Homework 2 – Solutions

Nathan Friedrichsen

Question 1 (5 pts)

This question is Question 4.2 from the textbook and has been reproduced here. The dataset below contains the results from a poll based on a random sample with two variables: response, indicating their response to the poll question, and political, reporting their self-reported political ideology.

Nine-hundred and ten (910) randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country.

```
## Copy and run this code to create tables
library(ggplot2)
library(dplyr)
immigration <- read.csv("https://collinn.github.io/data/immigrationpoll.csv")

# Proportion table (overall)
table(immigration) %>% proportions()
```

```
##                         political
## response                 conservative     liberal    moderate
##   Apply for citizenship  0.062637363 0.110989011 0.131868132
##   Guest worker           0.132967033 0.030769231 0.124175824
##   Leave the country      0.196703297 0.049450549 0.138461538
##   Not sure               0.016483516 0.001098901 0.004395604
```

```
# Proportion table (conditioned on political ideology)
table(immigration) %>% proportions(margin = 2) %>% addmargins(1)
```

```
##                         political
## response                 conservative     liberal    moderate
##   Apply for citizenship  0.153225806 0.577142857 0.330578512
##   Guest worker           0.325268817 0.160000000 0.311294766
##   Leave the country      0.481182796 0.257142857 0.347107438
##   Not sure               0.040322581 0.005714286 0.011019284
##   Sum                    1.000000000 1.000000000 1.000000000
```

Use the appropriate tables to answer the following questions:

a) What percent of these Tampa, FL voters identify themselves as conservatives?

40.1%

b) What percent of these Tampa, FL voters are in favor of the citizenship option?

30.5%

c) What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

6.02%

d) What percent of these Tampa, FL voters who identify themselves as conservatives are in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

15.3%, 33.1%, 57.7% (these are conditional probabilities – conditioned on ideology)

e) Do political ideology and views on immigration appear to be associated? Explain your reasoning.

Yes. Knowing a voters political ideology tells us about their likely views on citizenship.

Numerical Summaries

Question 2 – Conceptual Questions (1 pt each)

- **Part A** How does outlier classification differ between boxplots and histograms?

Outliers in histograms are where there are gaps in the data. Outliers in boxplots are determined by if the observation value is 1.5IQR above Q3 or below Q1.

- **Part B** What does it mean to say a statistic is *robust*?

It is not affected by skew or outliers very much.

- **Part C** Why do we use median and IQR (instead of mean and standard deviation) for the center and spread when we have skews, outliers, or both?

Because they are robust statistics. Mean and standard deviation are not good measures of center with skews/outliers because they change too much.

- **Part D** If a distribution is right-skewed with no outliers, which of these will be larger: the median or the mean?

The mean. The large values in the skew will make the mean larger than the median.

Question 3 (4 pts)

This question is 5.10-Question #22 from the textbook and has been reproduced here.

The first histogram below shows the distribution of the yearly incomes of 40 patrons at a college coffee shop. Suppose two new people walk into the coffee shop: one making $225,000 and the other $250,000. The second histogram shows the new income distribution. Summary statistics are also provided, rounded to the nearest whole number.

a) Would the mean or the median best represent what we might think of as a typical income for the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

The median best represents the typical income. The median is more robust because it gives us a better idea of center when outliers are present. The mean is not robust because outliers affect it a lot.

b) Describe the distribution of incomes for the coffee shop's 42 patrons using the bottom histogram ('After'). You may assume the value for Q3 is $68,000 (this may or may not be necessary).

The distribution is multimodal and symmetric with a couple of very large outliers. The median is $65,352 and the IQR is $7162 (68000 - 60838).

Study Design

Question 4 – Conceptual Questions (1 pt each)

- **Part A** What does it mean for a sample to be *representative* of a population?

A sample is representative if it has the same characteristics / looks the same as the population.

- **Part B** What is the difference between an Experiment and an Observational Study?

Experiments use random assignment.

- **Part C** What type of randomness do we need in a study in order to make cause-and-effect statements?

Random assignment.

- **Part D** What name do we give to variables that interfere with our ability to make cause-and-effect statements?

Confounding / Lurking variables.

Question 5 (4 pts)

The following is Chapter 2.5, Q3 from the textbook recreated.

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. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. (Ritz et al. 2000)

- **Part A** Identify the population of interest and the sample in this study.

Population: (Something like): All preterm babies in Southern California or All babies in the US.

Sample: 143,196 preterm babies in Southern California between 1989 and 1993.

- **Part B** Comment on whether 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.

Generalization: No, there was no random sample or otherwise stated reason why this is expected to be representative.

Causal relationships: No, this was an observational study and not an experiment.

Question 6 (3 pts)

The following is Chapter 2.5, Q12 from the textbook recreated.

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.

- **Part A** What type of study is this?

Observational Study.

- **Part B** Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

No. (There was no random assingment)

- **Part C** State possible confounding variables that might explain the observed relationship between increased stress and muscle cramps.

Answers will vary.