IS 606 - Chapter 1 Homework

Charley Ferrari

Wednesday, September 02, 2015

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.

Sex, Age, Marital, Gross Income (categorical), smoke (yn), amt Weekends, amt weekdays

sex: Categorical, not ordinal

Age: Numerical, discrete

Marital: Categorical, not ordinal

grossIncome: 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.

We'd have to know more about how the children were picked to know if the results can be generalized to the population. If the researchers only picked children within the vicinity of their school for example, there could be biases in the sample. The findings can't be used to establish causal relationships without more tests. This sort of study will only find correlations between qualities of the students and their cheating behavior.

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......

Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.

We cannot conclude for sure that smoking causes dementia later in life, only that smoking is associated with smoking. There may be confounding variables that we are not seeing. Because the study seems to be prospective however, it's a bit more conclusive than a retrospective study.

b. Another article titled "The School Bully is Sleeping" 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?

Once again, correlation is not the same as causation. This study does show that being a bully and having sleep disorders are associated. It's not clue which is causing which, or if there's a confounding variable influencing both.

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? This is a randomized experiment
b. What are the experimental and control treatments in this study? In this case, the subjects instructed to exercise are the experimental group, whereas the control group includes those subjects told not to exercise (although, I would think the better control would be to not give instruction on whether or not to exercise, to help avoid a situation where someone who exercises stops exercising.)
c. Does this study make use of blocking? If so, what is the blocking variable? Yes, blocking is used in this study. Age is the blocking variable.
d. Has blinding been used in this study? Blinding has not been used in this study. The patients are given clear instructions, and know whether or not they are exercising.
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 thte population at large.

One problem I mentioned above is the effect of whether or not patients already exercise. I would argue a better study design would be to give the experimental group instructions to exercise, and give the control group no instructions (allowing them to exercise if they would have). Alternatively, participants can be observed over time, taking measurements before being told to exercise and then after being told to exercise (in this case perhaps blinding can be used in the initial observation).

f. Suppose youare given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

My reservations were explained above in whether or not this study can be used to establish a causal relationship.

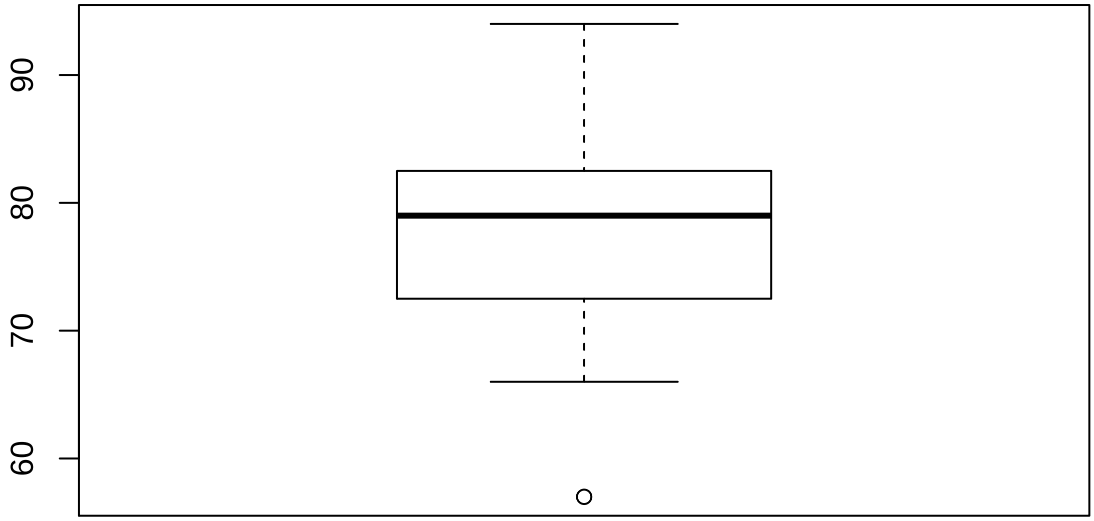
1.48: Stats scores

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.

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.

a. Bawed on the mosaic plot, is survival independent of whether or not the patient got a transplant?

According to the mosaic chart, those who received a transplant had greater odds of surviving. Because of this, it would suggest the two variables aren't independent.

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

According to the box plots, there was a relatively even distribution of increased survival times with the introduction of treatment. This gives more information than the mosaic plot, as we're able to see that the distribution was stretched outward. People seem to have a natural distribution of survival times, and given the treatment, every quartile survived for longer.

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

Using the mosaic's plot definition of "survived", (meaning survived to the end of the study), it would appear that roughly 1/8 of the control group survived, and roughly 1/3 of the treatment group survived.

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

e. What are the claims being tested? 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.

I found this wording very difficult to match without knowing the numbers of the studies. To reword this paragraph:

Out of the total population in this study, we have a proportion of participants who survived. We write *alive* on this number of cards, and *dead* on the rest. Then, we shuffle these cards and split them into two groups, one group **the size of the original treatment group** representing treatment, and another group **the size of the original control group** 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 his 100 times to build a distribution centered at **0**. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are **greater than than the observed difference in proportions**. 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?

Based off of my estimations in the mosaic chart, the difference in proportions is 1/3 - 1/8 = 0.2083. A large majority of the simulations produced proportions below this result, suggesting the we can reject the null hypothesis, and that heart transplant really does have an effect on survival rates.