homework_ch1_JimmyNg

Jimmy Ng

September 4, 2018

Question 1.8

(a) Each row represents a single participant. This is like a typical data frame. Each row represents a user/participant, and each column represents a variable.
(b) There are 1691 participants.
(c) First variable "sex" is categorical and nominal (in other words, not ordinal); second variable "age" is numerical and discrete (in this case, someone can be 42 or 43 but cannot be 42.5 or 42.75); third variable "marital" is categorical and nomial (or not ordinal); fourth variable "grossIncome" is categorical and ordinal (as we can compare higher to lower income); fifth variable "smoke" is categorical and nominal; sixth and seventh variables "amtWeekends" and "amtWeekdays" are numerical and discrete, i.e. someone can smoke on average 5 cigarettes per day but not 5.125 cigarettes per day.

Question 1.10

(a) The population of interest would be children between the ages of 5 and 15. The sample is composed by 160 children within this age range.
(b) There are many confounding variables that need to be controlled, such as IQ, social-demographic status (such as hhi - household income, parental education), academic performance, etc. It's unlikely that this study can be generalized to the population, especially when we consider cross-cultural significance of cheating or lying. In addition, the finding cannot be used to establish a causal relationship - there's no direcitonal indication or explicit measure of the magnitude about the "instruction not to cheat" would "cause" less cheating, or "the absence of instruction" would result in more cheating. We merely observe group differences (such as gender and/or age) - we can speculate about how these groups performed statistically significantly different, but we cannot draw a causal relationship from the finding. This study would need to be replicated in different control settings in order to demonstrate its validity and reproducibility.

Question 1.28

(a) This is an observational study. We cannot choose or decide who smoke or level of smoking (such as pack-a-day or two packs a day). We draw conclusion by simply running a correlation between levels of smoking and observation of dementia. We do not choose our participants and we do not intervene their lifestyle. As a result, we cannot draw a uni-directional conclusion by stating that smoking causes dementia. In fact, not all smokers have dementia; not all dementia patients smoke.
(b) This is an observational study. We neither choose our participants (decide who is a bully or not) nor make any lifestyle intervention. Being identified as a bully is independent of sleep disruption. In order to understand the relationship, we could have made this an experimental study by random sampling children and controlling their levels of sleep (such as 2 hours, 4 hours, 8 hours a day and so on) and subsequently observe their physical and psychological changes before and after the intervention (although this study will unlikely get approval from any Human Subject Board review for ethical reason). The study described in this paragraph is only an observational/correlation study. We can easily turn this around by saying, "bullying or demonstration of any physical, emotional or verbal abuses to others would cause psychological distress and subsequently the feeling of guilt would be unhinged and surfaced in a child's consciousness during sleep and therefore result in sleeping disorder" and this statement would still make sense (and can neither be proved nor disproved).

