the code book for the data set you have selected.

After choosing a data set, you should:

1. Identify a specific topic of interest
2. Prepare a codebook of your own (i.e., print individual pages (or copy screen and paste into a new document) from the larger codebook that includes the questions/items/variables that measure your selected topics.)

Example:

While nicotine dependence is a good starting point, I need to determine what it is about nicotine dependence that I am interested in. It strikes me that friends and acquaintances that I have known through the years that became hooked on cigarettes did so across very different periods of time. Some seemed to be dependent soon after their first few experiences with smoking and others after many years of generally irregular smoking behavior. I decide that I am most interested in exploring the association between level of smoking and nicotine dependence. I add to my codebook variables reflecting smoking levels (e.g. smoking quantity and frequency).

During a second review of the codebook for the dataset that you have selected, you should:

1. Identify a second topic that you would like to explore in terms of its association with your original topic
2. Add questions/items/variables documenting this second topic to your personal codebook.

Following completion of the steps described above, show your instructor and peer mentor (in class) a hard copy of your personal codebook. **Keep this in a folder or binder for your own use throughout the course.**

Chapter 4

Conducting a Literature Review

At this point you have (1) generated a personal codebook reflecting variables of interest to you from your data set; and (2) selected an association that you would like to test. You are now ready to conduct a literature review using **primary source** journal articles (i.e. those reporting original research findings).

This [video](#) describes the nature and content of primary source journal articles. It highlights the importance of conducting a literature review before initiating a research project.

You should start your search using key words based on the two topics you have selected (note: search for their presence in the title of articles). You can then narrow your search as necessary based on the amount of relevant literature that you find. Although some libraries have extensive paper collections of journals, you should focus on articles available **online**. Secondary source literature including review articles and theoretical papers should be used only for needed background on a topic.

It is important to identify and review primary sources either through your on-line search or by using the reference list from primary or secondary sources.

It may also be useful to limit your search to journal articles published in the past 5 years.

Note that as you read the literature, there should be an exchange between your research question and what you are learning. **The literature review may cause you to add to the complexity of your research question, further focus that question, or even abandon the question for another.**

Literature Review

During your literature review, you should:

1. Identify primary source articles that address the association that you have decided to examine
2. Download relevant articles.
3. Read the articles that seem to test the association most directly.
4. Identify replicated and equivocal findings in order to generate a more focused question that may add to the literature. Give special attention to the “future research” sections of the articles that you read
5. Based on the literature, select additional questions/items/ variables that may help you to understand the association of interest. In doing so, further refine your research question. Add relevant documentation (i.e. code book pages) to your personal codebook.

Example:

Given the association that I have decided to examine, I use such keywords as nicotine dependence, tobacco dependence, and smoking. After reading through several titles and abstracts, I notice that there has been relatively little attention in the research literature to the association between smoking exposure and nicotine dependence. I expand a bit to include other substance use that provides relevant background as well.

References:

Caraballo, R. S., Novak, S. P., & Asman, K. (2009). Linking quantity and frequency profiles of cigarette smoking to the presence of nicotine dependence symptoms among adolescent smokers: Findings from the 2004 National Youth Tobacco Survey. *Nicotine & Tobacco Research*, 11(1), 49-57.

Chen, K., Kandel, D., (2002). Relationship between extent of cocaine use and dependence among adolescents and adults in the United States. *Drug & Alcohol Dependence*. 68, 65-85.

Chen, K., Kandel, D. B., Davies, M. (1997). Relationships between frequency and quantity of marijuana use and last year proxy dependence among adolescents and adults in the United States. *Drug & Alcohol Dependence*. 46, 53-67.

Decker, L., He, J. P., Kalaydjian, A., Swendsen, J., Degenhardt, L., Glantz, M., Merikangas, K. (2008). The importance of timing of transitions for risk of regular smoking and nicotine dependence. *Annals of Behavioral Medicine*, 36(1), 87-92.

Decker, L. C., Donny, E., Tiffany, S., Colby, S. M., Perrine, N., Clayton, R. R., & Network, T. (2007). The association between cigarette smoking and DSM- IV nicotine dependence among first year college students. *Drug and Alcohol Dependence*, 86(2-3), 106-114.

Lessov-Schlaggar, C. N., Hops, H., Brigham, J., Hudmon, K. S., Andrews, J. A., Tildesley, E., . . . Swan, G. E. (2008). Adolescent smoking trajectories and nicotine dependence. *Nicotine & Tobacco Research*, 10(2), 341-351.

Riggs, N. R., Chou, C. P., Li, C. Y., & Pentz, M. A. (2007). Adolescent to emerging adulthood smoking trajectories: When do smoking trajectories diverge, and do they predict early adulthood nicotine dependence? *Nicotine & Tobacco Research*, 9(11), 1147-1154.

Van De Ven, M. O. M., Greenwood, P. A., Engels, R., Olsson, C. A., & Patton, G. C. (2010). Patterns of adolescent smoking and later nicotine dependence in young adults: A 10-year prospective study. *Public Health*, 124(2), 65-70.

Based on my reading of the above articles as well as others, I have noted a few common and interesting themes:

1. While it is true that smoking exposure is a necessary requirement for nicotine dependence, frequency and quantity of smoking are markedly imperfect indices for determining an individual's probability of exhibiting nicotine dependence (this is true for other drugs as well)
2. The association may differ based on ethnicity, age, and gender (although there is little work on this)
3. One of the most potent risk factors consistently implicated in the etiology of smoking behavior and nicotine dependence is depression I have decided to further focus my question by examining whether the association between nicotine dependence and depression differs based on how much a person smokes. I am wondering if at low levels of smoking compared to high levels, nicotine dependence is more common among individuals with major depression than those without major depression.

I add relevant depression questions/items/variables to my personal codebook as well as several demographic measures (age, gender, ethnicity, etc.) and any other variables I may wish to consider.

Citation Assignment Describe the association that you have decided to examine and key words you found helpful in your search. List at least 5 of the more appropriate references that you have found and read (TO RECEIVE CREDIT, YOU MUST USE ZOTERO FOR THIS ASSIGNMENT). Describe findings and interesting themes that you have uncovered and list a tentative research question or two that you hope to pursue. Be brief and use bullets to cover these details. **The example above is a model for this assignment.** See the [LiteratureReview directory](#) for an example of the **Citation Assignment** done in R Markdown.
