Grando-1 Homework

John Grando

January 31, 2017

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 which can also referred to as a unit of observation or observational unit.

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

Answer: 1,691 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, nominal

age - numerical, discrete

marital - categorical, nominal

grossIncome - categorical, ordinal

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.

Answer: The population of interest is all children and the sample is 160 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.

Answer: Since this is an experiment, and not an observational study, then causal relationships could possibly be determined. However, this experiment deals directly with cheating (e.g. honesty) and it does not appear that the study has a way to separate the factor of self-control. Therefore, it is unclear if a students reporting incorrect numbers are dishonest or honest with a low level of self-control due to the reward. Possibly, if the reward were varied then data could be collected determining how often a child incorrectly reported results between reward types, which would show how many children would lie for something they wanted rather than just in general. Additionally, it is unclear whether 160 children is a large enough sample size or if there was a selection bias (e.g. all children from the same school). Additionally, these results could only be generalized to children between the ages of 5 and 15.

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: "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.

Answer: No, this is an observational study. In order to determine causal connection, an experiment must be performed. Additionally, this was a voluntary survey which means that there could be a bias in the population of respondents.

(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?

Answer: No, this is not justified. I would say that the study shows there is an association between sleep disorders and bullying but it does not prove one is caused by the other. Also, there may be some other variable (confounding variable) which may be the cause of both issues which could not be determined from this data.

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?

Answer: This is a blocked randomized experiment.

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

Answer: The treatment group is the participants who receive instructions to exercise. The control group is the pariticipants who are told not to exercise.

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

Answer: Yes, age is the blocking variable.

(d) Does this study make use of blinding?

Answer: No, the control group and administrators of the survey know who they are and no placebo is used.

(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.

Answer: The study does state that it randomized and blocked for age to determine causal relationships. It appears to be sufficient in design; however, since the control group was instructed to not exercise, rather than randomly selecting people and choosing those who do not already exercise, it is unclear whether the respondents will record poor mental health scores due to the fact that they are now not doing what they normally do. This in fact, this would give me concern on how well of a reference point the control group is.

(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?

Answer: Yes, my reservations would be that the control group has not been properly instructed which means they may not be a sufficient reference point.

1.48 Stats scores.

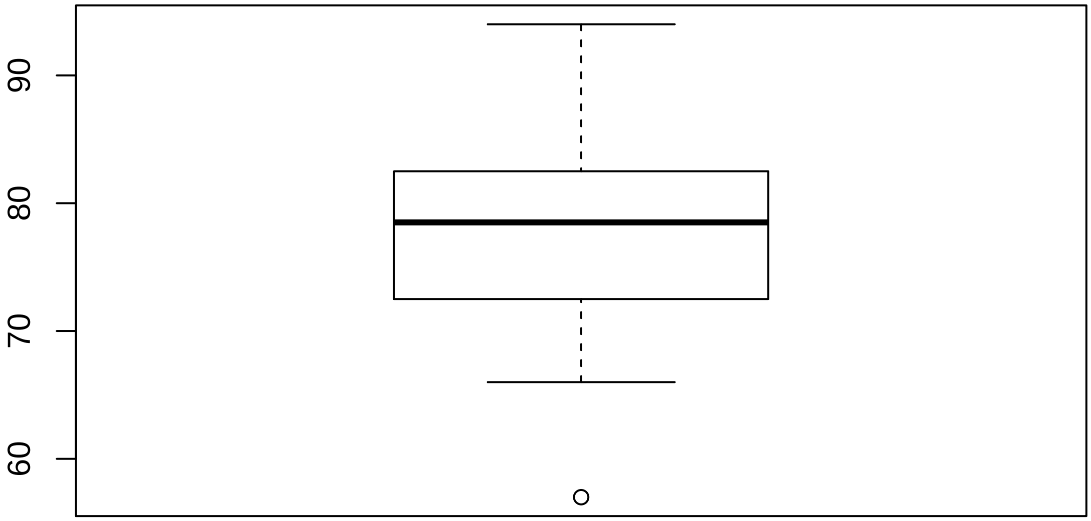
Below are the final exam scores of twenty introductory statistics students.

