STAT 1293 - Midterm I

Gordon Lu

7/13/2020

Problem 1: Objects (25 points)

1a) Create a vector as follows. Call it `v`. (2 points)

Solution:

```
v <- c(0, 1, -2, 2, -1, 0)
v
```

```
## [1]  0  1 -2  2 -1  0
```

1b) Create a matrix as follows, call it `M`. Don't forget the column names. (5 points)

Solution:

```
c1 <- c("1", "2", "3", "4", "5")
c2 <- letters[21:25]
c3 <- c("TRUE", "TRUE", "FALSE", "FALSE", "FALSE")
M <- cbind(c1,c2,c3)
M
```

```
##      c1  c2  c3
## [1,] "1" "u" "TRUE"
## [2,] "2" "v" "TRUE"
## [3,] "3" "w" "FALSE"
## [4,] "4" "x" "FALSE"
## [5,] "5" "y" "FALSE"
```

1c) Create a data frame as follows. Call it `df`. (4 points)

Solution:

```
df <- data.frame(c1,c2,c3)
df
```

```
##   c1 c2    c3
## 1  1  u  TRUE
## 2  2  v  TRUE
## 3  3  w FALSE
## 4  4  x FALSE
## 5  5  y FALSE
```

1d) Create a list as follows. Call it `mylist`. (5 points)

Solution:

```r
mylist <- list(v, M[, c(1,3)], df)
mylist
```

```
## [[1]]
## [1]  0  1 -2  2 -1  0
##
## [[2]]
##      c1  c3
## [1,] "1" "TRUE"
## [2,] "2" "TRUE"
## [3,] "3" "FALSE"
## [4,] "4" "FALSE"
## [5,] "5" "FALSE"
##
## [[3]]
##   c1 c2    c3
## 1  1  u  TRUE
## 2  2  v  TRUE
## 3  3  w FALSE
## 4  4  x FALSE
## 5  5  y FALSE
```

1e) Pull the third row of `df`. (2 points)

Solution:

```r
df[3, ]
```

```
##   c1 c2    c3
## 3  3  w FALSE
```

1f) Pull out the second sub-list in `mylist` (2 points)

Solution:

```r
mylist[[2]]
```

```
##      c1  c3     
## [1,] "1" "TRUE" 
## [2,] "2" "TRUE" 
## [3,] "3" "FALSE"
## [4,] "4" "FALSE"
## [5,] "5" "FALSE"
```

1g) Convert `v` to a factor-type vector (call it `v.f`) and redefine the levels as "Strongly Disagree","Disagree", "Neutral","Agree","Strongly Agree", with -2 being "Strongly Disagree" and 2 being "Strongly Agree". (5 points)

Solution:

```r
v.f <- factor(v, levels = -2:2)
levels(v.f) <- c("Strongly Disagree","Disagree", "Neutral","Agree","Strongly Agree")
v.f
```

```
## [1] Neutral           Agree             Strongly Disagree Strongly Agree   
## [5] Disagree          Neutral          
## Levels: Strongly Disagree Disagree Neutral Agree Strongly Agree
```

Problem 2: Probability Distributions (40 points)

2a) Find `P(-3 < X < 3)` where `X ~ N(2,2)` (3 points)

Solution:

```r
pnorm(3, mean = 2, sd = 2) - pnorm(-3, mean = 2, sd = 2)
```

```
## [1] 0.6852528
```

2b) Find `P(X = 15)` where `X ~ Binomial(20, 0.8)` (2 points)

Solution:

```r
dbinom(x = 15, size = 20, prob = 0.8)
```

```
## [1] 0.1745595
```

2c) Generate 100 observations from the t_{20} and calculate the mean and standard deviation of them. (5 points)

Solution:

```r
t_20 <- rt(100, 20)
mean(t_20)
```

```
## [1] 0.1455049
```

```r
sd(t_20)
```

```
## [1] 1.118598
```

2d) Find the 99th percentile of the χ^2_5 distribution. (2 points)

Solution:

