IS 606 - Statistics and Probability for Data Analytics – Fall 2016

Sharon Morris

Homework 1 – August 24, 2016

1.8 Smoking habits of UK residents. A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey.

(a) What does each row of the data matrix represent? *Each row in the data matrix represents information recorded about a single UK resident in the study and the associated variables in the columns*

(b) How many participants were included in the survey? *There are 1,691 participants included in the survey*

(c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal. Sex – Categorical - not ordinal Age – Numercial - discrete Marital – Categorical - not ordinal GrossIncome – Categorical - ordinal Smoke – Categorical, nor ordinal amtWeekends – Numerical discrete

1.10 Cheaters, scope of inference. Exercise 1.5 introduces a study where researchers studying the relationship between honesty, age, and self-control conducted an experiment on 160 children between the ages of 5 and 15. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. Di???erences were observed in the cheating rates in the instruction and no instruction groups, as well as some di???erences across children?s characteristics within each group.

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

Children of both genders between the ages of 5 and 15.

(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.

The results of this study cannot be gerneralized to the population. There is no mention as to whether the sampling methodolgy and the sample sizes might too small to be stable. There are several other variables that would need to be controled for that were not collected based on the description. Even though there is a control and experimental group there isn't enough information about the variables that are controlled to measure self control.

1.28 Reading the paper. Below are excerpts from two articles published in the NY Times: (a) Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.

We cannot conclude that smoking causes dementia. There appears to be an association but not causation. The sample for this study may not be representative of the population of smokers – it was not randomly selected. Since the sample was not randomly selected and even with the adjustments made by the researchers there could be confunding varialbles at play.

(b) Another article titled The School Bully Is Sleepy states the following

A friend of yours who read the article says, ?The study shows that sleep disorders lead to bullying in school children.? Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study?

No, the statement is not justified. It is not clear that sleep disorders causes bullying versus bullying causes sleep disorders. There maybe some association but factors were not controlled to determine the direction. Additionally, there can be some bias introduced by the parents and/or researchers.

1.36 Exercise and mental health. A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of

18-30, 31-40 and 41- 55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

(a) What type of study is this? Stratified sampling

(b) What are the treatment and control groups in this study? The treatment group that exercised twice a week The control group is the group that was told not to exercise

(c) Does this study make use of blocking? If so, what is the blocking variable? Yes, this study does use blocking. The blocking variable is age.

(d) Does this study make use of blinding? No this study does not make use of blinding – subjects were instructed to exercise or not to excise

(e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.

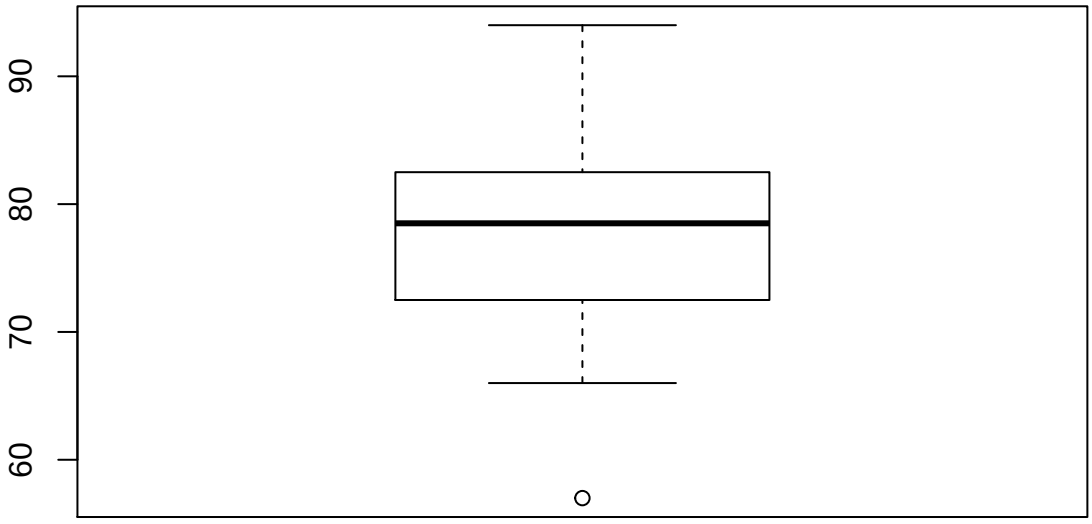
A causal relationship between exercise and mental health cannot be generalized to the larger population. The methodolgy used in this study is simplistic and there could be confunding variables that were not controlled for. The conclusions of this study cannot be gerneralized to the population at large.

(f) Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal? Yes, I would have reservations, My reservations would be around the study design and lack of control.

1.48 Stats scores. Below are the final exam scores of twenty introductory statistics students.

```
scores <- c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94)
```

```
boxplot(scores)
```




```
summary(scores)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   57.00   72.75   78.50   77.70   82.25   94.00
```

1.50 Mix-and-match. Describe the distribution in the histograms below and match them to the box plots.

a) This histogram is is unimodal and looks like a normal distribution – with most of the frequency counts bunched in the middle and the counts dying off out in the tails. a matches 2

b) This histogram is multimodal wit several frequency counts bunched b matches 3

c) This histogram is skewed to the right c matches 1

1.56 Distributions and appropriate statistics, Part II. For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

The data are skewed to the right The mean or median will be the best observation of the data since it is skewed. The IQR can be used for similar reasons

(b) Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.

The data reparesent a normal distribution. The mean, mode or media can be used since the date are normalluy distributed. The standard diviation or IQR can be used to describe the data.

(c) Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don?t drink since they are under 21 years old, and only a few drink excessively.

The data are skewed to the right since few students drink. The median and IQR are thebest measures for this data.

(d) Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than the all other employees.

The data are skewed to the right. The median is the best measure to describe the data.

1.70 Heart transplants. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable transplant indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Another variable called survived was used to indicate whether or not the patient was alive at the end of the study.

(a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant?

Based on the mosiac plot, those who received a transplant had a greater chanse of surviving. Thus, the 2 variables are not independent.

(b) What do the box plots below suggest about the efficacy (e???ectiveness) of the heart transplant treatment.

The box plots suggest there was a relatively even distribution of increased survival times with the introdcution of the treatment. The box plot also provide more information than the mosiac plot, it illustrates the distribution is stretched outward. Survival times applear to be a normal distribution.

(c) What proportion of patients in the treatment group and what proportion of patients in the control group died?

Approximately, 1/8th of the control group survived, and roughtly 1/3rd of the treatment group survived.

(d) One approach for investigating whether or not the treatment is e???ective is to use a randomization technique.

i. What are the claims being tested?

The claims being tested are the null hypothese, having a transplant has no effect on survival reates and the alternative hypotheses, having a transplant leads to higher survival rates.

ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

We write alive on alive *some* cards representing patients who were alive at the end of the study, and dead on *the remaining* cards representing patients who were not. Then, we shu???e these cards and split them into two groups: one group of size *of the original* representing treatment, and another group of size *original control* representing control. We calculate the di???erence between the proportion of dead cards in the treatment and control groups (treatment - control) and record this value. We repeat this 100 times to build a distribution centered at *0*. Lastly, we calculate the fraction of simulations where the simulated di???erences in proportions are *greater than the observed difference*. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis should be rejected in favor of the alternative.

iii. What do the simulation results shown below suggest about the e???ectiveness of the transplant program?

The simulation results shown below suggest the difference in proportions is 1/3rd-1/8th = 0.2083. The majority of the proportions are below the proportions above – we reject the null hypothese, heart transplant does have an impact on survival rates.