HW1

Bin Lin

2016-9-4

1.8 Smoking habits of UK residents.

a. What does each row of the data matrix represent? Each row represents the observation of an individual UK resident's smoking habits.

b. How many participants were included in the survey? 1691

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 Age: Numerical, discrete Marital: Categorical Gross income: Categorical, ordinal Smoke: Categorical antWeekends: numerical, discrete amtWeekdays: numerical, discrete

1.10 Cheaters, scope of inference.

a. Identify the population of interest and the sample in this study. Population of interest: All people Sample: 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.

I don't think the result of the study can be generalized to the population, because the experiment only included children whose age is between 5 and 15. It does not include any subjects that are in different age groups. Let us say the adults maybe are less prone to the award, therefore, they are more honest. Or for instance, the adults are maybe less afraid of getting punished by parents, therefore, they are more likely to cheat to gain more personal interest. that is why I think the result can not be generalized. The study can be used to establish causal relationship, because it is a experiemental study. Researchers assign children ramdonly into two different groups. They were trying to establish causal connect among honesty, age, and sel-control.

1.28 Reading the paper.

a. Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning. We can not conclude that smoking causes dementia later in life. Because the study is a observational study.Researchers just collect data, but they do not interfere with how the data arises by imposing treatments. They can only establish association between smoking and dementia.

b. 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?

This statement is not justified. Reason is same as part (a). The conclusion should be there is an association between sleep disorders and disruptive behaviors.

1.36 Exercise and mental health.

a. What type of study is this? Experimental study

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

Treatment groups:The group that exercise twice a week Control groups: The group that do not exercise

c. Does this study make use of blocking? If so, what is the blocking variable? No.

d. Does this study make use of blinding? No

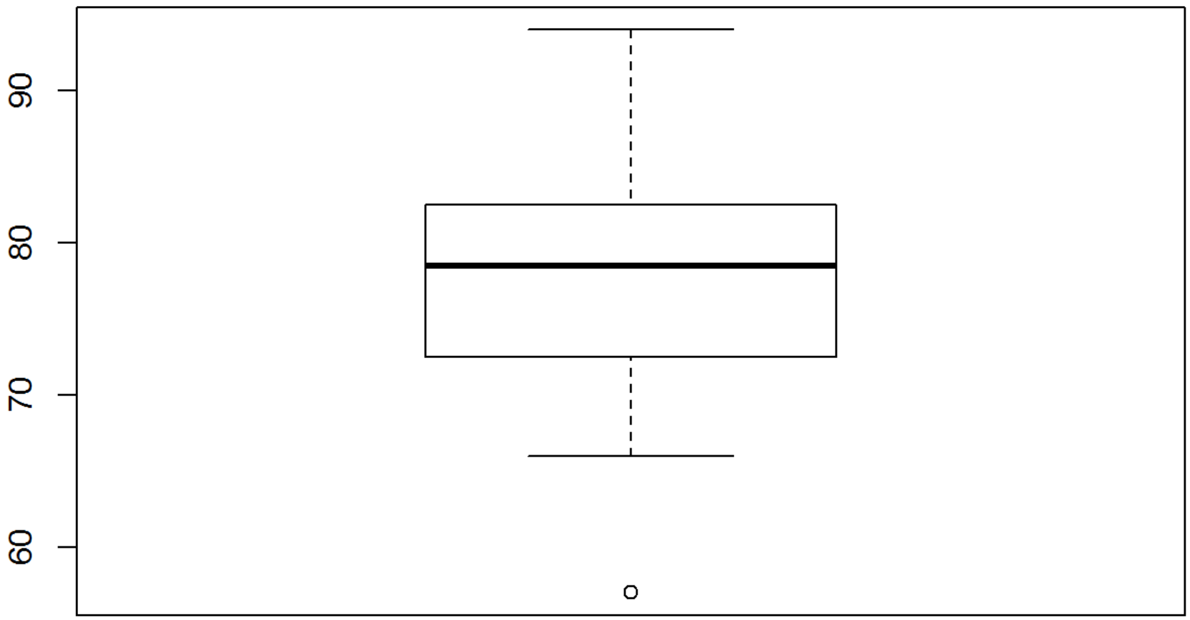
e. Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.

This is an experimental study, so this study can be used to establish a causal relationship between exercise and mental health. The conclusion can be generalized to population in large, because the experimental units that were recruited were selected using stratified random sampling.

