Introduction to Data

CUNY MSDA - IS606 - Homework 1

Completed by: Duubar Villalobos Jimenez mydvtech@gmail.com

February 5, 2017

Homework

OpenIntro Statistics

`Practice: 1.7 (available in R using the data(iris) command), 1.9, 1.23, 1.33, 1.55, 1.69`
`Graded: 1.8, 1.10, 1.28, 1.36, 1.48, 1.50, 1.56, 1.70`

For 1.48, the following R code will create a vector scores that can be used to answer the question:

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

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?

Answer: Each row represents a case study.

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

Answer: 1691 participants were 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.

Answer:

Sex: Categorical Variable.

Age: Numerical Continuous Variable.

Marital: Categorical Variable.

Gross Income: Numerical Discrete.

Smoke: Categorical Variable.

amtWeekends: Numerical Discrete.

amtWeekedays: 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.

Answer: The population of interest are Children between ages 5 and 15. The 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.

Answer: These results can not be generalized since we don't know if the sample was chosen randomly and also the assignment was not random; No casual conclusion can be made since the correlation statement is only valid for the sample study.

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.

Answer: We can NOT conclude that smoking causes dementia since we need to take a few extra factors:

- The sample was not Random but voluntary, so the results could be biased.
- We don't now if other factors could be contributing for example genetics, alcoholism or drugs.

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

62 "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?

Answer: In this case we can NOT make that conclusion because:

- The sample was not representative from the population but local.
- Are there some "things" happening at the time that is keeping the students awake at night?
- There are other factors not considered for example as to why are those students having sleep disorders? was it at only one point in time? noise level?

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 old 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?

Answer: Prospective Study (It identifies individuals and collects information as events unfold).

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

- Treatment Group: Patients that exercise twice a week.
- Control Group: Patients for whom advice was given as to not to exercise.

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

Yes, the blocking variable is the age.

(d) Does this study make use of blinding?

Yes (the experimental cases don't know whether they are in the control group or the treatment group).

(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.

Yes the results can be used to establish a casual relationship between exercise and mental health since the sampling is random and the assignments were random; in this case the results 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?

No, I would not have any reservations about the study proposal. I think is statistically well presented since it complies with the Principles of Experimental Design:

- Control
- Randomize
- Replicate
- Block

1.48 Stats scores.

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

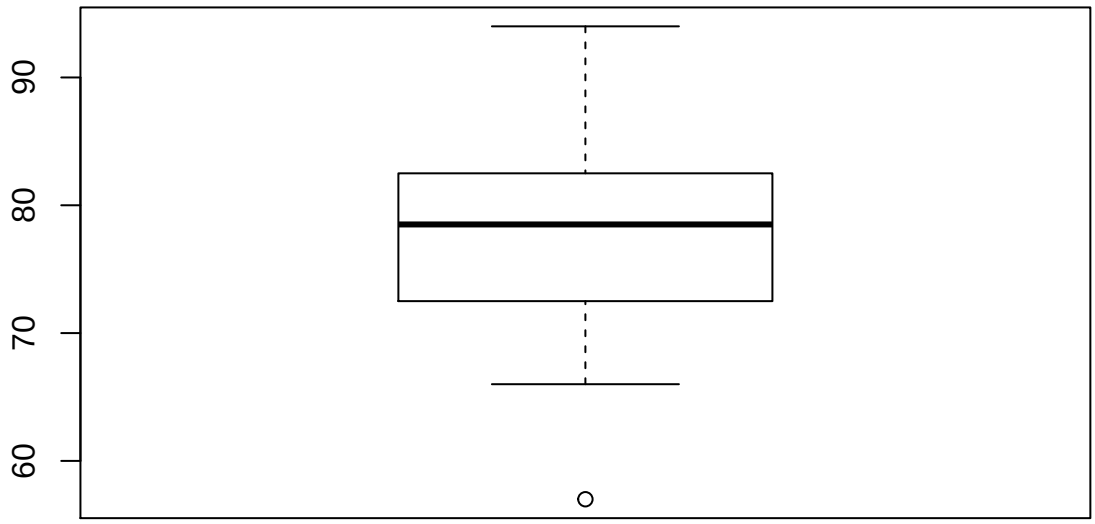
57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94

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

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

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

```
boxplot(scores)
```



1.50 Mix-and-match.

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

a) Symmetrical distribution; the match will be box plot #2.

b) Multimodal distribution; the match will be the box plot #3.

c) Right Skew distribution; the match will be the box plot #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.

Answer: This will be a left skewed distribution; in this case the Median will be a better choice and the IQR as well.

(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.

Answer: This will be represented with a symmetrical distribution; but since there's the risk of altering our numbers due to outliers, is better to take the Median and 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.

Answer: This will be a right skewed distribution; in this case the Median and IQR is better to employ.

(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.

Answer: This will be a symmetrical distribution; in this case the median salary will be best in order to control the outliers and also 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.

Answer: If we see the mosaic plot, we can conclude that the survival is not independent since the expectancy of life is bigger for the patients who got the heart transplant.

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

Answer: The box plot suggest that the heart transplant increases the survival rate for a longer period of time.

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

Answer:

From the article we can find as follows:

Control Group

alive = 4

dead = 30

total control = 34

Treatment Group

alive = 24

dead = 45

treatment total = 69

Control Group died proportion:

$$\frac{Control\ Dead}{Total\ Control} = \frac{30}{34}$$

Treatment Group died proportion:

$$\frac{Treatment\ Dead}{Total\ Treatment} = \frac{45}{69}$$

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

i. What are the claims being tested?

Answer:

H0: We start with a null hypothesis that represents the status quo.

HA: We also have an alternative hypothesis that represents our research question (Survival due to transplant).

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.

Answer:

We write alive on 28 cards representing patients who were alive at the end of the study, and dead on 75 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 $\frac{45}{69} - \frac{30}{34} = -0.230179$. 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 on the 100 simulations, we can conclude as follows:

(1) We conclude that the study results do provide strong evidence against the NULL hypothesis. That is, we do have sufficiently strong evidence to conclude the heart transplant was a success since the difference in between the 100 simulations is centered near zero.

(2) We conclude that the evidence is sufficiently strong to reject **H0** and assert that there was a success survival rate due to heart transplant. When we conduct formal studies, usually we reject the notion that we just happened to observe a rare event 0.50 So in this case, we reject the independence model in favor of the alternative. That is, we are concluding the data provide strong evidence of survival due to hearth transplant.