606-01 - Homework 1_Introduction to data

Shyam BV

September 4, 2016

1. Excercise 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.

```
smokinghabits <- read.csv("https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20
```

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

```
head(smokinghabits)
```

```
##   gender age maritalStatus highestQualification nationality ethnicity
## 1   Male  38      Divorced     No Qualification     British     White
## 2 Female  42        Single     No Qualification     British     White
## 3   Male  40       Married               Degree     English     White
## 4 Female  40       Married               Degree     English     White
## 5 Female  39       Married         GCSE/O Level     British     White
## 6 Female  37       Married         GCSE/O Level     British     White
##        grossIncome    region smoke amtWeekends amtWeekdays    type
## 1   2,600 to 5,200 The North    No          NA          NA
## 2      Under 2,600 The North   Yes          12          12 Packets
## 3 28,600 to 36,400 The North    No          NA          NA
## 4 10,400 to 15,600 The North    No          NA          NA
## 5   2,600 to 5,200 The North    No          NA          NA
## 6 15,600 to 20,800 The North    No          NA          NA
```

Each row represents a single observation. In this dataset, it means the data of single UK resident. It shows his detail and the amount which he smokes.

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

```
nrow(smokinghabits)
```

```
## [1] 1691
```

Totally 1691 participants were included in this 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.

```
str(smokinghabits)
```

```
## 'data.frame':    1691 obs. of  12 variables:
##  $ gender              : Factor w/ 2 levels "Female","Male": 2 1 2 1 1 1 2 2 2 1 ...
##  $ age                 : int  38 42 40 40 39 37 53 44 40 41 ...
##  $ maritalStatus       : Factor w/ 5 levels "Divorced","Married",..: 1 4 2 2 2 2 2 4 4 2 ...
##  $ highestQualification: Factor w/ 8 levels "A Levels","Degree",..: 6 6 2 2 4 4 2 2 3 6 ...
##  $ nationality         : Factor w/ 8 levels "British","English",..: 1 1 2 2 1 1 1 2 2 2 ...
##  $ ethnicity           : Factor w/ 7 levels "Asian","Black",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ grossIncome         : Factor w/ 10 levels "10,400 to 15,600",..: 3 9 5 1 3 2 7 1 3 6 ...
##  $ region              : Factor w/ 7 levels "London","Midlands & East Anglia",..: 6 6 6 6 6 6 6 6 6 6
##  $ smoke               : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 2 1 2 2 ...
##  $ amtWeekends         : int  NA 12 NA NA NA NA 6 NA 8 15 ...
##  $ amtWeekdays         : int  NA 12 NA NA NA NA 6 NA 8 12 ...
##  $ type                : Factor w/ 5 levels "","Both/Mainly Hand-Rolled",..: 1 5 1 1 1 1 5 1 4 5 ...
```

gender - Categorical age - numerical -> Continuous maritalStatus - Categorical highestQualification - Categorical nationality - Categorical ethnicity - Categorical grossIncome - Categorical -> Ordinal region - Categorical smoke - Categorical amtWeekends - numerical -> discrete amtWeekdays - numerical -> discrete type - Categorical

2. Excercise 1.10

(a) Identify the population of interest and the sample in this study.

Answer: In this research or study, the population of interest is all the childrens between the ages of 5 to 15.

The sample used in this study is the 160 children between the ages of 5 to 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.

Answers may way. The results of this study cannot be generalized to the population. Because we did not know the region of this 160 children. Also compared to the population, the sample size is small. So we need more sample size to make generalize the results to whole population.

Yes, this is an experiment. And the findings from this study can be used to establish causal relationships.

Excercise 1.28

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

This is an observational study. There was no treatment and control group. We cannot derive causal relationship on observational study. Although the numbers show that the smoking causes dementia, we are not sure about the external factors(cofounding variable) which is involved. So we cannot conclude that smoking causes dementia.

