Homework\_1

Ruoyuan Qian

2019/9/15

# Problem 1

## a)

I found that there are two rows of variable names which may cause some errors when reading the data into R. As a result, I use skip= to skip first row and read others into R.

library(readxl)  
  
#raw\_data <- read.table('C:\\Users\\hp\\Desktop\\Biostats\\homework\\Exercise 11.txt',header =TRUE,skip = 1)  
 raw\_data<-read\_excel("C:\\Users\\hp\\Desktop\\Biostats\\homework\\Exercise.xlsx",skip =1)  
   
#check variables and values in raw\_data  
 str(raw\_data)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 72 obs. of 20 variables:  
## $ Group : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Age : num 57 65 61 54 64 41 59 57 67 44 ...  
## $ Gender : num 1 1 1 2 1 2 1 2 1 2 ...  
## $ Race : num 1 1 1 2 1 1 1 2 1 1 ...  
## $ HTN : num 1 1 1 1 1 0 1 0 1 0 ...  
## $ T2DM : num 0 1 0 1 1 0 1 0 1 0 ...  
## $ Depression: num 0 0 0 0 0 1 0 0 0 0 ...  
## $ Smokes : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ PRE...9 : num 160 126 120 140 148 116 142 123 160 106 ...  
## $ POST...10 : num 163 92 121 121 123 130 121 113 124 111 ...  
## $ PRE...11 : num 102 59 67 81 63 82 69 77 81 65 ...  
## $ POST...12 : num 107 57 68 65 58 87 68 65 80 65 ...  
## $ PRE...13 : num 33 25.7 26.1 41.6 29.2 31.7 27.6 25.7 36.6 32.3 ...  
## $ POST...14 : num 32.7 25.7 25.3 39.7 28.4 31.3 28.4 25.8 35.6 31.9 ...  
## $ PRE...15 : num 60 40 88 44 48 66 37 69 40 45 ...  
## $ POST...16 : num 62 43 67 24 52 56 44 73 35 49 ...  
## $ PRE...17 : num 110 133 114 112 63 62 89 117 77 126 ...  
## $ POST...18 : num 107 96 98 75 58 86 81 129 73 131 ...  
## $ PRE...19 : num 96 106 92 401 96 75 66 96 113 91 ...  
## $ POST...20 : num 105 132 95 162 216 92 72 71 101 92 ...

#check if there is a missing value  
 anyNA(raw\_data)

## [1] FALSE

#only 3 kinds of race in the data  
 table(raw\_data$Race)

##   
## 1 2 3   
## 51 19 2

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(arsenal)  
  
#rename some values into a readable way   
 fixed\_data <- raw\_data %>% mutate(  
 Group = recode(Group, "0" = "Intervention", "1" = "Control"),  
 Gender = recode(Gender, "1" = "male", "2" = "female"),  
 HTN = recode(HTN, "0" = "no", "1" = "yes"),  
 T2DM = recode(T2DM, "0" = "no", "1" = "yes"),  
 Smokes = recode(Smokes, "0" = "no", "1" = "yes"),  
 Depression = recode(Depression, "0" = "no", "1" = "yes"),  
 Race = recode(Race, "1" = "African American",   
 "2" = "Hispanic",  
 "3" = "African American"))  
  
#draw the descriptive table  
 sum\_data <- arsenal::tableby(Group ~ Age + Gender + Race + Depression + Smokes + HTN+T2DM,   
 data = fixed\_data,  
 test = FALSE,   
 total = FALSE,  
 numeric.stats = c("meansd" ,"median" ) )  
 summary(sum\_data,text = TRUE)

##   
##   
## | | Control (N=36) | Intervention (N=36) |  
## |:-------------------|:--------------:|:-------------------:|  
## |Age | | |  
## |- Mean (SD) | 53.583 (9.581) | 51.500 (10.809) |  
## |- Median | 55.500 | 51.000 |  
## |Gender | | |  
## |- female | 20 (55.6%) | 20 (55.6%) |  
## |- male | 16 (44.4%) | 16 (44.4%) |  
## |Race | | |  
## |- African American | 31 (86.1%) | 22 (61.1%) |  
## |- Hispanic | 5 (13.9%) | 14 (38.9%) |  
## |Depression | | |  
## |- no | 26 (72.2%) | 23 (63.9%) |  
## |- yes | 10 (27.8%) | 13 (36.1%) |  
## |Smokes | | |  
## |- no | 31 (86.1%) | 31 (86.1%) |  
## |- yes | 5 (13.9%) | 5 (13.9%) |  
## |HTN | | |  
## |- no | 14 (38.9%) | 16 (44.4%) |  
## |- yes | 22 (61.1%) | 20 (55.6%) |  
## |T2DM | | |  
## |- no | 23 (63.9%) | 17 (47.2%) |  
## |- yes | 13 (36.1%) | 19 (52.8%) |