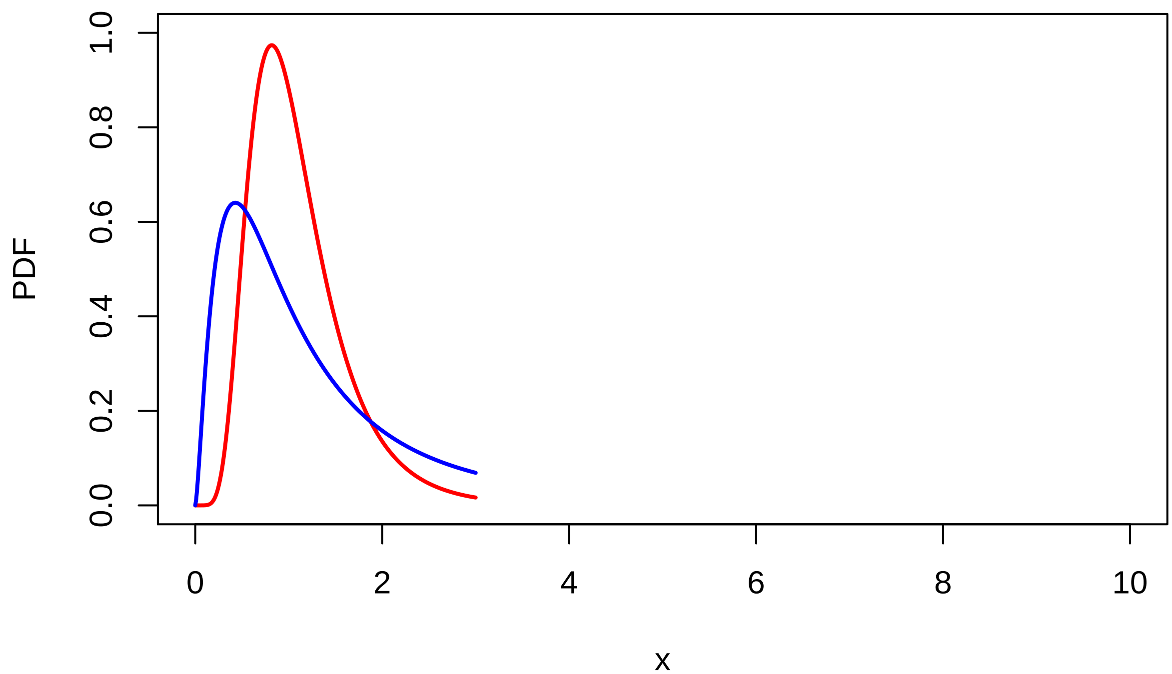
```r
qchisq(0.99, 5)
```

```
## [1] 15.08627
```

2e) Plot the density functions of $F_{20,20}$ and $F_{5,5}$ distributions. (10 points)

Solution:

```r
x <- seq(0, 3, 0.01)
d20.20 <- df(x, 20, 20)
d5.5 <- df(x, 5, 5)
plot(x, d20.20, type = "l", lwd = 2, col = 2, xlab = "x", ylab = "PDF", ylim = c(0,1), xlim = c(0, 10))
lines(x, d5.5,lwd = 2, col = 4)
```

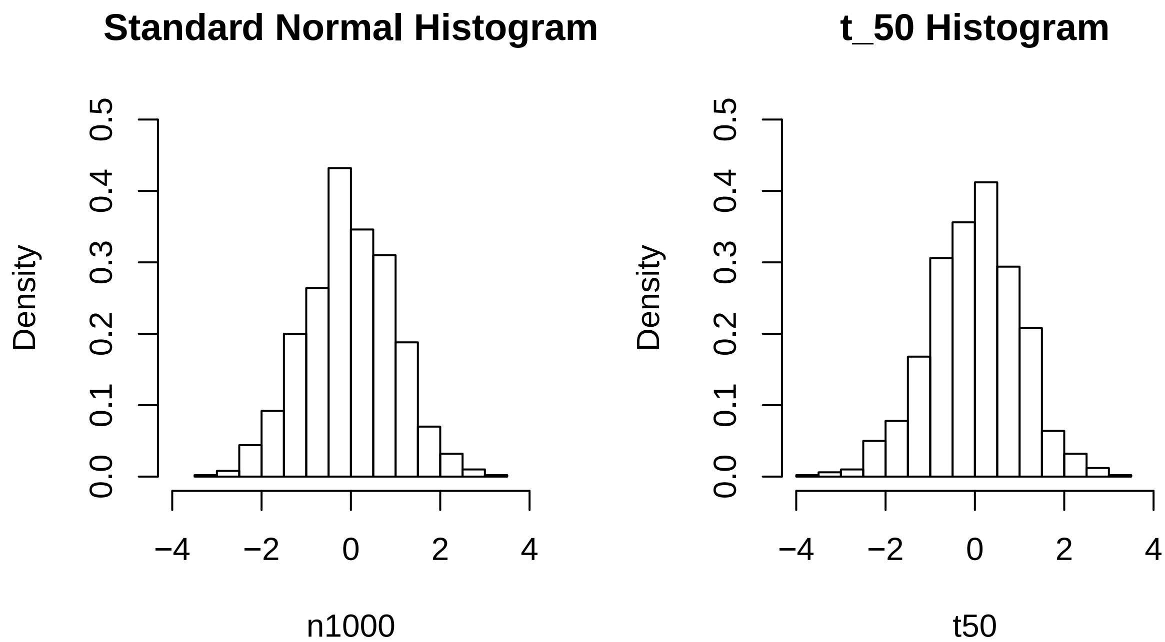


2f) Generate 1000 observations from the standard normal distribution and the t_{50} distribution, respectively. Create a histogram for each of the two samples. Is there any obvious difference between the two histograms. (10 points)

