HW1

Dhananjay Kumar

September 2, 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 in the data matrix represent? Each row in the data matrix represents a case, a.k.a. a unit of observation or observational unit (I've always said just "observation")

b. How many participants were included in the survey? From the row labels, it looks like 1691.

c. Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, iindicate if the variable is ordinal.

Solution: a) Each row represents a record or observation related to the smoking habits of UK residents. b) 1691 participants c) Sex : categorical, not ordinal age: numerical, discrete marital: categorical, not ordinal gross income: categorical, ordinal smoke: categorical, not ordinal amtWeekends: numerical, discrete amtWeekdays: 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 researches 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. Differences were observed in the cheating rates in the instruction and no instructions groups, as well as some differences across children's characteristics within each group.

a. Identify the population of interest and the sample in this study - In this case the population of interest are all children 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.

Solution: a) The population of interest are children aged between 5 and 15. The sample in this study are 160 children. b) It is difficult to generalize the results, because it depends on how the sample was chosen. To generalize the results, the sample must be a true representation of the population. There isn't sufficient information to validate if the sample was unbiassed. We cannot establish the causal relationships by the findings, we can only establish that there is association between them.

1.28: Reading the paper

Below are exerpts from two articles published in the NY Times:

(a) An article titled Risks: Smokers Found More Prone to Dementia states the following: "Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer's disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a- day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs."Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.

(b) Another article titled The School Bully Is Sleepy states the following: "The University of Michigan study, collected survey data from parents on each child's sleep habits and asked both parents and teachers to assess behavioral concerns. About a third of the students studied were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues and those who were identified as bullies were twice as likely to have shown symptoms of sleep disorders." 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?

Solution: a) We cannot conclude smoking causes dementia later in life, however there seems correlation between them. Correlation doesn't necessary mean causation. b) The statement is not justified. It just shows an association between sleep disorders and bullying in school. There may be other factors involved in bullying however we cannot establish a causal relationship between the two.

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 proporations 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?

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

c. Does this study make use of blocking? If so, what is the blocking variable?

d. Has blinding been used in this study?

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.

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?

Solution: a) The study is a randomized experiment. b) *Treatment group:* group asked to exercise *control group:* group asked not to exercise c) Yes this study makes use of blocking. Age is the blocking variable. d) No, this study doesn't make use of blinding. e) As long as the sample is large enough and true representation of the population, we can make inference out of the results. f) As long as condition mentioned in Answer (e) is satisfied, I would favor funding for the proposed study.

