Illapani_Homework_CH1

Srini Illapani

September 6, 2015

Chapter 1 - Introduction to Data

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. Note that "£" stands for British Pounds Sterling, "cig" stands for cigarettes, and "N/A" refers to a missing component of the data.

(a) What does each row of the data matrix represent?

- Each row of the data represents a single case which has multiple variables representing the demographic of smokers in UK and their smoking habits. The information is a sample size and has variables from gender, age, income to smoking habits.

(b) How many participants were included in the survey?

- 1961

(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 - Nominal, Age: Numerical- Discrete, Marital: Categorical - Nominal, Gross Income: Numerical - Continuous, Smoke: Categorical - Nominal, 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 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. Differences were observed in the cheating rates in the instruction and no instruction 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.

- Children between the ages of 5 and 15. Sample size is 160.

(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.

- While there is a corelation between the results and the two groups, the results do not indicate or establish a causal relationship between the findings and the sample population. The population is not representative of all children.

1.28

Reading the paper. Below are excerpts from two articles published in the NY Times:

(a) An article titled Risks: Smokers Found More Prone to Dementia states the following:61 "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-aday 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. - No, the study does not establish any causal relationship between smoking and dementia. The study was observational and based on the analysis of the data collected over years from volunteer participants.

(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?

- No, the conclusion by my friend is not accurate. The study was based on the survey results from parents and teachers, who were asked to idenntify on two aspects, sleep habits and behavior concerns of the students. These are two different variables and there is no other data to support that one variable is responsible for the other.

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?

- Effects of exercise on mental health

(b) What are the treatment and control groups in this study?

- The treatment group is the one with instructions to exercise and the control group is the one with instructions not to exercise.

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

- There is no mention if the study included both men and women. But the groups based on the age blocks are equaly represented in the control and treatment groups, hence the blocking.

(d) Does this study make use of blinding?

- No, there is no use of a placebo or something similar to keep the population in the blind.

(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.

- This is a random experiment performed on a sample population and is not representative of the total population as the subjects were from only 3 age groups. We do not know here about the findings but there could be a corelation between exercise and mental health, also there could be a causal relationship just with in the sample population. This cannot 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?

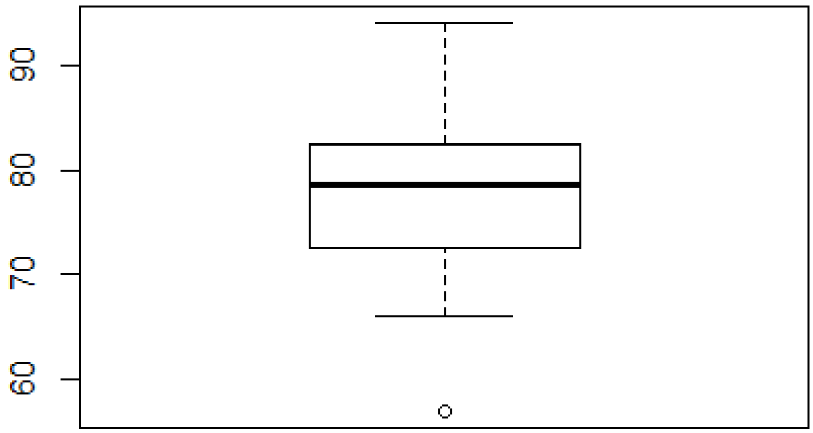
- Yes, I would have some reservations. What are the variables for measuring the mental health? How does one quantify them and do a pre and post experiment comparison?

1.48

(a) Below are the final exam scores of twenty introductory statistics students. 57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94 Create a box plot of the distribution of these scores.

- Here is the R code and the resulting box plot:

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



1.50

(a) Mix-and-match. Describe the distribution in the histograms below and match them to the box plots.

- a -> 2 (unimodal)
- b -> 3 (uniform)
- c -> 1 (bimodal)

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 distribution of housing prices is likely symmetric, as a meaningful number of houses are proced above the third quartile giving the symmetric curve rather than a steep one. Therefore the center would be best described by the mean, and variability would be best described by the standard deviation.

(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 distribution of housing prices is likely right skewed, as only few houses are at the top end price. Therefore the center would be best described by the median, and variability would be best described by the IQR.

(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 distribution of drinks consumed by college students is likely right skewed, as majority of them are assumed to be underage and hence conusme no or very little. Therefore the center would be best described by the median, and variability would be best described by the IQR.

(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 distribution of employees is likely right skewed, as only a few of them are making much higher salaries compared to the rest. Therefore the center would be best described by the median, and variability would be best described by the IQR.

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? Explain your reasoning.
   - Based on the masaic plot, Survival rate is higher in the treatment group compared to the control group. Hence survival is dependent on transplant.

(b) What do the box plots below suggest about the efficacy (effectiveness) of the heart transplant treatment.
   - The effectiveness of heart transplant in treatment group is clearly visible and is almost 3 times the control group rates of survival. It is symmetric for treatment group and skewed for the control group.

(c) What proportion of patients in the treatment group and what proportion of patients in the control group died?
   - Approximately 90% patients died in control group compared to 70% in the treatment group

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

i. What are the claims being tested?

   - H0: The heart transplant and death rates are independent. They have no relationship, and the difference in death rates between the groups is due to chance. HA: The heart transplant and death rates are not independent.The difference in the death rates between the groups is not due to chance and death is associated with heart transplants done or not done.

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 **index** cards representing patients who were alive at the end of the study, and dead on **index** cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of **equal** size representing treatment, and another group of size **equal** 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 **0**. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are **plotted**. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance andthat 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?

   - In the actual study, we observed about 70% death rate in the treatment group. In the 100 simulations under the independence model, we observed somewhat less than 25% in every single simulation, which suggests that the actual results did not come from the independence model. That is, the variables do not appear to be independent, and we reject the independence model in favor of the alternative. The actual study's results provide convincing evidence that heart transplant is associated with an decreased risk of deaths.