Solution:

```
n1000 <- rnorm(1000)
t50 <- rt(1000, 50)

par(mfrow = c(1,2), pty = "s")
hist(n1000, xlim = c(-4, 4), ylim = c(0, 0.5), freq = F, main = "Standard Normal Histogram")
hist(t50, xlim = c(-4, 4), ylim = c(0, 0.5), freq = F, main = "t_50 Histogram")
```



No, there is no obvious difference between the two histograms.

2g) Find the 90% quantile of the t_{10} distribution. (2 points)

Solution:

```
qt(0.90, 10)
```

```
## [1] 1.372184
```

2h) Find $P(t_{15} > 3)$ (3 points)

Solution:

```
1 - pt(3, 15)
```

```
## [1] 0.004486369
```

2i) Find $P(X > 15)$ where `X ~ Binomial(20, 0.8)` (3 points)

Solution:

Since `pbinom()` calculates $P(X \leq 15)$, and $P(X > 15)$ is analogous to asking: $1 - P(X \leq 15)$, we can do `1 - pbinom(15, 20, 0.8)`.

```
1 - pbinom(15, 20, 0.8)
```

```
## [1] 0.6296483
```

Problem 3: Sampling Distributions (25 points)

Part a:

3ai) Generate 1000 samples of each of three sample sizes: n = 1, n = 10, and n = 30. Calculate the sample means of the 1000 samples of each sample size. (3 points)

Solution:

```
m <- 1000 #number of samples

#sample size 1
# x_bar1 <- rnorm(m, 100, 15) #N(100, 15)
#
# n <- 10 #Sample size 10
# x_bar10 <- rep(0, m)
# for(i in 1:m){x_bar10[i] <- mean(rnorm(n, 100, 15))}
#
# n <- 30 #Sample size 30
# x_bar30 <- rep(0, m)
# for(i in 1:m){x_bar30[i] <- mean(rnorm(n, 100, 15))}

#Now store all x_bar into a single vector:
x_bar <- matrix(0, 3, m)

n <- c(1, 10, 30) #Sample sizes are 1, 10, 30

for(i in 1:3) #Fill each of n = 1, n = 10, and n = 30 up
{
  for(j in 1:m) #Fill each with 1000 samples
  {
    #Calculate sample mean for 1000 samples of size 1, 10, 30
    x_bar[i, j] <- mean(rnorm(n[i], 100, 15))
  }
}
```

3aii) Calculate the mean of sample means and standard deviation of sample means for each sample size. Do you think the two basic properties of $\overline{X}$ are true? Why? Show your evidence. (6 points)

Solution:

```
apply(x_bar, 1, mean) #calculate mean
```

```
## [1] 100.5519 100.0196 100.0775
```

```
rep(100, 4) #compare to theoretical mean
```

```
## [1] 100 100 100 100
```

```
apply(x_bar, 1, sd) #calculate standard error
```

```
## [1] 14.797549  4.699681  2.732785
```

```
15/sqrt(n) #compare to theoretical standard error
```

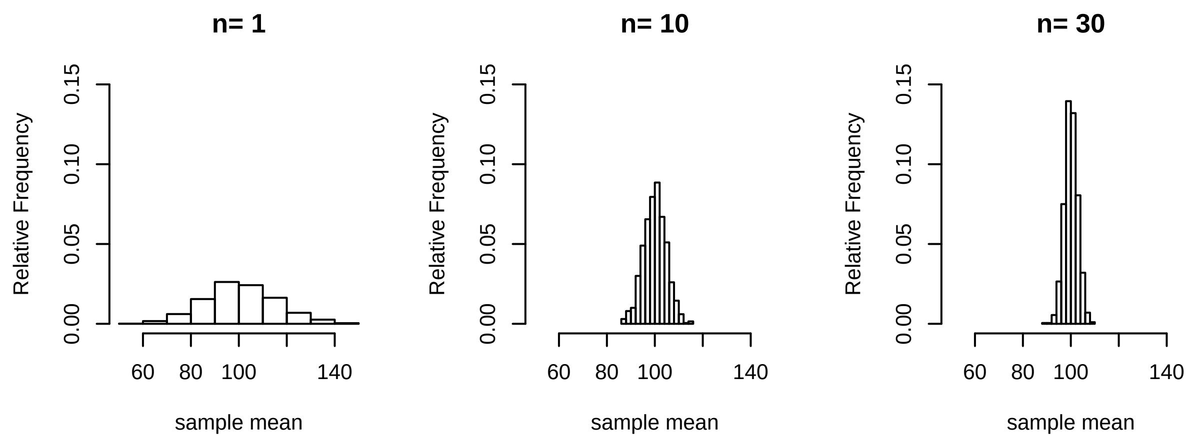
```
## [1] 15.000000  4.743416  2.738613
```