## b)

### b) i

# merge BMI\_PRE and BMI\_POST into one column  
 a <- dplyr::select(fixed\_data, c(Group,PRE...13 ))  
 b <- dplyr::select(fixed\_data, c(Group,POST...14))  
 bmi\_pre <- rename(a,BMI=PRE...13 )  
 bmi\_post <- rename(b,BMI=POST...14)  
   
# add a new variable that helps to illustrate time status of BMI, either baseline or post-measurement  
 len\_pre <- length(bmi\_pre$Group)  
 len\_post <- length(bmi\_pre$Group)  
 data\_b <- rbind(cbind(bmi\_pre, time=rep(0,len\_pre )),  
 cbind(bmi\_post,time=rep(1,len\_post)))  
 data\_bmi <- data\_b %>% mutate(  
 time = recode(time, "0" = "baseline",   
 "1" = "post-measurement"))  
 int\_pre<-filter(data\_bmi, Group=="Intervention" ,time=="baseline")  
 int\_post<-filter(data\_bmi, Group=="Intervention" ,time=="post-measurement")  
 con\_pre<-filter(data\_bmi, Group=="Control" ,time=="baseline")  
 con\_post<-filter(data\_bmi, Group=="Control" ,time=="post-measurement")  
  
 mean\_int\_pre<- mean( int\_pre$BMI)  
 mean\_int\_post<- mean( int\_post$BMI)  
 sd\_int\_pre<-sd(int\_pre$BMI)  
 sd\_int\_post<-sd(int\_post$BMI)  
   
 mean\_con\_pre<- mean( con\_pre$BMI)  
 mean\_con\_post<- mean( con\_post$BMI)  
 sd\_con\_pre<-sd(con\_pre$BMI)  
 sd\_con\_post<-sd(con\_post$BMI)  
  
 abs(max((mean\_int\_pre+sd\_int\_pre),(mean\_int\_post+sd\_int\_post))-  
 min((mean\_int\_pre-sd\_int\_pre),(mean\_int\_post-sd\_int\_post)))

## [1] 12.413

### b) ii

# draw side by side boxplot of BMI  
library(ggplot2)  
 ggplot(data\_bmi,aes(x = Group, y = BMI, fill = time)) +  
 geom\_boxplot() +  
 scale\_fill\_manual(values = c("blue", "red"),   
 labels = expression("baseline", "post-measurement")) +   
 xlab("") +   
 ylab("BMI ") +   
 theme(legend.position = "top", legend.title = element\_blank())

ii-1.png)

# merge DLD\_PRE and DLD\_POST into one column  
 a1 <- dplyr::select(fixed\_data, c(Group,PRE...17 ))  
 b1 <- dplyr::select(fixed\_data, c(Group,POST...18))  
 ldl\_pre <- rename( a1,LDL = PRE...17 )  
 ldl\_post <- rename( b1,LDL = POST...18)  
   
# add a new variable that helps to illustrate time status of DLD, either baseline or post-measurement  
 len\_pre <- length( ldl\_pre$Group )  
 len\_post <- length( ldl\_pre$Group )  
 data\_l <- rbind( cbind(ldl\_pre,time=rep(0,len\_pre)),  
 cbind(ldl\_post,time=rep(1,len\_post)))  
 data\_ldl <- data\_l %>% mutate(  
 time = recode(time, "0" = "baseline",   
 "1" = "post-measurement"))  
  
 # draw side by side boxplot of LDL  
 ggplot(data\_ldl,   
 aes(x = Group, y = LDL, fill = time)) +  
 geom\_boxplot() +   
 scale\_fill\_manual(values = c("blue", "red"),  
 labels = expression("baseline", "post-measurement")) +   
 xlab("") +  
 ylab("LDL ") +   
theme(legend.position = "top", legend.title = element\_blank())

ii-2.png)

### b) iii

As for the BMI in the study, in general, the values in intervention group are higher than the control group’s while there is little difference of values within each group.

As for the LDL, for both groups, there is a trend that the values in post measurement are lower than the baseline’s, and it is more obvious in control. Besides, in intervention , the values are more concentrated whereas values in control are more dispersed.

## c)

It is an interventional clinical trial, and from the perspective of Demographics, the two groups are basically balanced, which is favorable for the study. However, there is some potential issues as well. For instance, the medians of baselines of BMI in two groups are of great difference, which may make the two groups incomparable. Besides, as for the LDL, the variances of two groups are extremely different, which may cause the similar problem like BMI.

# Problem 2

The probability is The probability is(0.6×0.001)/(0.6×0.001+0.05×0.999)=0.012,