Homework 2

Jia-Ru Chung

Octobor 4, 2017

**Note:**  
All answers have to be obtained using R code!

#### Max points = 80

### Question 1 {20 pts}.

**1.1** Load the two datasets DrugA.rdata and DrugB.csv. Each dataset contains patient information from two treatment groups, treated with either DrugA or DrugB. Which treatment group had the most responders to treatment? (see 'Trt\_Response' column, 1=Responder, 0=non-responder) For the treatment-group with highest number of responders, what proportion of patients responded to treatment? {10 pts}

DrugB <- read.table("DrugB.csv", header = TRUE, sep = ",")  
head(DrugB)

## AnonID Time Delta\_Severity Age Trt\_Response bmi\_day1 bmi\_week12  
## 1 205575 Week 0 12.8 47 0 26.49417 27.11191  
## 2 205585 Week 0 27.6 21 0 24.05777 24.74745  
## 3 205603 Week 0 23.3 52 1 25.68007 26.24742  
## 4 205617 Week 0 21.6 38 1 26.04167 26.46684  
## 5 205621 Week 0 8.8 37 0 33.10229 32.76623  
## 6 205641 Week 0 11.5 27 1 23.92486 23.92486  
## SampleNames bmi\_day1NA alc\_drinks\_per\_week  
## 1 601638 26.49417 1  
## 2 607214 NA 0  
## 3 603068 NA 0  
## 4 606874 26.04167 0  
## 5 603166 33.10229 3  
## 6 600468 NA 0

load(file = "DrugA.rdata")  
head(DrugA)

## AnonID Time Delta\_Severity Age Trt\_Response bmi\_day1 bmi\_week12  
## 1 205567 Week 0 14.7 53 1 29.38776 29.38776  
## 2 205582 Week 0 15.6 61 1 29.41406 29.10156  
## 3 205587 Week 0 40.9 47 1 35.50714 35.35842  
## 4 205594 Week 0 13.1 28 1 25.18006 24.81994  
## 5 205607 Week 0 10.5 51 1 29.05475 29.95868  
## 6 205615 Week 0 10.6 40 0 38.71442 38.45730  
## SampleNames bmi\_day1NA alc\_drinks\_per\_week  
## 1 602594 29.38776 0  
## 2 598830 29.41406 0  
## 3 602084 35.50714 1  
## 4 600418 NA 3  
## 5 606854 29.05475 0  
## 6 609666 38.71442 0

ls()

## [1] "DrugA" "DrugB"

DrugAresponder <- subset(DrugA, Trt\_Response > 0)  
dim(DrugAresponder) #92

## [1] 92 10

DrugAnonresponder <- subset(DrugA, Trt\_Response < 1)  
dim(DrugAnonresponder) #42

## [1] 42 10

DrugBresponder <- subset(DrugB, Trt\_Response > 0)  
dim(DrugBresponder) #84

## [1] 84 10

DrugBnonresponder <- subset(DrugB, Trt\_Response < 1)  
dim(DrugBnonresponder) #41

## [1] 41 10

# DrugA has highest number of responders: 92>84.  
nrow(DrugAresponder)/nrow(DrugA) #92/134

## [1] 0.6865672

**Answer:**

**1.2** Create the new column 'Treatment' in each data set (should be "DrugA" for data DrugA and 'DrugB' for DrugB). Combine (append) the two datasets. What was the proportion of Responders for both treatments combined? Round your answer to 2 decimal places. {10 pts}

DrugAresponder["Treatment"] <- "DrugA"  
DrugBresponder["Treatment"] <- "DrugB"  
  
DrugABresponder <- merge(DrugAresponder, DrugBresponder, all = T)  
dim(DrugABresponder) #176

## [1] 176 11

DrugABnonresponder <- merge(DrugAnonresponder, DrugBnonresponder, all = T)  
dim(DrugABnonresponder) #83

## [1] 83 10

DrugABall <- merge(DrugABresponder, DrugABnonresponder, all = T)  
dim(DrugABall) #259

## [1] 259 11

x <- nrow(DrugABresponder)/nrow(DrugABall) #176/259  
round(x, digits = 2)

## [1] 0.68

**Answer:**

### Question 2 {20 pts}.

**2.1** Upload 'Trt\_Groups.csv' file. Check for duplicates. Remove duplicated observations if there are any.  
Trt\_Response is coded as 1/0. Create a new variable (Trt\_ResponseC) that will have 'R' if Trt\_Response=1 and 'NR' otherwise.  
Also create a character variable for Age (AgeC): if age is less than 40 then AgeC = 'Younger', and 'Older' otherwise. Ensure that Trt\_ResponseC ('R'=first level) and AgeC ('Younger'=firstu level) are factors.  
What percentage of patients were 'Younger' and responded to treatment? (2 decimal places) {20 pts}