Yes, the two basic properties of $\overline{X}$ do hold. Notice how the simulated mean and standard error are fairly close to the theoretical mean and standard error. In particular, observe how the sample mean of means, gets closer to the theoretical mean, 100, as the sample size increases, in other words, the differences between the theoretical mean and simulated mean become smaller as the sample size increases. Additionally, observe that in regard to the standard error of the mean, the simulated standard error, when compared vertically to the theoretical standard error, are very close to each other. Thus, the two basic properties of $\overline{X}$ do appear to hold.

3aiii) Create a histogram of the sample mneans for each sample size. (4 points)

Solution:

```
par(mfrow = c(1,3), pty = "s")
for(i in 1:3)
{
  #Plot each of the respective histograms
  hist(x_bar[i,],ylab="Relative Frequency",freq=F,
       main=paste("n=",n[i]),xlab="sample mean", xlim = c(50, 150), ylim = c(0, 0.15))
}
```



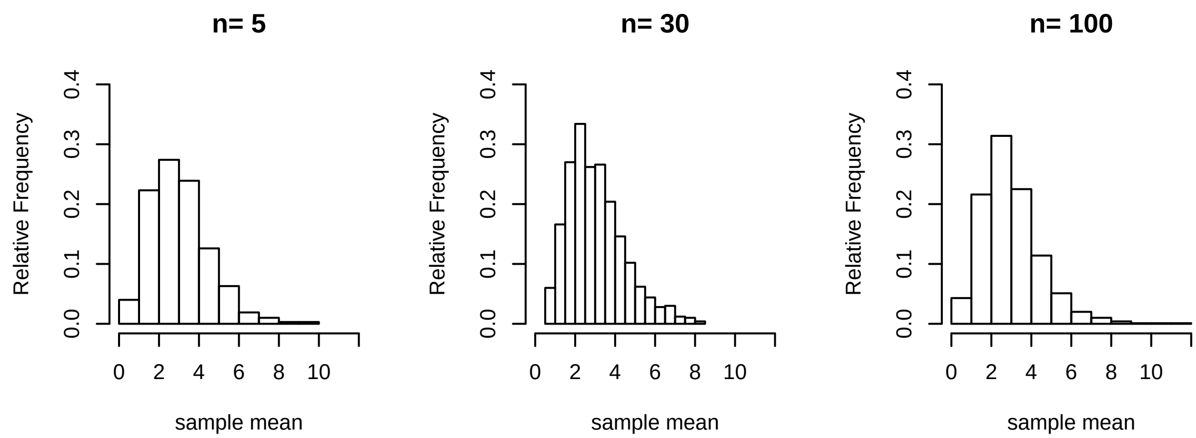
Part b: This part is to verify the Central Limit Theorem.

3bi/3bii/3biii) Simulate 1000 random samples from the χ^2_3 distribution with 3 sample sizes (5, 30, 100) (2 points), Calculate sample means (2 points), and Create histograms of the sample means (8 points).

Solution:

```r
par(mfrow = c(1,3), pty = "s")
m <- 1000
x_bar <- matrix(0, 3, m)
n <- c(5, 30, 100)

for(i in 1:3)
{
  for(j in 1:m)
  {
    x_bar[i,j] = mean(rchisq(n, 3))
  }
  hist(x_bar[i,], freq = F, ylab = "Relative Frequency", main = paste("n=", n[i]),
       xlab = "sample mean", xlim = c(0, 12), ylim = c(0, 0.40))
}
```



Problem 4: Numerical Summary and Graphical Display of One Numerical Variable (25 points)

4a) Download the data set (call it `iq`) and show the top 5 rows of it. (2 points)

Solution:

```
iq <- read.table("C:/Users/gordo/Desktop/iq.txt", header = TRUE) #read in iq.txt
head(iq, 5)
```

```
##    IQ
## 1 111
## 2 107
## 3 100
## 4 107
## 5 115
```

4b) Calculate the mean, standard deviation, and variance of IQ. (3 points)

Solution:

```r
mean(iq$IQ)
```

```
## [1] 108.9231
```

```r
sd(iq$IQ)
```

```
## [1] 13.17097
```

```r
var(iq$IQ)
```

```
## [1] 173.4745
```

4c) Calculate the five-number summary of IQ. (2 points)

Solution:

