

Data 606: Chapter-1 Homework

Student Name : Sachid Deshmukh

Date : 09/09/2018

- GitHub Location for rmd file
 - GitHub Location for pdf file
 - RPub location of published file
-

1. 1.8 Smoking habits of UK residents

- Each row of the data matrix represents an observation within a sample collected as part of the smoking habit study of UK resident
- 1691 participants were included in the survey
- Variable Types
 - sex - Categorical
 - age - Numerical, discrete
 - marital - Categorical
 - grossIncome - Categorical, ordinal
 - smoke - Categorical
 - amtWeekends - Numerical, discrete
 - amtWeekdays - Numerical, discrete

2. 1.10 Cheaters, scope of inference

- Population of interest : All children between ages of 5 to 15. Sample : 160 observations collected as part of experiment
- Result of the study can be generalized to the population and finding of the study can be used to establish causal relationship if assumption listed below are true
 - Sample of 160 children represents the population proportion of
 - children between 5 to 15 ages adequately.
 - Sample is collected randomly without any bias.

3. 1.28 Reading the paper

- Based on the study we can not conclude that smoking certainly causes dementia later in life. Not 100% of the participant experienced dementia later in life, however study suggests that smokers were more likely to cause dementia compare to non smoker. Risk factors for smokers were 37% to 44% higher compared to non smokers for having dementia later in life.
- The statement “Sleep disorder leads to bullying in school children is not justified. Based on the study it can be concluded that there is a positive correlation between behavioral issues and sleep disorders. Children with behavioral issues are 200% more prone to sleep disorders compared to normal children

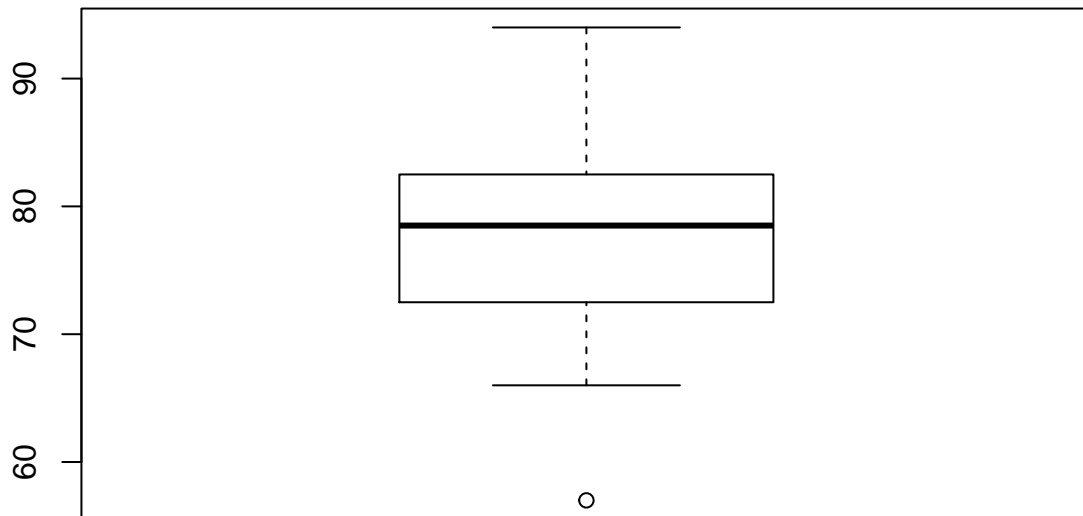
4. 1.36 Exercise and mental health

- Randomized Experiment
- Treatment Group - Doing exercise, Control Group - Not doing exercise
- Study makes use of blocking. Age is the blocking variable
- This study doesn't make use of blinding

- This is a well designed randomized experiment. we are making sure sample represents population proportions effectively. We are also taking care of randomization to reduce bias. Use of blocking makes sure we are studying population with it's true representation. This study can be used to establish a causal relationship between exercise and mental health. Conclusions of the study can be generalized to population at large
- I am in opinion that the study should get formal funding. I will emphasize more on the type of exercise participant went through while study to add more clarity on what type of exercise are helpful to mental health

5. 1.48 Stat Scores

```
Stat.Score = c(57,66,69,71,72,73,74,77,78,78,79,79,81,81,82,83,83,88,89,94)
boxplot(Stat.Score)
```



6. 1.50 Mix and match

- Diagram a represents normal distribution and matches with boxplot 3
- Diagram b represents uniform distribution and matches with boxplot 2
- Diagram c represents skewed distribution with skew in right side. It matches with boxplot 1

7. 1.56 Distributions

- For this scenario, mean is greater than median (\$450,000). This represents right skewed distribution
- For this scenario mean is same as median (600,000). This represents normal distribution with no skew
- This scenario represents range. No drinkers and heavy drinkers. This represents a uniform distribution
- For this scenario, mean is greater than median. This represents right skewed distribution

8. 1.70 Heart Transplants

- Based on the mosaic plot it is evident that survival proportion in in the treatment group is significantly

higher than the control group This suggests that survival is dependent on whether patient got a heart transplant.

- Box plot suggests that mean survival time for the treatment group is higher than control group. It also suggests that, patient in treatment groups have longer survival time compared to patient in control group.

•

```
prop.test(c(30,45), c(34,69))
```

```
##
## 2-sample test for equality of proportions with continuity
## correction
##
## data:  c(30, 45) out of c(34, 69)
## X-squared = 4.9891, df = 1, p-value = 0.02551
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.05215768 0.40820037
## sample estimates:
##      prop 1      prop 2
## 0.8823529 0.6521739
```

88.23% patients in control groups died compared to 65% patients in treatment group

- Randomization Technique
 - Experimental heart transplant results in increased lifespan for heart transplant eligible designated patients or not.
 - We write alive on **alive** cards representing patients who were alive at the end of the study, and dead on **dead** cards representing patients who were not. Then we shuffle these cards and split them in to two groups. One group of size **69** representing treatment, and other group of size **34** representing control. We calculate the difference between 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 **zero** Lastly we calculate the fraction of simulations where the simulated differences in proportions are **zero**. 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 alternate hypothesis
 - The simulation results suggests that the fraction of observations where difference in proportions of dead patient in treatment and control group is zero is less. This suggests that we can safely reject null hypothesis and accept alternate hypothesis. We can conclude that heart transplant helps in increasing lifespan of the patients.