STAT 140: Homework 1

YOUR NAME HERE

Due Date: Wednesday, 9/18/2019

BOOK EXERCISES (OpenIntro Statistics, Fourth Edition)

Section 1.2

1.3) **Air pollution and birth outcomes, study components.** Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide (CO) were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was collected for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.

a) Identify the main research question of the study.

b) Who are the subjects in this study, and how many are included?

c) What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

1.8) **Sinusitis and antibiotics, Part II.** Exercise 1.2 introduced a study exploring the effect of antibiotic treatment for acute sinusitis. Study participants received either a 10-day course of an antibiotic (treatment) or a placebo similar in appearance and taste (control). At the end of the 10-day period, patients were asked if they experienced improvement in symptoms. What are the explanatory and response variables in this study?

1.9) **Fisher's irises.** Sir Ronald Aylmer Fisher was an English statistician, evolutionary biologist, and geneticist who worked on a data set that contained sepal length and width, and petal length and width from three species of iris flowers (*setosa*, *versicolor*, and *virginica*). There were 50 flowers from each species in the data set.

a) How many cases were included in the data?

b) How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.

c) How many categorical variables are included in the data, and what are they? List the corresponding levels (categories).

Section 1.3

1.13) **Air pollution and birth outcomes, scope of inference.** Exercise 1.3 introduces a study where researchers collected data to examine the relationship between air pollutants and preterm births in Southern California.

a) Identify the population of interest and the sample in this study.

b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.17) **Relaxing after work.** The General Social Survey asked the question, "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours. Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter.

a) An American in the sample.

b) Number of hours spent relaxing after an average work day.

c) 1.65

d) Average number of hours all Americans spend relaxing after an average work day.

1.22) **Stressed out, Part I.** A study that surveyed a random sample of otherwise healthy high school students found that they are more likely to get muscle cramps when they are stressed. The study also noted that students drink more coffee and sleep less when they are stressed.

a) What type of study is this?

b) Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

c) State possible confounding variables that might explain the observed relationship between increased stress and muscle cramps.

1.24) **Random digit dialing.** The Gallup Poll uses a procedure called random digit dialing, which creates phone numbers based on a list of all area codes in America in conjunction with the associated number of residential households in each area code. Give a possible reason the Gallup Poll chooses to use random digit dialing instead of picking phone numbers from the phone book.

1.27) **Sampling strategies.** A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Various research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect.

a) He randomly samples 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.

b) He gives out the survey only to his friends, making sure each of them fills out the survey.

c) He posts a link to an online survey on Facebook and asks his friends to fill out the survey.

d) He randomly samples 5 classes and asks a random sample of students from those classes to fill out the survey.

Section 1.4

1.32) **Music and learning.** You would like to conduct an experiment in class to see if students learn better if they study without any music, with music that has no lyrics (instrumental), or with music that has lyrics. Briefly outline a design for this study.

1.37) **Chia seeds and weight loss.** Chia Pets - those terra-cotta figurines that sprout fuzzy green hair - made the chia plant a household name. But chia has gained an entirely new reputation as a diet supplement.

In one 2009 study, a team of researchers recruited 38 men and divided them randomly into two groups: treatment or control. They also recruited 38 women, and they randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant difference between the groups in appetite or weight loss.

a) What type of study is this?

b) What are the experimental and control treatments in this study?

c) Has blocking been used in this study? If so, what is the blocking variable?

d) Has blinding been used in this study?

e) Comment on whether or not we can make a causal statement, and indicate whether or not we can generalize the conclusion to the population at large.

1.40) **Income and education in US counties.** The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.

a) What are the explanatory and response variables?

b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.

c) Can we conclude that having a bachelor's degree increases one's income?

R Exercises

Leada. This site (https://www.teamleada.com/courses/r-bootcamp) has some free introductory lessons for R and R Studio. You still need to sign up, but you do not need to pay for anything. Complete the first lesson (An Introduction to R) for a more comprehensive introduction to R basics. Include a statement here indicating that you have completed the lesson.

Other Exercises. For these exercises, you will use a data set about peanut allergies that is provided through your textbook. Here is the description:

"In the late 1990s, it was believed that young children should exclude peanuts from their diets to reduce the chance of developing an allergy. In 2008, researchers were no longer so sure. This experiment evaluates whether peanut exposure is helpful or harmful. A peanut diet regimen (consume or avoid) was assigned to over 500 young children during years 2-5, ages during which children had previously been told not to eat peanuts. The key outcome was testing for a peanut allergy when each child turned 5."

1) **Exercise 1.** For this exercise, you should create a "setup" R chunk. This means that you should load any required packages and read in the data. To access the peanut allergy data, use the link: https://www.openintro.org/stat/data/?data=peanut_allergy.

- There are two ways you can read the data into your R session. For this exercise, I want you to download the CSV file and save it on your computer. Read it into your R session using the read.csv() function. The general structure is commented out in the chunk below:
- Load the ggplot2 package.
- Include comments for each line of code that you write, explaining what you are doing.

```
# peanut_allergy <- read.csv(file="", header=TRUE)
```

2) **Exercise 2.** Perform an exploratory analysis for the peanut allergy data set. Include comments for each line of code, as before. Using your code output from the R chunk below, answer the questions about the data set that follow.

- How many observations are included in this data set?
- How many variables are there? What types of variables are there and what are their names?

3) **Exercise 3.** The table() function allows us to summarize categorical data in a table. You can use it to find out how many observations of a categorical variable fall into one category versus another. You can also use it to look at the relationships between two categorical variables. Consider the following toy example.

Suppose I served two entrees, steak and fish, at my restaurant, and some of my guests reported nausea after they ate. I collected entree information on 50 randomly sampled patrons, and recorded whether they reported nausea. I want to examine the relationships between reported nausea (Yes/No) and entree (Steak/Fish). (Note, I made up these data and I don't own a restaurant, so you don't need to worry about getting food poisoning from it!)

```
## Creating the data set (you do not need to be able to do this on your own)
food_poisoning <- data.frame(Patron=1:50, Entree=sample(c("Steak","Fish"), 50, replace=TRUE),
                             Nausea=c(rep("Yes", 15), rep("No", 35)))

## Use table() to determine how many reports of nausea I have
table(food_poisoning$Nausea)
```

```
##
##  No Yes
##  35  15
```

```
## recall that the $ placed between the data set name and
##column/variable name lets us look at that column alone (this is a vector)

## Use table() to examine the relationship between nausea and entree choice
table(food_poisoning$Nausea, food_poisoning$Entree)
```

```
##
##       Fish Steak
##   No    15    20
##   Yes    8     7
```

```
## later in the semester we will learn how to test
## formally for a relationship for this kind of problem
```

- Now, for the peanut_allergy data set, use table() to examine the relationship between 'had_early_risk' and 'regimen'. What do you notice about the allocation of individuals to regimen based on risk? Do you think there is any evidence of stratified random sampling? If so, how were the strata determined? Include your R code in the chunk below, along with appropriate comments.

- You can also look at the relationships between all three variables, 'had_early_risk', 'regimen', and 'allergic', using the table command. Simply add 'allergic' into the table command as you did with the other variables (there should be three arguments to table() now). Describe what this gives you. Include your R code in the chunk below, along with appropriate comments.