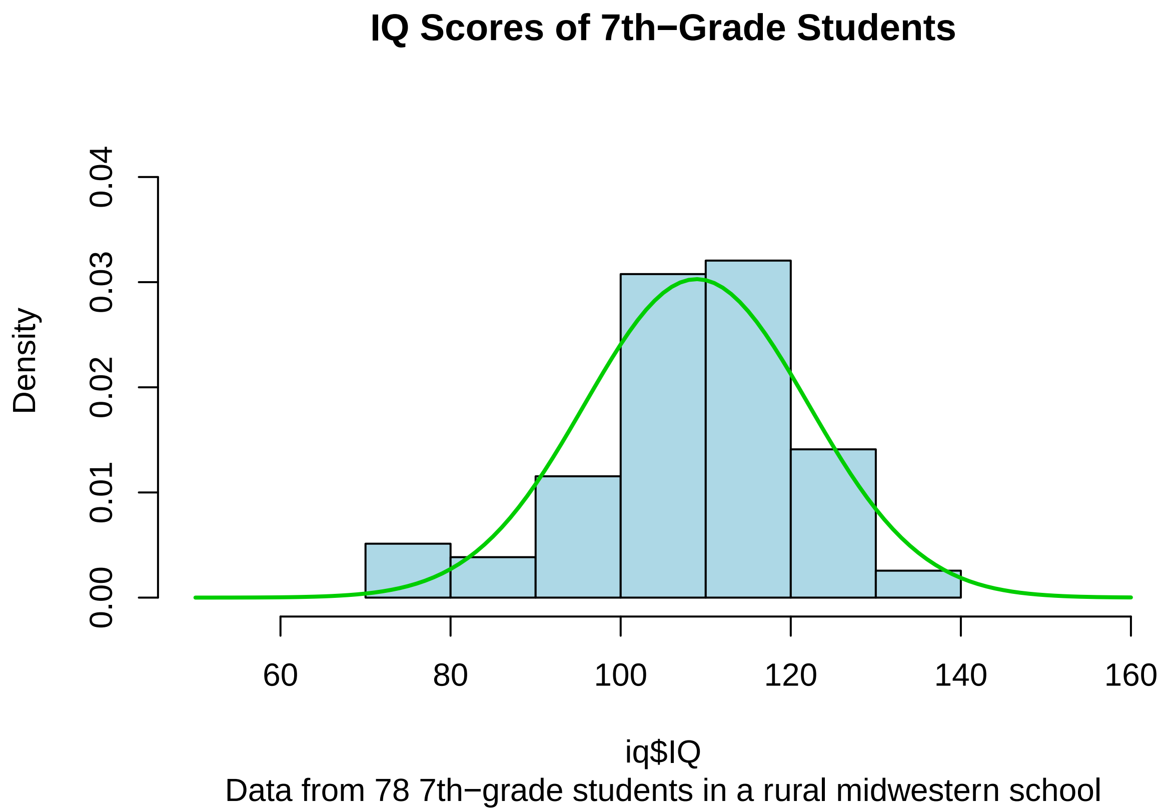
```r
summary(iq$IQ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    72.0   103.0   110.0   108.9   117.5   136.0
```

4d) Create a histogram of `IQ`. Superimpose a normal density curve on it. (10 points)

Solution:

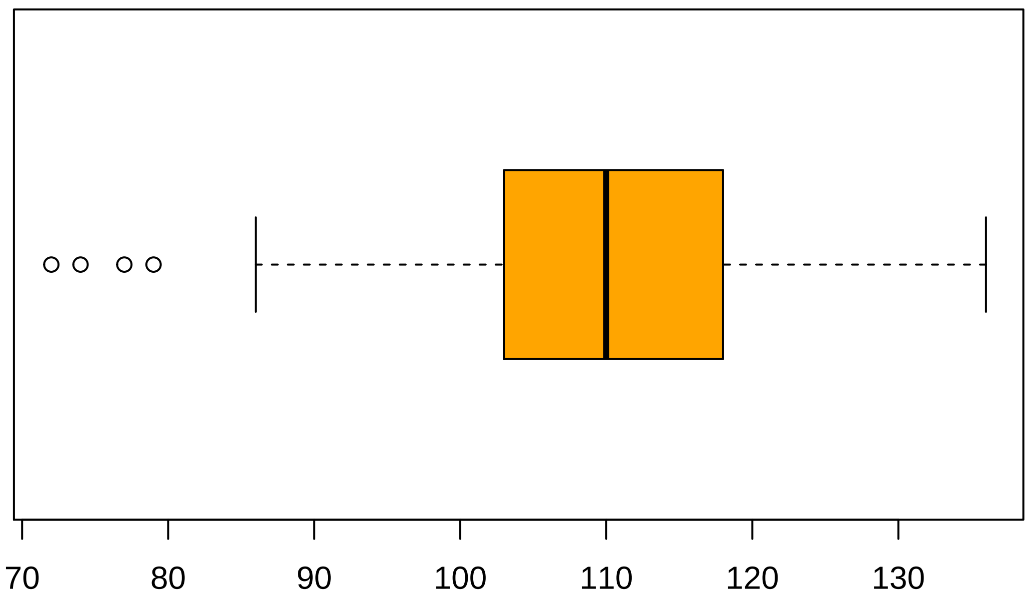
```r
hist(iq$IQ, col = "lightblue", main = "IQ Scores of 7th-Grade Students",
     sub = "Data from 78 7th-grade students in a rural midwestern school",
     freq = F, xlim = c(50, 160), ylim = c(0, 0.045))
y = seq(50, 160)
lines(y, dnorm(y, mean(iq$IQ), sd(iq$IQ)), col = 3, lwd = 2) #overlay density curve
```