Sample popuation: 23123 25% - Dementia (includes 1136 - Alzheimer, 416 - Vascular dementia) Total persons had Dementia: 5781

(b) Another article titled The School Bully Is Sleepy states the following:62

This is an observational study. The statement is not a valid statement. Because the study shows that the children who had behavior issues are likely to show sleep disorders. But it is not other way around. We cannot state that the sleep disorders lead to bullying in school children.

So we can't make causal relationship on observational study.

Excrecise 1.36 :

(a) What type of study is this?

This is an experimental study.

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

Treatment group contains random half the subjects from all the ages (18-30, 31-40 and 41- 55 year).

Control group contains random half the subjects from all the ages (18-30, 31-40 and 41- 55 year).

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

Yes. This study makes use of blocking. The blocking variable is age.

(d) Does this study make use of blinding?

No. It does not explicitly mention that the study is using blinding.

(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. It can use used to find out the causal relationship. It depends on the sample size. The study can be generalized to the population if the sample size is large enough.

(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?

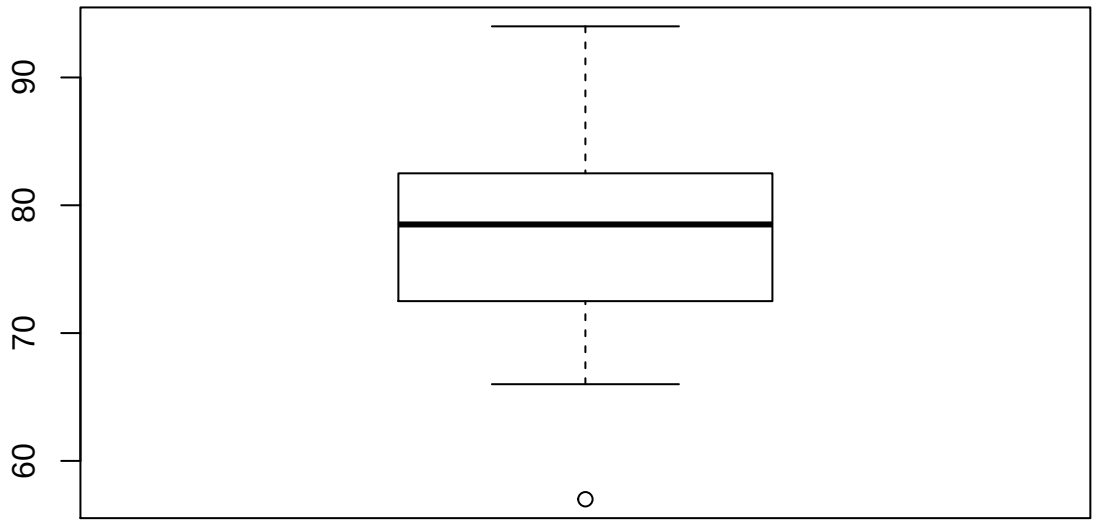
I would recommend to fund the study. I would recommend a good number of sample size for this study.

Excercise 1.48

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

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

boxplot(statscores)
```



Excercise 1.50

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

(a) The match box plot number is 2. The distribution is unimodel which has one peak. And the histogram is symmetric.

(b) The match box plot number is 3. The distribution is multimodel which has many peak. And the histogram is symmetric.

(c) The match box plot number is 1. The distribution is unimodel which has one peak. And the histogram is right skewed.

Excercise 1.56

(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.

(1) Right skewed
(2) Median - Because the study is right skewed.
(3) Standard Deviation - Because we need to best represent the variablity.

(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.

(1) Symmentrical distribution

(2) Mean - Because the study is symmentrical
(3) IQR - Because all the data can be showed in a single chart with variability.

(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.

(1) Symmentrical distribution
(2) Median - Because the study is multimodel
(3) Standard Deviation

(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.

(1) Right skewed
(2) Median - Because the study is right skewed.
(3) Standard Deviation - Because we need to best represent the variablity.

Excercise 1.70