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? The intervention of this study is to exercise twice a week for the experimental group. However, it does not specify how much and how long that group should exercise. That is something I qill qeustion about it before I provide the funding.

1.48 Stats scores.

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



1.50 Mix-and-match.

(a)–(2) (b)–(3) (c)–(1)

1.56 Distributions and appropriate statistics, Part II .

(a)The distribution is right skewed.Because there is a natural boundary on the housing price at 0 and only a few houses that cost millions. The median would best represent the typical observation in the data, and the variability of observations would be best represented using IQR.

(b)The distribution is very closed to symmetric. because if we attempt to draw a boxplot, the center is at $600,000, the 25% is symmetric to 75%, 0% is also symmetric to 100% with very few incidents of extreme values which represent home prices that is greater than $1.2 million.The mean would best represent the typical observation, and the standard deviation can best represent the variability of observations.

(c)The distribution will be right skewed. Because the natural boundary at 0 and very few student drink excessively as given by the question. Therefore, median can best represent the typical observation in the data, and IQR can best represent the variability of observations.

(d)The distribution will be right skewed. Because the natural boundary at 0 and very few employees earn very high salaries given by the question. Therefore, median can best represent the typical observation in the data, and IQR can best represent the variability of observations.

1.70 Heart transplants.

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

The survival is dependent on whether or not the patient got a transplant. Because according to the mosaic plot, the patient who received treatment have significant higher survival rate.

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

The boxplot suggests the treatment group has much higher survival time in general compared to control group, so it proves that the heart transplant treatment is very efficacious.

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

88.23% of patient in the control group died, while only 65.22% patient in the treatment group died.

```
getwd()
```

```
## [1] "C:/Users/blin261/Documents/R/DATA606"
```

```
setwd('C:/Users/blin261/Documents/Lab1/more')

raw_data<-read.csv("heartTr.csv",sep=",")
head(raw_data)
```

```
##   id acceptyear age survived survtime prior transplant wait
## 1 15         68  53     dead        1    no    control   NA
## 2 43         70  43     dead        2    no    control   NA
## 3 61         71  52     dead        2    no    control   NA
## 4 75         72  52     dead        2    no    control   NA
## 5  6         68  54     dead        3    no    control   NA
## 6 42         70  36     dead        3    no    control   NA
```

```
str(raw_data)
```

```
## 'data.frame':    103 obs. of  8 variables:
##  $ id        : int  15 43 61 75 6 42 54 38 85 2 ...
##  $ acceptyear: int  68 70 71 72 68 70 71 70 73 68 ...
##  $ age       : int  53 43 52 52 54 36 47 41 47 51 ...
##  $ survived  : Factor w/ 2 levels "alive","dead": 2 2 2 2 2 2 2 2 2 2 ...
##  $ survtime  : int  1 2 2 2 3 3 3 5 5 6 ...
##  $ prior     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ transplant: Factor w/ 2 levels "control","treatment": 1 1 1 1 1 1 1 2 1 1 ...
##  $ wait      : int  NA NA NA NA NA NA NA 5 NA NA ...
```

```
table(raw_data$survived, raw_data$transplant)
```

```
## 
##        control treatment
##   alive       4        24
##   dead       30        45
```

```
#Proportion of patient in control group died
30/(30+4)
```

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

```
#Proportion of patient in treatment group died
45/(24+45)
```

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

d.

i. What are the claims being tested? H0: Heart transplant treatment is not effective H1: Heart transplant treatment is effective

ii: 28 alive card 75 dead card 52 treatment size 51 control size distributin centered at 0 Difference are 23.01%

```
dim(raw_data)
```

```
## [1] 103   8
```

```
#Number of alive card
4+24
```

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

```
#Numer of dead card
30+45
```

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

```
#Difference of death rate between treatment group and control group
0.8823-0.6522
```

```
## [1] 0.2301
```

```
#According to the graph, based on the actual study, the proportion of patient who died in treatment is about 23% less than t
hat of control group. Proportion of 23% fall on the right side of the bell curve generated from the simulation result, with
 mass majority of the data fall under the 23%. This suggest the heart tranplant program is really effective. 23% reduction i
n terms of death rate is not due to chance, it has something to do with the heart transplant program. Therefore, we reject t
he null hypothesis, so that we can conclude the heart transplant treatment is effecttve.
```