4e) Create a boxplot of `IQ`. (5 points)

Solution:

```
boxplot(iq$IQ, horizontal = TRUE, col = "orange", outline = TRUE)
```



4f) Create a stem-and-leaf plot of IQ and specify the four outliers in the boxplot. (3 points)

Solution:

```
stem(iq$IQ)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    7 | 2479
##    8 | 69
##    9 | 01336778
##   10 | 0022333344555666777789
##   11 | 0000111122222333444455688999
##   12 | 003344677888
##   13 | 026
```

The outliers of the data are 72, 74, 77, and 79. In particular any value that is less than 81.25 is consider a lower outlier, and anything above 139.25 is also considered an outlier.

Problem 5: Numerical Summary and Graphical Display of Grouped Numerical Data (25 points)

5a) Download the data set `mitt2` and randomly sample 10 rows to view. (3 points)

Solution:

```
mitt2 <- read.table("C:/Users/gordo/Desktop/mitt2.txt", header = TRUE) #read in mitt2.txt
attach(mitt2)
```

```
## The following object is masked from package:datasets:
##
##     attitude
```

```
mitt2[sample(nrow(mitt2), 10), ]
```

```
##    trustworthiness attitude
## 9              3.9        1
## 58             5.4       -1
## 56             3.3       -1
## 47             3.3       -1
## 15             3.9        1
## 5              3.4        1
## 54             4.1       -1
## 37             4.2       -1
## 22             3.9        1
## 33             4.4       -1
```

5b) Convert the `attitude` variable to a factor-type vector and use labels "positive" and "negative". Use the `table()` function to summarize the factor variable. (6 points)

Solution:

```
detach(mitt2)
mitt2 <- transform(mitt2, attitude = factor(attitude, labels = c("positive", "negative")))
attach(mitt2)
```

```
## The following object is masked from package:datasets:
##
##     attitude
```

```
table(mitt2$attitude) #summarize attitude variable
```

```
##
## positive negative
##       29       29
```

5c) Compare the five-number summaries and the means and standard deviations of the two groups. (10 points)

Solution:

```r
by(mitt2, mitt2$attitude, summary)
```

```
## mitt2$attitude: positive
##  trustworthiness      attitude
##  Min.   :1.70    positive:29
##  1st Qu.:3.30    negative: 0
##  Median :3.80
##  Mean   :3.61
##  3rd Qu.:4.30
##  Max.   :5.40
## ------------------------------------------------------------
## mitt2$attitude: negative
##  trustworthiness      attitude
##  Min.   :2.600   positive: 0
##  1st Qu.:3.300   negative:29
##  Median :3.900
##  Mean   :3.917
##  3rd Qu.:4.500
##  Max.   :5.700
```

```r
mitt2.positive <- mitt2$trustworthiness[mitt2$attitude == "positive"]
mitt2.negative <- mitt2$trustworthiness[mitt2$attitude == "negative"]
#compute respective sd's.
sd(mitt2.positive)
```

```
## [1] 0.912745
```

```r
sd(mitt2.negative)
```

```
## [1] 0.7960029
```

Which group has higher ratings in general? Which group has a larger range and interquartile range ($Q_3 - Q_1$)? How about standard deviation? It appears that the `Negative` group tends to have higher ratings. In regards to range, it appears that the `Positive` group has a larger range with `3.7`, while the `Negative` group has a lower range with `3.1`. In regards to standard deviation, the `Positive` group appears to have a larger standard deviation than the `Negative` group.

Mean and Median: The `Negative` group has a higher mean with `3.917`, while `Positive` group has lower mean with `3.61`. In regard to median, the `Negative` group has a higher median with `3.900`, while `Positive` group has a lower median with `3.80`.

