DATA 606 Homework 1

Ilya Kats

February 3, 2017

1.8 Smoking habits of UK residents.

a. Each row of the data matrix represents a case which in this study is information about one participants.
b. The study included 1,691 participants.
c. Variables tracked in the study are as follows:
   - Sex - categorical
   - Age - numerical, discrete
   - Marital status - categorical
   - Gross income - categorical, ordinal
   - Smoker - categorical
   - Cigarettes smoked during weekends - numerical, discrete
   - Cigarettes smoked during weekdays - numerical, discrete

1.10 Cheaters, scope of inference

a. The population is all children between the ages of 5 and 15. The sample is 160 children selected for the study.
b. Provided the sample is representative of the population the results can be generalized to the population (but sample selection is critical here). This study is an experiment because participants were assigned to two groups - with exlicit instructions and without - so it can be used to establish causal relationship between some variables.

1.28 Reading the paper.

a. The study is an observational study and not an experiment; therefore, it cannot be used to establish causal relationship. It can only demonstrate dependency of variables.
b. Similarly to above, this study is not an experiment, so the statement that "sleep disorders lead to bullying in school children" is not justified. I believe the more accurate statement would be "The study shows that there is a link between disruptive behavior or bullying and sleep disorders."

1.36 Exercise and mental health

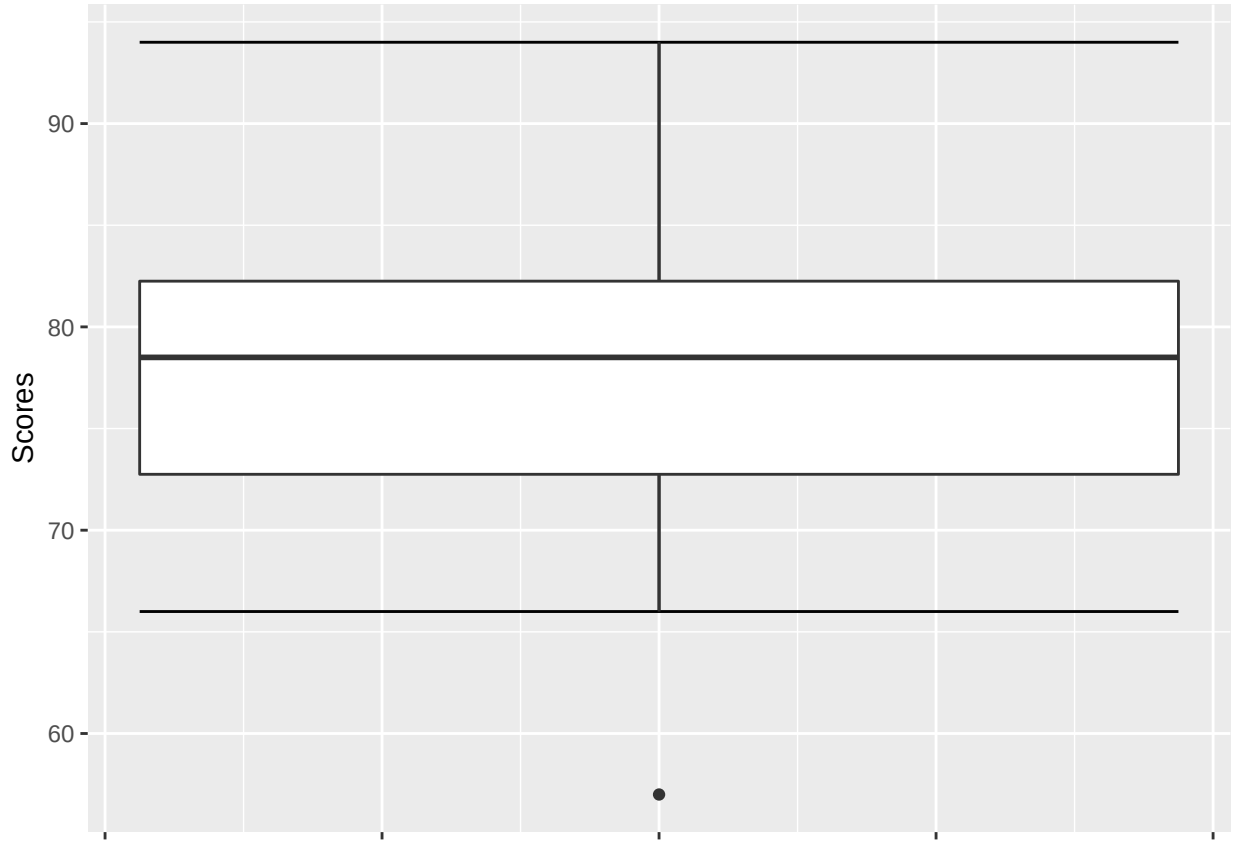
a. This study is an experiment.
b. The treatment group consists of subjects selected to exercise and the control group consists of subjects instructed not to exercise.
c. Yes. Age is the blocking variable, so subjects are divided into age groups.
d. No. Each subject knows what group he is in, so the study is not blind.
e. Since the study is an experiment, it can be used to show causal relationship between exercise and mental health, but only within selected age groups. I believe under 18 and over 55 age groups may show significant differences from selected stratas, so conclusion cannot be generalized to the entire population.
f. I would have some reservations about funding this study. I am not sure if it is possible to design this study to make it blind, but that would be desirable. And I think the study should eliminate other factors that may influence mental health.