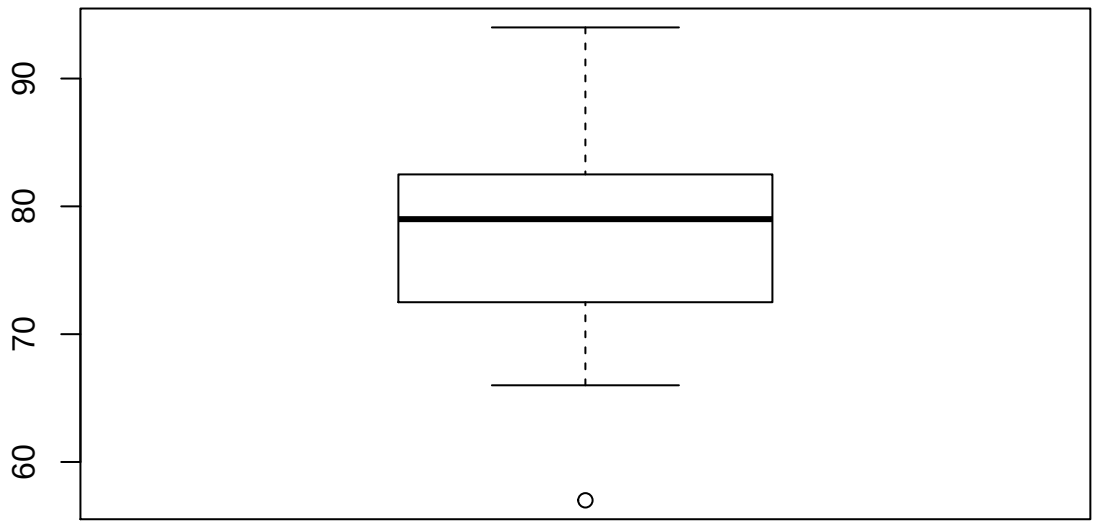
1.48: Stats scores

Below are the final exam scores of twenty introductory statistics students:

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

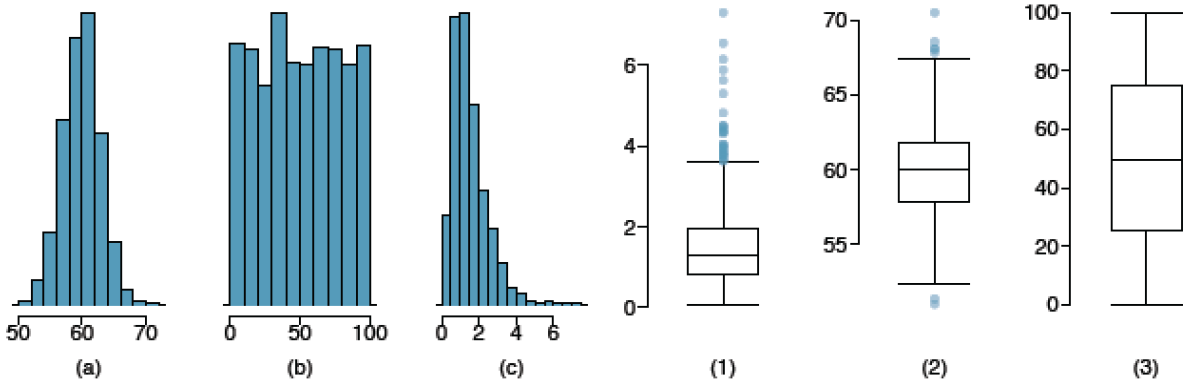
Create a box plot of the distribution of these scores. The five number summary provided below may be useful:

```
boxplot(scores)
```



1.50: Mix-and-match.

Describe the distribution in the histograms below, and match them to the box plots.



Solution: a matches with 2, b matches with 3, and c matches with 1.

a is unimodal, b is multimodal, and c is unimodal but skewed to the right.

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 of 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.

This data is skewed to the right. Because of the outlier houses, it's probably better to use the median when looking for a typical observation above the mean (which whould be skewed by the outliers) and better to use the IQR for the same reason.

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.

This data seems very evenly distributed, mean and median should be roughly the same, and the SD or IQR can both be used to accurately 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.

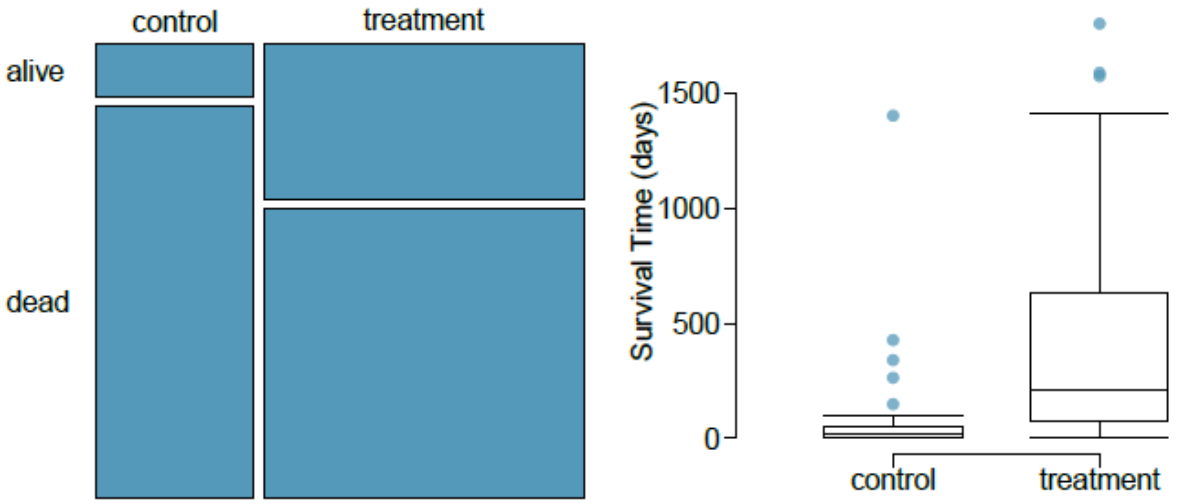
Because of the outliers, once again IQR and Median should be used over the Standard Deviation and the mean. The data is going to be skewed to the right because of the few students who drink excessively.