Trt\_Groups <- read.csv(file = "Trt\_Groups.csv")  
head(duplicated(Trt\_Groups), 12)

## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [12] FALSE

table(duplicated(Trt\_Groups)) #8 dups

##   
## FALSE TRUE   
## 259 8

dups <- Trt\_Groups[duplicated(Trt\_Groups), "AnonID"]  
dups

## [1] 206596 205732 206173 206422 206803 206647 206939 205985

head(subset(Trt\_Groups, AnonID %in% dups))

## AnonID Time Delta\_Severity Age Trt\_Response bmi\_day1 Treatment  
## 2 206596 Week 0 20.0 31 1 20.88276 DrugB  
## 5 206803 Week 0 17.7 41 1 26.98523 DrugB  
## 19 206422 Week 0 21.1 59 1 24.15882 DrugB  
## 30 206596 Week 0 20.0 31 1 20.88276 DrugB  
## 43 205732 Week 0 25.2 49 1 30.82483 DrugA  
## 49 206173 Week 0 11.1 37 0 36.91358 DrugA  
## bmi\_day1NA alc\_drinks\_per\_week Delta\_IL17A Delta\_IL17C Delta\_IL8  
## 2 NA 3 -0.3176884 -0.4570256 -8.864649  
## 5 26.98523 3 -0.5655564 -0.6579140 -9.202606  
## 19 NA 0 -0.2080723 -0.8942011 -9.238284  
## 30 NA 3 -0.3176884 -0.4570256 -8.864649  
## 43 30.82483 3 -0.9696022 -1.7608355 -1.618041  
## 49 36.91358 28 -0.8253469 -1.2655442 -1.053831

Trt\_GroupsNoDup <- Trt\_Groups[!duplicated(Trt\_Groups), ]  
table(duplicated(Trt\_GroupsNoDup))

##   
## FALSE   
## 259

Trt\_ResponseC <- ifelse(Trt\_GroupsNoDup$Trt\_Response > 0, "R", "NR")  
table(Trt\_ResponseC) # R 176 NR 83

## Trt\_ResponseC  
## NR R   
## 83 176

RinTrt\_ResponseC <- subset(Trt\_ResponseC, Trt\_GroupsNoDup$Trt\_Response > 0)  
table(RinTrt\_ResponseC) # R 176

## RinTrt\_ResponseC  
## R   
## 176

AgeC <- ifelse(Trt\_GroupsNoDup$Age < 40, "Younger", "Older")  
table(AgeC) # Older 156 Younger 103

## AgeC  
## Older Younger   
## 156 103

AgeCyoungresponder.subset <- subset(AgeC, Trt\_GroupsNoDup$Age < 40, Trt\_GroupsNoDup$Trt\_Response >   
 0)  
table(AgeCyoungresponder.subset) #Younger 103

## AgeCyoungresponder.subset  
## Younger   
## 103

x <- length(AgeCyoungresponder.subset)/length(duplicated(Trt\_GroupsNoDup))  
x

## [1] 0.3976834

round(x, digits = 2)

## [1] 0.4

# percentage: 103/259= 0.4

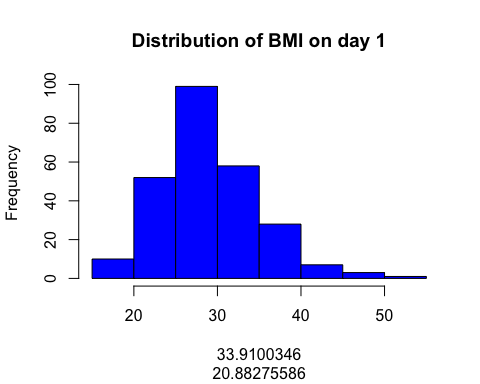
**Answer:**

### Question 3 {40 pts}.

Using the dataset created in 2.1 create plots to explore distribution of bmi (*bmi\_day1*).  
Note: you can use either *{ggplot2}* or *base* graphics.

