Homework\_1

Priya Shaji

February 7, 2019

## 1.8 Smoking habits of UK residents.

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

Each row repressents the responses of each individual respondants who answered the survey questions.

1. How many participants were included in the survey?

smoking <- read.csv(url("https://raw.githubusercontent.com/jbryer/DATA606Spring2019/master/data/os3\_data/Ch%201%20Exercise%20Data/smoking.csv"), header = FALSE)  
nrow(smoking)

## [1] 1692

Answer (b)

1691 participants

1. 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.

Answer (c)

gender - categorical  
age - numerical (discrete) maritalStatus - categorical highestQualification - categorical(ordinal) nationality - categorical ethnicity - categorical grossincome - categorical (ordinal) region - categorical smoke - categorical  
amtWeekends - numerical (discrete) AmtWeekdays - numerical (discrete) type - categorical

## 1.10 Cheaters, scope of inference.

1. Identify the population of interest and the sample in this study.

Answer (a)

The population of interest are childern, the sample is 160 children aged between 5 and 15

1. 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

Answer (b)

The results of the study cannot be generalised since it’s not sure whether the population of analysis were chosen based on certain criteria or based on certain factors. The findings of the study are based on experimental analysis, therefore it cannot be used to build casual relationships.

## 1.28 Reading the paper

Answer (a) according to the data given: pack-a-day 37% more likely two-pack-a-day 44% more likely more -than-two twice the risks Therefore, by these observations , we can conclude that smoking causes dimentia later in life and risks are likely to increase with increase in smoking rate.

Answer (b) Sleep disorders maynot lead to bullying in children. A child can be a bully based on various other factors, like, various family issues that disturbs the mental well being of a child, bullying child might have gone through same bullying experiance , child abuse, etc. Therefore, corelating sleep disorder only with bullying in children is not justified.

## 1.36 Exercise and mental health.

1. What type of study is this? Answer (a)

it is a randomized study

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

treatment group: the group which is told to exercise twice a week control group: the group who is told not to exercise.(half the subject)

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

Yes, this study makes use of age groups as a blocking variable.

1. Does this study make use of blinding? Answer (d)

No blinding is used. both the respondants and researchers are aware of the group who are exercising and not exercising.

1. 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. Answer (e)

There was random sampling of the population and also each group were assigned tasks to do, therefore the results of the study can be used to establish a causal relationship between exercise and mental health.

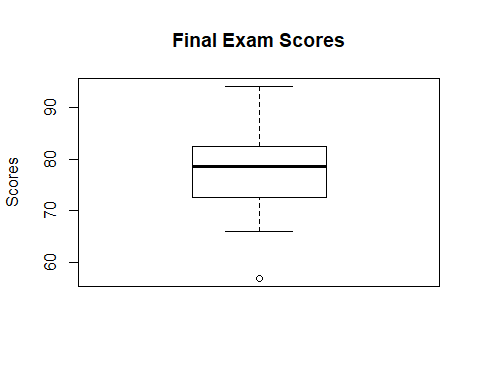
Conclusion can be generalised to the population , since this is a randomized experiment.

1. Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal? Answer (f)

No,I would not have any restrictions on the study proposal, since the analysis is done based on factors which are analytically strong ,like, randomized sample, blocking age groups, compairing results, clear instructions given for the experiment.

## 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))  
boxplot(scores, main = "Final Exam Scores", ylab = "Scores")



## 1.50 Mix-and-match.

histogram (A) matches box plot (2) : it is symmetrical, unimodal and uniformly distributed

histogram (B) matches with box plot (3): it is symmetrical, multimodal and a rough uniform distribution

histogram (C) matches with box plot (1): it is right skewed distribution, unimodal

## 1.56 Distributions and appropriate statistics, Part II .

1. Distribution will be left skewed: the 3rd quartile is less densely populated than the first two quartiles.

Median would best represent this typical observation in the data: it will lessen the effect of extreme values

variability of observations would be best represented by IQR, because SD can be sensitive to extreme values.