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

Similarly to the above, this data is skewed to the right, and median based measures should be used over the standard deviation and mean 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 teh 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.Of the 34 patients in the control group, 30 died. Of the 69 people in the treatment group, 45 died.



a. Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

Solution: Based on Mosaic chart, it seems survival is not independent of whether or not the patient got a transplant. Those who received a transplant had greater chance of surviving thus suggesting that the two variables aren't independent.

b. What do the box plots below suggest about the efficacy of th heart transplant treatment?

Solution: According to the box plots, with the treatment the survival time significantly increated which was also supported by Mosaic Plot

c. What proporation of patients in the treatment group and what proporation of patients in the control group died?

Solution: Proportion(Treatment): 45/69 or 0.652 Proportion(Control): 30/34 or 0.882

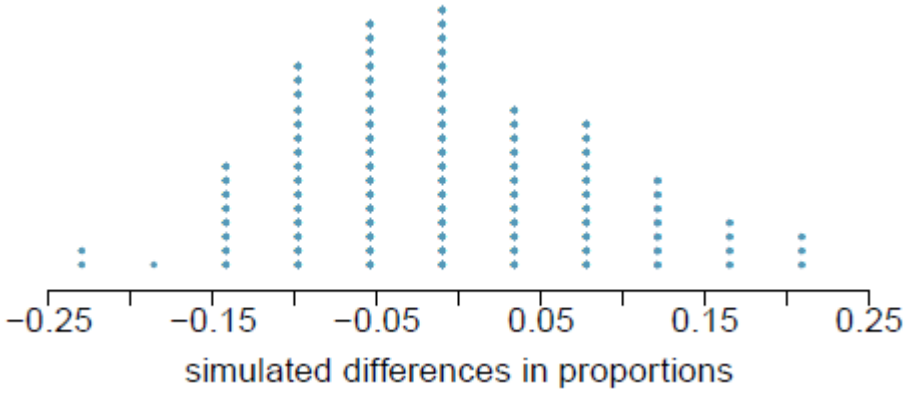
d. One approach for investigating whether or not th treatment is effective is to use a randomization technique.

i. What are the claims being tested? **Solution:** The claims being tested are the **H__o** that having a transplant has no effect on the survival rates at the end of the study and the **H__a** that having a transplant leads to greater survival rates at the end of the study.

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 **28 (since 28 is th total live people in the sample)** cards representing patients who were alive at the end of the study, and dead on **75(103-28)** cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size **69** representing treatment, and another group of size **34** representing control. We calculate the difference 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 **zero**. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are **greater than 24/69-4/34 = 0.23** . 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 effectiveness of the transplant program?



Solution: Per my estimations, the difference in proportions is 24/69 - 4/34 = 0.23, i.e. 23%. And large majority of the simulations produced proportions below this result, suggesting rejection of null hypothesis i.e. heart transplant really does have an effect on survival rates.