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.

Answer:

```
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 Mix-and-match.

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

Answer:

(a) The box plot that matches this histogram is (2). The distribution of this data is unimodal, symmetric, and appears it could be normally distributed.

(b) The box plot that matches this histogram is (3). The distribution of this data might be symmetric but it does not look normally distributed. It is possible it is uniform but a smaller bin width would have to be selected to see.

(c) The box plot that matches this histogram is (1). The distribution of this data is unimodal and has a right skew.

1.56 Distributions and appropriate statistics, Part II.

For each of the following, statewhether 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 distribution would be right skewed. The median would be the best to represent a typical observation and the IQR would best represent the variability since there are a number of very expensive houses compared to most of the data set which would shift the mean and standard deviation signfiicantly more.

(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 distribution would be symmetric since the few expensive houses wouldn't skew the data much, and they are not extraordinarly more expensive than the main data set. The mean and standard deviation would be acceptable to predicte the typical observation and variability since the most expensive houses are not large in number and not extraordinarily more expensive. It appears mediand and IQR would produce similar results.

(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 distribution is right skewed. Since there isn't a significant number of excessive drinkers, median and standard deviation would be good selections. Again, median and IQR would be acceptable in this example as well.

(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: As with the housing example (a), there appear to be a few highly outling salaries so the median and IQR should be used.

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: No, it appears survival is dependent on whether the patient recieved a transplant. The control group's "dead" area appears to be significantly taller than the treatment group, the opposite is true for the survival areas. It is noted that many more people appear to be in treatment group than the control group, which may cause some issues with determining whether the treatment has a significant affect on survival rate.

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

Answer: The heart transplant treatment significantly increases the survival time (days) for patients. it appears all measurements of the data (e.g. Q1, median, etc.) are significantly better for the treatment group than the control group.

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

Answer: From the class github site, I found that there are exercise data sets. I was able to download this one by using the library and data functions

```
setwd("~/Documents/Masters/DATA606/Week1/Homework")
library(openintro)
```

```
## Please visit openintro.org for free statistics materials
```

```
##
## Attaching package: 'openintro'
```

```
## The following object is masked from 'package:datasets':
##
##     cars
```

```
data("heartTr")
```

I then calculated the number of patients in each control group and which ones died.

```
patient_control_dead <- nrow(subset(heartTr, heartTr$transplant ==
    "control" & heartTr$survived == "dead"))
patient_control <- nrow(subset(heartTr, heartTr$transplant ==
    "control"))
patient_treatement_dead <- nrow(subset(heartTr, heartTr$transplant ==
    "treatment" & heartTr$survived == "dead"))
patient_treatment <- nrow(subset(heartTr, heartTr$transplant ==
    "treatment"))
patient_control_dead_ratio <- patient_control_dead/patient_control
patient_treatment_dead_ratio <- patient_treatement_dead/patient_treatment
patient_control_dead_ratio
```

```
## [1] 0.8823529
```

```
patient_treatment_dead_ratio
```

```
## [1] 0.6521739
```

88.24% of the patients in the control group died by the end of th e study and 65.22% of the patients in the treatment group died by the end of the study.

(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: The claim being test is whether or not a heart transplant will increase a patient's lifespan.

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:

I will use the countss from the heartTr data

```
patient_alive <- sum(heartTr$survived == "alive")
patient_alive
```

```
## [1] 28
```

```
patient_dead <- sum(heartTr$survived == "dead")
patient_dead
```

```
## [1] 75
```

```
patient_treatment
```

```
## [1] 69
```

```
patient_control
```

```
## [1] 34
```

```
patient_treatment_dead_ratio - patient_control_dead_ratio
```

```
## [1] -0.230179
```

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 [approximately zero]. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are [more extreme than our determined result (-23.02%)]. 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?

Answer:

It appears that a difference of at least -23.02% due to chance alone would only happen about 2% of the time according to the figure. Such a low probability indicates a rare event.