(b)Distribution seems to be symmetric. There are only few expensive houses, which would not skew the data much and the quartile ranges are similar

Median would best represent this typical observation in the data: it will lessen the effect of extreme values

variability of observations would be best represented by IQR, because SD can be sensitive to extreme values.

1. It will be left-skewed distribution since most of the students will be at the minimum value of zero(non-drinkers) and since very few drink excessively.

The median would best represent the typical observation since it will lessen the effects of the all the non-drinkers and the excessive drinkers.

The variability would be best represented by the IQR because the SD would be sensitive to all the non-drinkers and the excessive drinkers.

1. It will be a symetrical distribution.

The median would best represent the typical observation since it will lessen the effect of the extreme values of the high-level executives.

The variability would be best represented by the IQR because the SD would be sensitive to the extreme values of the high-level executives.

## 1.70 Heart transplants.

heartTr<- read.csv(url("https://raw.githubusercontent.com/jbryer/DATA606Spring2019/master/data/os3\_data/Ch%201%20Exercise%20Data/heart\_transplant.csv"))  
heartTr

## 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  
## 7 54 71 47 dead 3 no control NA  
## 8 38 70 41 dead 5 no treatment 5  
## 9 85 73 47 dead 5 no control NA  
## 10 2 68 51 dead 6 no control NA  
## 11 103 67 39 dead 6 no control NA  
## 12 12 68 53 dead 8 no control NA  
## 13 48 71 56 dead 9 no control NA  
## 14 102 74 40 alive 11 no control NA  
## 15 35 70 43 dead 12 no control NA  
## 16 95 73 40 dead 16 no treatment 2  
## 17 31 69 54 dead 16 no control NA  
## 18 3 68 54 dead 16 no treatment 1  
## 19 74 72 29 dead 17 no treatment 5  
## 20 5 68 20 dead 18 no control NA  
## 21 77 72 41 dead 21 no control NA  
## 22 99 73 49 dead 21 no control NA  
## 23 20 69 55 dead 28 no treatment 1  
## 24 70 72 52 dead 30 no treatment 5  
## 25 101 74 49 alive 31 no control NA  
## 26 66 72 53 dead 32 no control NA  
## 27 29 69 50 dead 35 no control NA  
## 28 17 68 20 dead 36 no control NA  
## 29 19 68 59 dead 37 no control NA  
## 30 4 68 40 dead 39 no treatment 36  
## 31 100 74 35 alive 39 yes treatment 38  
## 32 8 68 45 dead 40 no control NA  
## 33 44 70 42 dead 40 no control NA  
## 34 16 68 56 dead 43 no treatment 20  
## 35 45 71 36 dead 45 no treatment 1  
## 36 1 67 30 dead 50 no control NA  
## 37 22 69 42 dead 51 no treatment 12  
## 38 39 70 50 dead 53 no treatment 2  
## 39 10 68 42 dead 58 no treatment 12  
## 40 35 71 52 dead 61 no treatment 10  
## 41 37 70 61 dead 66 no treatment 19  
## 42 68 72 45 dead 68 no treatment 3  
## 43 60 71 49 dead 68 no treatment 3  
## 44 62 71 39 dead 69 no control NA  
## 45 28 69 53 dead 72 no treatment 71  
## 46 47 71 47 dead 72 no treatment 21  
## 47 32 69 64 dead 77 no treatment 17  
## 48 65 72 51 dead 78 no treatment 12  
## 49 83 73 53 dead 80 no treatment 32  
## 50 13 68 54 dead 81 no treatment 17  
## 51 9 68 47 dead 85 no control NA  
## 52 73 72 56 dead 90 no treatment 27  
## 53 79 72 53 dead 96 no treatment 67  
## 54 36 70 48 dead 100 no treatment 46  
## 55 32 71 41 dead 102 no control NA  
## 56 98 73 28 alive 109 no treatment 96  
## 57 87 73 46 dead 110 no treatment 60  
## 58 97 73 23 alive 131 no treatment 21  
## 59 37 71 41 dead 149 no control NA  
## 60 11 68 47 dead 153 no treatment 26  
## 61 94 73 43 dead 165 yes treatment 4  
## 62 96 73 26 alive 180 no treatment 13  
## 63 90 73 52 dead 186 yes treatment 160  
## 64 53 71 47 dead 188 no treatment 41  
## 65 89 73 51 dead 207 no treatment 139  
## 66 24 69 51 dead 219 no treatment 83  
## 67 27 69 8 dead 263 no control NA  
## 68 93 73 47 alive 265 no treatment 28  
## 69 51 71 48 dead 285 no treatment 32  
## 70 67 73 19 dead 285 no treatment 57  
## 71 16 68 49 dead 308 no treatment 28  
## 72 84 73 42 dead 334 no treatment 37  
## 73 91 73 47 dead 340 no control NA  
## 74 92 73 44 alive 340 no treatment 310  
## 75 58 71 47 dead 342 yes treatment 21  
## 76 88 73 54 alive 370 no treatment 31  
## 77 86 73 48 alive 397 no treatment 8  
## 78 82 71 29 alive 427 no control NA  
## 79 81 73 52 alive 445 no treatment 6  
## 80 80 72 46 alive 482 yes treatment 26  
## 81 78 72 48 alive 515 no treatment 210  
## 82 76 72 52 alive 545 yes treatment 46  
## 83 64 72 48 dead 583 yes treatment 32  
## 84 72 72 26 alive 596 no treatment 4  
## 85 71 72 47 alive 630 no treatment 31  
## 86 69 72 47 alive 670 no treatment 10  
## 87 7 68 50 dead 675 no treatment 51  
## 88 23 69 58 dead 733 no treatment 3  
## 89 63 71 32 alive 841 no treatment 27  
## 90 30 69 44 dead 852 no treatment 16  
## 91 59 71 41 alive 915 no treatment 78  
## 92 56 71 38 alive 941 no treatment 67  
## 93 50 71 45 dead 979 yes treatment 83  
## 94 46 71 48 dead 995 yes treatment 2  
## 95 21 69 43 dead 1032 no treatment 8  
## 96 49 71 36 alive 1141 yes treatment 36  
## 97 41 70 45 alive 1321 yes treatment 58  
## 98 14 68 53 dead 1386 no treatment 37  
## 99 26 69 30 alive 1400 no control NA  
## 100 40 70 48 alive 1407 yes treatment 41  
## 101 34 69 40 alive 1571 no treatment 23  
## 102 33 69 48 alive 1586 no treatment 51  
## 103 25 69 33 alive 1799 no treatment 25