**3.1** Make a histogram of *bmi\_day1*. Is BMI uniformly distributed (the probability/frequency of each BMI is equal)? {10 pts}

hist(Trt\_GroupsNoDup$bmi\_day1, main = "Distribution of BMI on day 1", xlab = Trt\_GroupsNoDup$bmi\_day1,   
 col = "blue", breaks = 10)

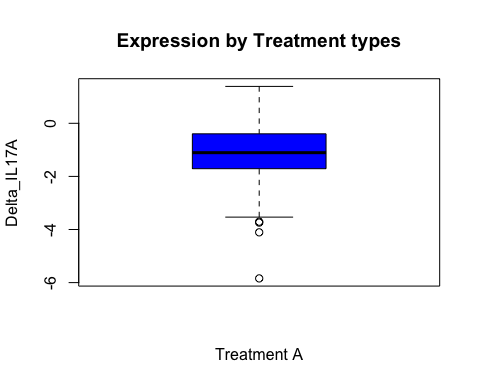


# Yes the frequency is equal.

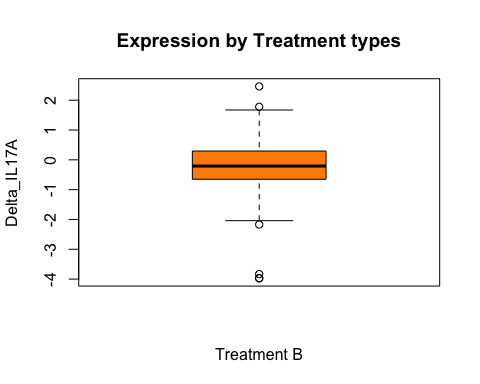
**Answer:**

**3.2** Make boxplots for the change (Delta) in IL17A expression *Delta\_IL17A* in both treatment groups *Treatment*. Based on these plots, which treatment has the biggest effect on IL17A expression? {10 pts}

Trt\_GroupsNoDupDrugA <- subset(Trt\_GroupsNoDup, Treatment == "DrugA")  
Trt\_GroupsNoDupDrugB <- subset(Trt\_GroupsNoDup, Treatment == "DrugB")  
  
  
boxplot(Trt\_GroupsNoDupDrugA$Delta\_IL17A, main = "Expression by Treatment types",   
 xlab = "Treatment A", ylab = "Delta\_IL17A", col = "blue")



boxplot(Trt\_GroupsNoDupDrugB$Delta\_IL17A, main = "Expression by Treatment types",   
 xlab = "Treatment B", ylab = "Delta\_IL17A", col = "darkorange")

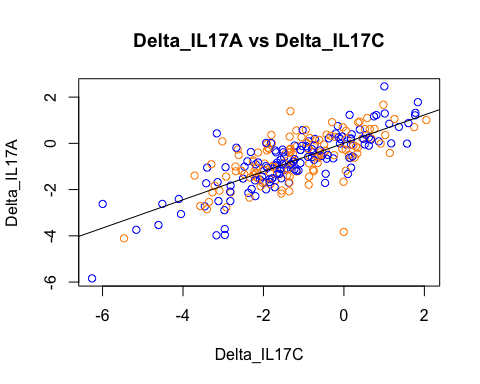


# Treatment B has the biggest effect on IL17A expression since it has higher  
# Delta\_IL17A response.

**Answer:**

**3.3** Make a scatter plot of *Delta\_IL17A* and *Delta\_IL17C*. Add a regression line (in ggplot use *+ geom\_smooth(method='lm')*) to assess if IL17A expression changes IL17C expression. {10 pts}

plot(Trt\_GroupsNoDup$Delta\_IL17A ~ Trt\_GroupsNoDup$Delta\_IL17C, main = "Delta\_IL17A vs Delta\_IL17C",   
 ylab = "Delta\_IL17A", xlab = "Delta\_IL17C", col = c("blue", "darkorange"))  
abline(lm(Trt\_GroupsNoDup$Delta\_IL17A ~ Trt\_GroupsNoDup$Delta\_IL17C))



**3.4** Use a plot of your choice to demostrate the relationship between change in IL17A expression *Delta\_IL17A* and the change in disease severity *Delta\_Severity* by *Treatment*. Make sure your plot is publication quality: has appropriate title, axes labels, distinct colors/shapes and proper legend (if any). {10 pts}

library(ggplot2)  
  
ggplot(data = Trt\_GroupsNoDup, aes(y = Delta\_IL17A, as.numeric(x = Delta\_Severity),   
 fill = Treatment)) + geom\_boxplot() + scale\_fill\_manual(values = c("blue",   
 "orange"), labels = c("DrugA", "DrugB")) + stat\_summary(fun.y = mean, colour = "red",   
 geom = "point", shape = 18, size = 3) + labs(title = "Changes by treatment types",   
 x = "Delta\_Severity", y = "Delta\_IL17A") + theme\_bw() + scale\_x\_continuous()

## Warning: position\_dodge requires non-overlapping x intervals