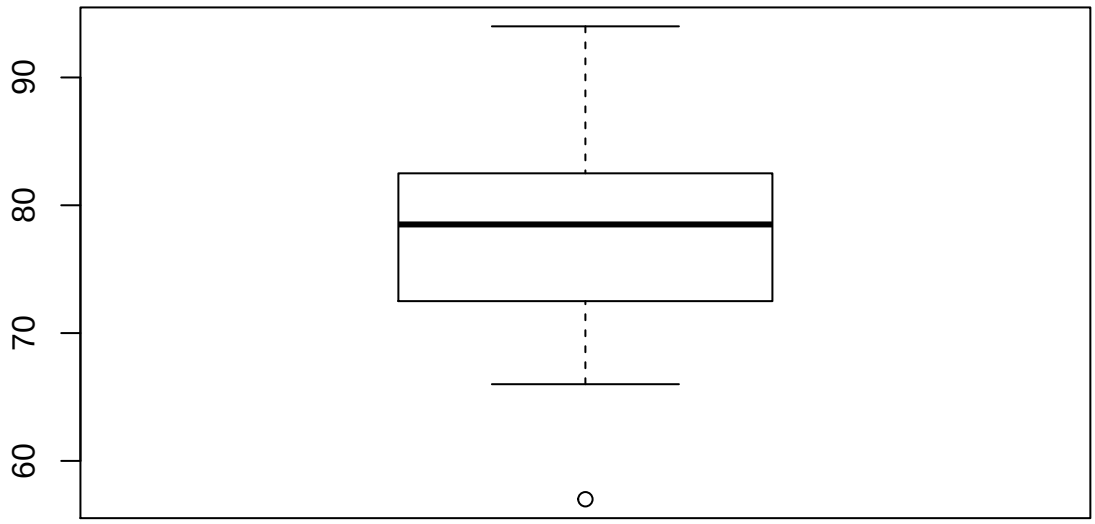
Question 1.36

(a) This is an experimental study using stratified random sampling method.
(b) The treatment group is the group which exercises, whereas the control group is the one who does not exercise.
(c) Blocking by age. This is a 2 x 3 between-subject experiment and we can run a 2 x 3 between-subject ANOVA to interpret group mean differenes. We have six separate groups (or blocks), i.e. 2 (treatment vs. control) x 3 (age 18-30, 31-40 and 41-55).
(d) No, the study does not use blinding. Subjects and experimenters are aware of the subjects' condition as the subjects know whether they have exercised or not.
(e) This is an experimental study and the result can be generated to the population if the study can have boarder control over other variables. Besides age, the study should also control other factors, such as subjects' weight, diet, physical fitness (perhaps someone already exercised ten times per week already), daily exericse level (perhaps someone's job requires intensive physical labor), mental health (such as people who are depressed should be excluded from the study), etc.
(f) I would have my reservation. It's because this type of study most likely have been done before, and it's difficult to see any value or conclusion drawn from this study that can be added to the scientific community. The study needs better control (such as variables listed above). A better way to secure funding is to come up with new set of physical exericse design that we can test and demonstrate any statistially significant improvement on mental health. We can partner with different departments (such as psychology, physical rehabilitation) in order to secure funding and promote copyright patents, so that we can (put it on the research grant proposal) say that the study has a monetary potential for future productizing, e.g. IQ test or those physical/rehabilitation assessment test.