1. As per the mosaic plot, the survival is not independent since the expectancy of life is bigger for the patients who got the heart transplant.
2. The box plot shows that that the heart transplant is effective for increase of life expectancy.

patientcont\_dead <- nrow(subset(heartTr, heartTr$transplant ==   
 "control" & heartTr$survived == "dead"))  
patientcont <- nrow(subset(heartTr, heartTr$transplant ==   
 "control"))  
patienttreat\_dead <- nrow(subset(heartTr, heartTr$transplant ==   
 "treatment" & heartTr$survived == "dead"))  
patienttreat <- nrow(subset(heartTr, heartTr$transplant ==   
 "treatment"))  
patientcont\_deadratio <- patientcont\_dead/patientcont  
patienttreat\_deadratio <- patienttreat\_dead/patienttreat  
patientcont\_deadratio

## [1] 0.8823529

patienttreat\_deadratio

## [1] 0.6521739

88.23% of the patients in the control group died by the end of the study and 65.22% of the patients in the treatment group died by the end of the study.

1. The claim being tested is whether or not a heart transplant will increase a patient’s lifespan.

alive <- sum(heartTr$survived == "alive")  
alive

## [1] 28

dead <- sum(heartTr$survived == "dead")  
dead

## [1] 75

patienttreat

## [1] 69

patientcont

## [1] 34

patienttreat\_deadratio - patientcont\_deadratio

## [1] -0.230179

28 75 69 34 -0.230179 at least as extreme or greater.

It seems that a difference of at least -23.02% due to chance alone would only happen about 2% of the time according to the figure. Such a low probability indicates a rare event.