1.48 Stats scores

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

```
##    0%   25%   50%   75%  100%
## 57.00 72.75 78.50 82.25 94.00
```

```
ggplot(, aes(scores, x = 1)) + stat_boxplot(geom = "errorbar") + geom_boxplot() +
  theme(axis.title.x=element_blank(), axis.text.x=element_blank()) + labs(y = "Scores")
```



1.50 Mix-and-match

- Histogram (a) is symmetric and unimodal, matches box plot (2)
- Histogram (b) is multimodal, matches box plot (3)
- Histogram (c) is right skewed and unimodal, matches box plot (1)

1.56 Distribution and appropriate statistics, Part II

a. The distribution of housing prices is likely **right skewed** since there is a concetration of house prices around $350,000 - $450,000 with a meaningful number significantly higher. Therefore the center would be best described by the **median**, and variability would be best described by the **IQR**.
b. The distribution of housing prices is likely **symmetric** as there is about equal number of houses sold per each $300,000 with just a few outliers past $1,200,000. Therefore the center would be best described by the **mean**, and variability by the **standard deviation**.

c. The distribution of number of drinks consumed by students is likely **right skewed** since there is a natural boundary at 0 and most student do not drink with very few students drinking excessively. Therefore the center would be best described by the **meadian**, and variability by the **IQR**.

d. The way the question is phrased, the distribution of annual salaries of the employees at a Fortune 500 company may be symmetric since there are few outliers and majority of employees should fall close together; however, I think it is more likely that the distribution if **right skewed** since there are more supporting personnel which early less than management, so as salaries grow the number of employees taper out. If right skewed, the center would be best described by the **meadian**, and variability by the **IQR**.

1.70 Heart transplants

a. The mosaic plot clearly shows that larger proportion of treatment patients survived than proportion of survived control patients. Therefore whether or not the patient got a transplant and survival are **dependent** variables.

b. The box plot for the control group shows not only that survival was low, but that there was no spread in how many days patients survived. It appears that majority of patients survived less than 100 days with a few outliers, and only 1 patient survived for more than 500 days. The box plot for the treatment group shows much wider distribution with more than half of patients surviving for longer than majority of control patients.

c. In the treatment group 45 patients died out of 69, so the proportion is $45/69 = 0.65$. In the control group 30 patients died out of 34, so the proportion is $30/34 = 0.88$.

d. Randomization technique:

   i. H_0: The survival of patients and whether or not they got a transplant are independent. They have no relationship, and the difference in survival rates is due to chance. H_A: The survival of patients and whether or not they got a transplant are not independent. The differentce in survival rates is not due to chance, and not getting a transplant readuces your chance of survival.

   ii. We write *alive* on **28** cards representing patiens 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 center at **0**. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are **at least the difference observed in the study outcome, 24/69 - 4/34 = 0.35 - 0.12 = 0.23**. 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. The simulation results show that there were only 2 simulations with a difference of at least 23% difference. This indicates a rare event and it is unlikely that the difference in the study outcome is due to chance. Although it is possible that the study demonstrated a rare event, it is unlikely and we should accept the altenative model that transplants increase survival rate of patients.