Question 1.48

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

boxplot(x)
```



Question 1.50

(a) It's a unimodal symmetric distribution and it's matching the box plot(2).
(b) It's a uniform distribution and it's matching the box plot(3).
(c) It's a right-skewed distribution and it's matching the box plot(1).

Question 1.56

(a) Housing prices represent a right skewed distribution. Only a small number of prices skewed to the right and drives the mean to a much higher number than the median. Therefore, the median would better represent the data; otherwise, the mean would be skewed up and misrepresent the housing market in general. The variability of observations would be best represented using IQR as it clearly presents the 25%-, 50%- and 75%-quartile of the data as described in the paragraph. If plotting in boxplot, we can clearly display outliners as well. That would give us a better picture of the housing market.
(b) Housing prices represent a symmetric distribution. The mean would be a better option to represent the data. In this case, standard deviation would be better option to capture the data variability, e.g. 2 SD plus or minus the mean would have captured 95% of the data in this symmetric distribution.
(c) Right skewed distribution for alcoholic drinks. Only a small number of people drink excessively and these people (as outliners) will skew the distribution to the right. Median would better present the picture; otherwise, people would have thought that on average college students drink more than they are (as the mean is inflated by the outliners and it's higher than the median). In fact, majority of students are under 21 and presumably don't drink. The IQR would better present the data as it will focus on the 25%, 50% (which is the median) and 75% of the distribution. It will presumably show us that majority of these students consume zero or very little alcoholic drinks in a given week.

(d) Annual salaries also presents a right skewed distribution. Median would be better option than mean to present the data (for similar reason described above). However, once we remove few outliners on the far right, the distribution would look more symmetrical and the mean and standard deviation would better capture the data variation. Right now, the IQR would better capture the variability as it shows the 25%-, 50%- and 75%-quartile of the data. These quartiles can display a picture of how much people make in general from this company.

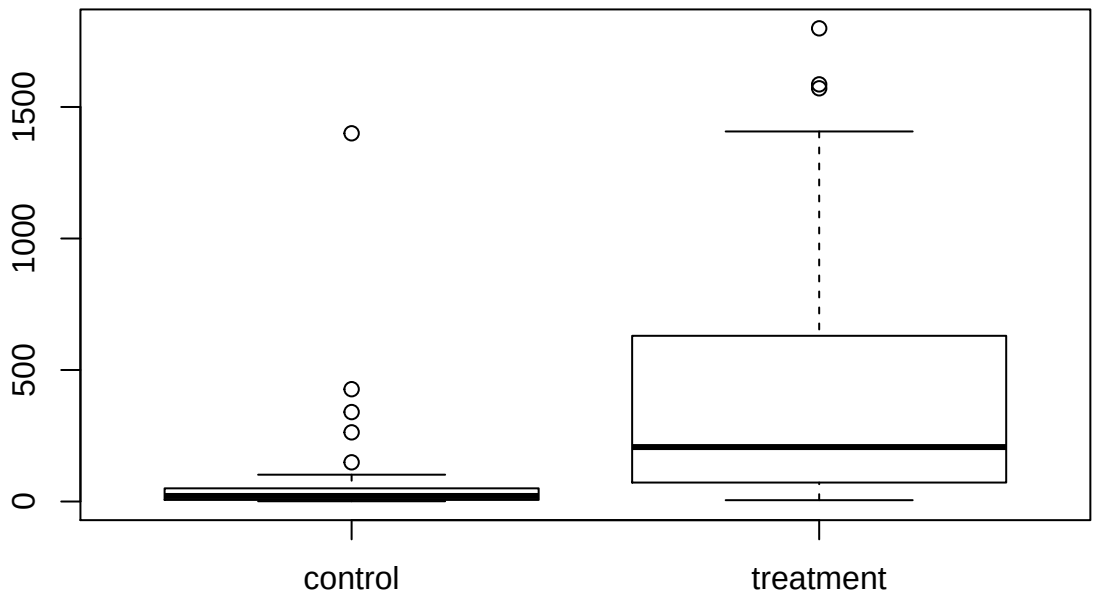
Question 1.70

```
data("heartTr", package = "openintro")
attach(heartTr)
round( prop.table(ftable(survived, transplant),
                  margin = 2),
      2 )
```

```
##          transplant control treatment
## survived
## alive                  0.12      0.35
## dead                   0.88      0.65
```

(a) The mosaic plot presents the sample size and relative frequency of above table. Only 12% of patients from control group survived, in contrast to 35% from treatment group. In other words, 88% of control patients were dead, in comparing to 65% from the treatment group. A much larger number of patients survived from the treatment group. Survival is not independent of heart transplant; getting a heart transplant is very critical of survival.

```
boxplot(survtime ~ transplant, data = heartTr)
```



(b) The boxplot displays the days of survival between control and treatment groups. First of all, the median of survival days for control is close to zero, whereas the median from treatment group is significantly higher! Second, there are only 5 outliners from control group that skew the data to the right. Without these 5 outliners, almost everyone survived in the control group do not last more than 100 days. Third, even the longest survivor from the control group is still within (or close to) the upper whisker in the treatmen group. Overall, treatment group patients have undenibly much longer survival time.

(c) From above table, we see that 88% control patients died, in contrast to 65% of treatment group patients.

(d.i) The claims being tested include treatment group has a statistically, significantly better survival rate, and longer survival time. Primarily, we are testing survival rate here by running a simulation. (d.ii) "survived", "dead", 50, 50, "resembling a normal distribution", "negative - death rate (treatment - control)" (d.iii) The death rate (treatment - control) in this study is 65% - 88% = -23%. According to this simulation result chart, this is an extremely rare event and therefore we can reject the null hypothesis and favor the alternative model, i.e. treatment demonstrated significantly different result (less death) than the control group.