```
heattrans <- read.csv("https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20from%
head(heattrans)
```

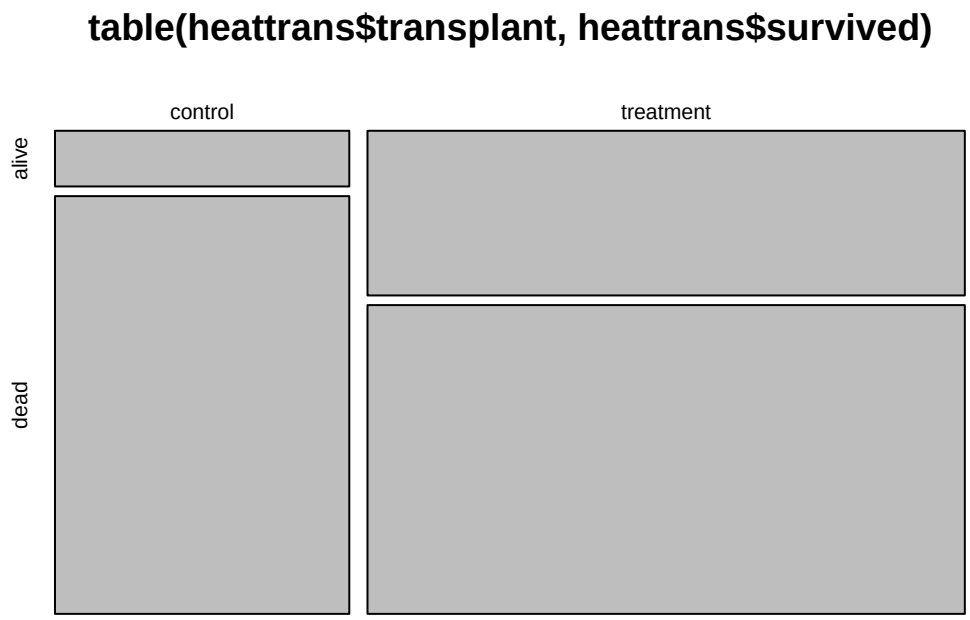
```
##   id acceptyear age survived survtime prior transplant wait
## 1 15         68  53     dead        1    no    control   NA
## 2 43         70  43     dead        2    no    control   NA
## 3 61         71  52     dead        2    no    control   NA
## 4 75         72  52     dead        2    no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

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

```
summary(heattrans)
```

```
##        id          acceptyear         age         survived
##  Min.   :  1.0   Min.   :67.00   Min.   : 8.00   alive:28
##  1st Qu.: 26.5   1st Qu.:69.00   1st Qu.:41.00   dead :75
##  Median : 49.0   Median :71.00   Median :47.00
##  Mean   : 51.4   Mean   :70.62   Mean   :44.64
##  3rd Qu.: 77.5   3rd Qu.:72.00   3rd Qu.:52.00
##  Max.   :103.0   Max.   :74.00   Max.   :64.00
##
##     survtime      prior        transplant      wait
##  Min.   :   1.0   no :91   control  :34   Min.   :  1.00
##  1st Qu.:  33.5   yes:12   treatment:69   1st Qu.: 10.00
##  Median :  90.0                          Median : 26.00
##  Mean   : 310.2                          Mean   : 38.42
##  3rd Qu.: 412.0                          3rd Qu.: 46.00
##  Max.   :1799.0                          Max.   :310.00
##                                          NA's   :34
```

```
mosaicplot(table(heattrans$transplant,heattrans$survived))
```



From the mosaic plot, the survival of the patient is dependet on transplant.

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

From the box plots, the efficacy of heart transplant treatment is very bad. Most(30) of the patients were dead. There were some outlies in this study. So some of them were alive might be due to the chance or error in experiment.

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

```
control_died <- (30/75)
control_died
```

```
## [1] 0.4
```

```
treatment_died <- (45/75)
treatment_died
```

```
## [1] 0.6
```

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

i. What are the claims being tested?

Whether the trasplant is successful or not.

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.

1. Random
2. Random
3. half
4. half
5. Mean or Zero
6. Low

iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?

From the simulation results, it shows that the effectiveness of happening is 3 times.