Homework 2

Student's Name Here

September 22, 2016

**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}

load("../Homework\_2/DrugA.rdata")  
DrugA = DrugA  
DrugB = read.csv("../Homework\_2/DrugB.csv")  
head(DrugA$Trt\_Response)

## [1] 1 1 1 1 1 0

DrugAR = sum(DrugA$Trt\_Response)  
DrugAR

## [1] 92

DrugBR = sum(DrugB$Trt\_Response)  
DrugBR

## [1] 84

DrugAR > DrugBR

## [1] TRUE

Total = length(DrugA$Trt\_Response)  
print(paste0("The proportion of responders for Drug A is ", signif(DrugAR/Total,   
 2), "%"))

## [1] "The proportion of responders for Drug A is 0.69%"

**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}

DrugA$Treatment = "DrugA"  
DrugB$Treatment = "DrugB"  
DrugBoth = rbind(DrugA, DrugB)  
  
Total = length(DrugBoth$Trt\_Response)  
Resp = sum(DrugBoth$Trt\_Response)  
print(paste0("The proportion of responders for both Drugs is ", signif(Resp/Total,   
 2), "%"))

## [1] "The proportion of responders for both Drugs is 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'=first level) are factors.  
What percentage of patients were 'Younger' and responded to treatment? (2 decimal places) {20 pts}

Data <- read.csv("../Homework\_2/Trt\_Groups.csv")  
  
dups <- Data[duplicated(Data$AnonID), "AnonID"]  
print(dups)

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

Data <- Data[!duplicated(Data$AnonID), ]  
dim(Data)

## [1] 259 12

Data$Trt\_ResponseC <- ifelse(Data$Trt\_Response == 1, "R", "NR")  
Data$Trt\_ResponseC <- factor(Data$Trt\_ResponseC, levels = c("R", "NR"))  
Data$AgeC <- factor(ifelse(Data$Age < 40, "Younger", "Older"), levels = c("Younger",   
 "Older"))  
levels(Data$AgeC)

## [1] "Younger" "Older"

levels(Data$Trt\_ResponseC)

## [1] "R" "NR"

table(Data[, c("AgeC", "Trt\_ResponseC")])

## Trt\_ResponseC  
## AgeC R NR  
## Younger 63 40  
## Older 113 43

DataYR <- subset(Data, Trt\_ResponseC == "R" & AgeC == "Younger")  
num = nrow(DataYR)  
print(paste0(signif(num/nrow(Data), 2), "% of patients were Younger and responded to treatment"))

## [1] "0.24% of patients were Younger and responded to treatment"

**Answer:**

### Question 3 {40 pts}.

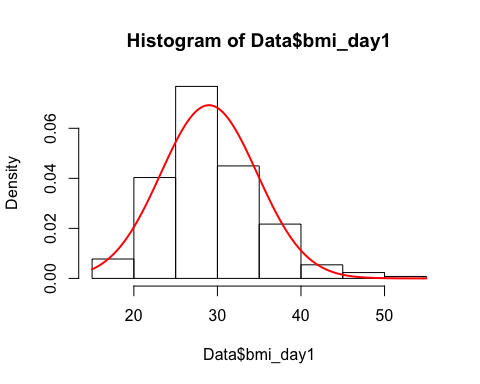
Using the dataset created in 2.1 (the one with duplicated removed and Trt\_reponse modified) create plots to explore distribution of bmi (*bmi\_day1*).  
Note: you can use either *{ggplot2}* or *base* graphics.

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

hist(Data$bmi\_day1, breaks = 10, freq = F)  
is.numeric(Data$bmi\_day1)

## [1] TRUE

curve(dnorm(x, mean = mean(Data$bmi\_day1, na.rm = T), sd = sd(Data$bmi\_day1,   
 na.rm = T)), add = T, col = "red", lwd = 2)



print("BMI is not uniformly distributed")

## [1] "BMI is not uniformly distributed"

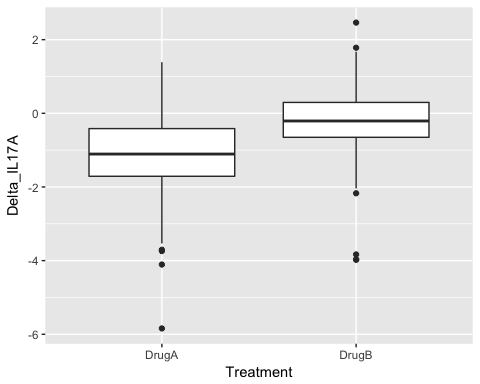
**Answer:**

**3.2** Make boxplots for the change 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}

require(ggplot2)

## Loading required package: ggplot2

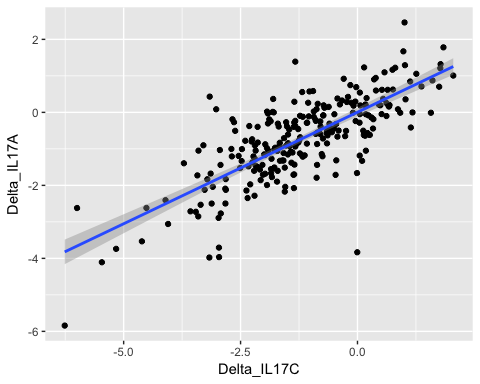
ggplot(Data, aes(x = Treatment, y = Delta\_IL17A)) + geom\_boxplot()



**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}

ggplot(Data, aes(x = Delta\_IL17C, y = Delta\_IL17A)) + geom\_point() + geom\_smooth(method = lm)



**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}

ggplot(Data, aes(x = Delta\_Severity, y = Delta\_IL17A, group = Treatment, color = Treatment)) +   
 geom\_point() + geom\_smooth(method = lm) + labs(title = expression(paste(Delta,   
 "Severity vs ", Delta, "IL17A-expression")), x = expression(paste(Delta,   
 "Severity")), y = expression(paste(Delta, "IL17A-expression"))) + theme(plot.title = element\_text(hjust = 0.5))