Range: The `Positive` group has a larger range with `3.7`, while the `Negative` group has a lower range with `3.1`.

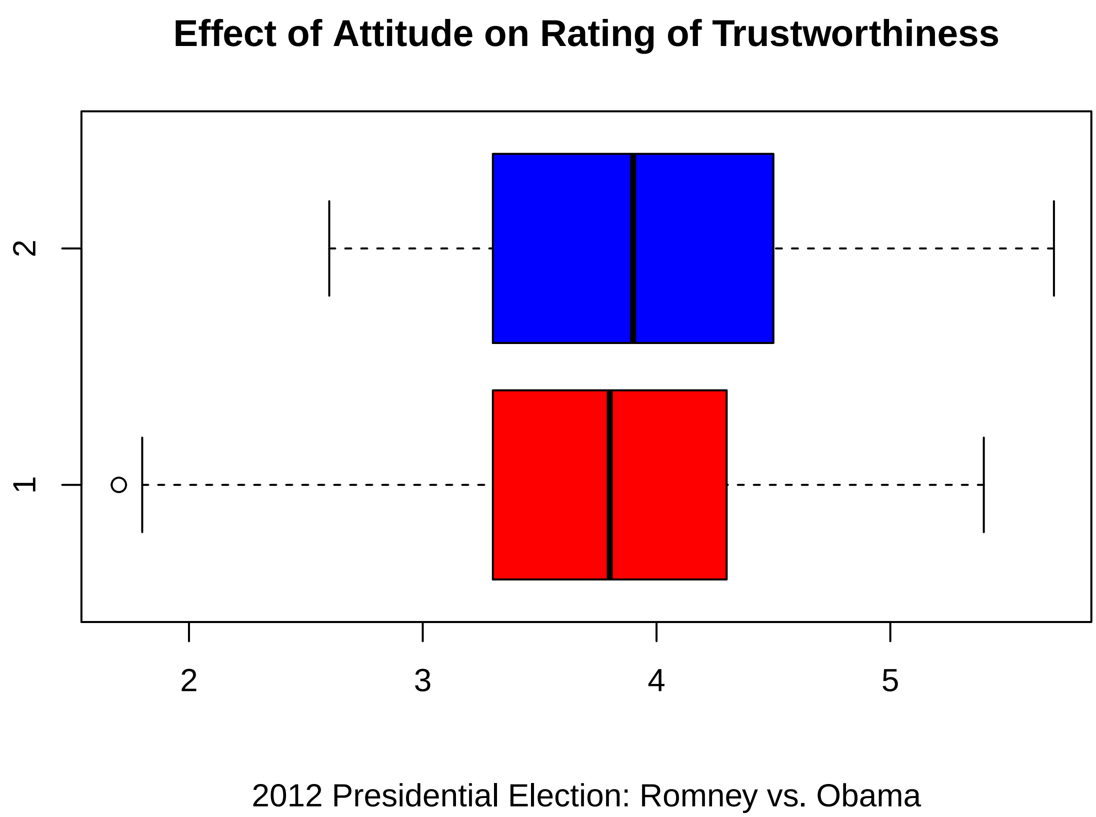
Interquartile range: The `Negative` group has a higher IQR with `1.2`, while the `Positive` group has a lower IQR with `1`.

Standard devaiation: The `Positive` group has a higher standard deviation with `0.912745`, while the `Negative` group has a lower standard deviation with `0.7960029`.

5d) Create a side-by-side boxplot. (6 points)

Solution:

```
boxplot(mitt2.positive, mitt2.negative, horizontal = TRUE, outline = TRUE, col = c(2,4))
title(main = "Effect of Attitude on Rating of Trustworthiness",
      sub = "2012 Presidential Election: Romney vs. Obama")
```




```
detach(mitt2)
```

Problem 6: Numerical Summary and Graphical Display of One Categorical Variable (25 points)

6a) Create a bar graph for the age distribution of Facebook users. Do the same for Twitter and LinkedIn. (11 points)

Solution:

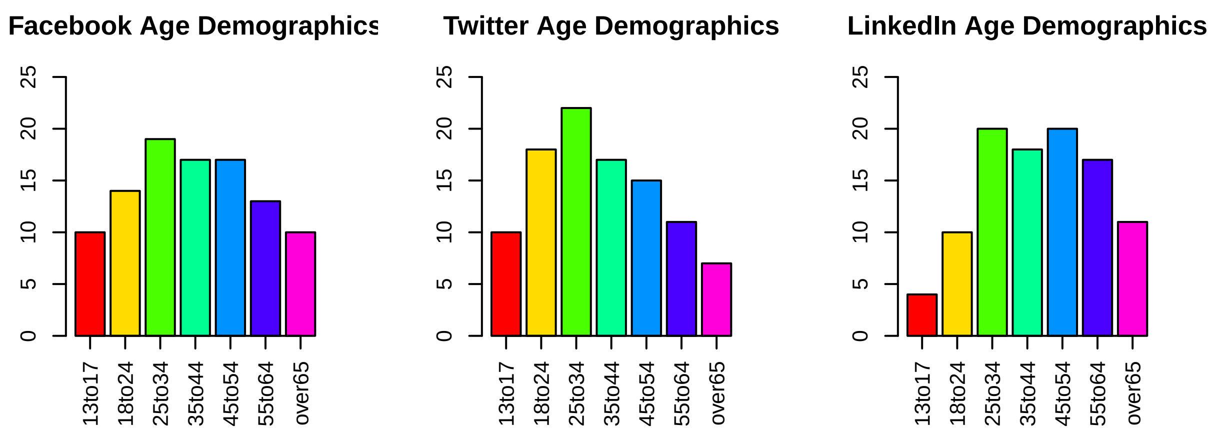
```
socialnt <- read.table("C:/Users/gordo/Desktop/socialnt.txt", header = TRUE) #read in socialnt.txt

par(mfrow = c(1,3), pty = "s")
```

```
facebook <- barplot(socialnt$FacebookPct, col = rainbow(7), ylim = c(0,25),
                    main = "Facebook Age Demographics") #set percent axis lim using ylim
axis(side = 1, labels = socialnt$Age, at = facebook, las = 2)

twitter <- barplot(socialnt$TwitterPct, col = rainbow(7), ylim = c(0,25),
                    main = "Twitter Age Demographics") #set percent axis lim using ylim
axis(side = 1, labels = socialnt$Age, at = twitter, las = 2)

linkedin <- barplot(socialnt$LinkedInPct, col = rainbow(7), ylim = c(0,25),
                    main = "LinkedIn Age Demographics") #set percent axis lim using ylim
axis(side = 1, labels = socialnt$Age, at = linkedin, las = 2)
```



6b) Compare the three barplots created in part (a). Which social networking site has relatively the youngest users in general? Which age the oldest? (4 points)

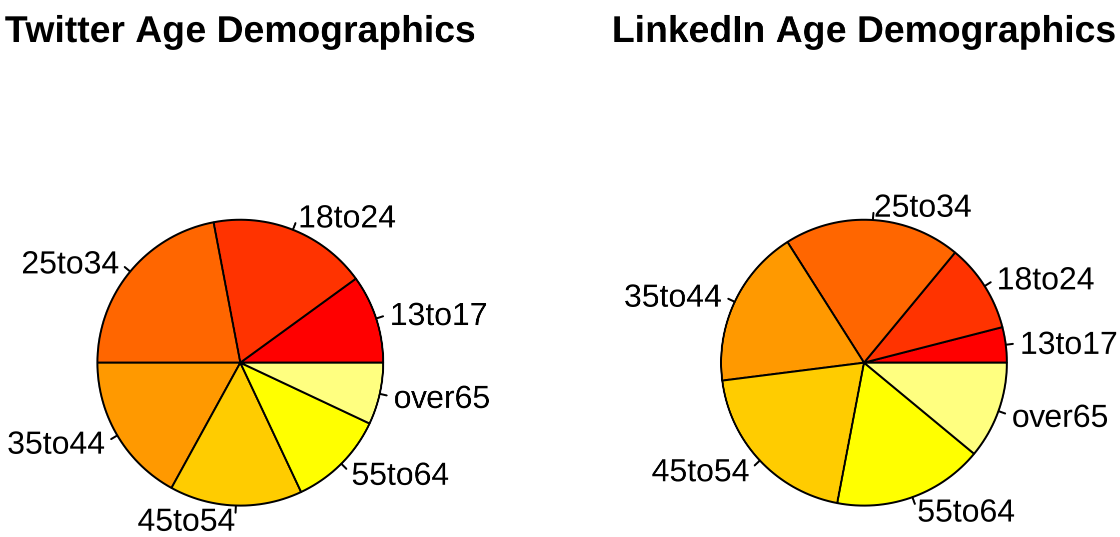
Solution:

Based on the above plots, it appears that `Twitter` has relatively more young users. If we consider the first 3 categories (13 to 17), to (25 to 34), there are fairly more young `Twitter` users than young `Facebook` users. Additionally, in regard, it appears that `LinkedIn` has relatively the oldest users in general. If we consider the last 3 categories (45 to 54) to (over 65), LinkedIn has more older users than `Facebook` and `Twitter` do. Alternatively, we can reach the same conclusion by comparing the medians of each respective bar plot. For `LinkedIn`, it is apparent that the median is higher than that of Twitter and `Facebook's` median. As for `Twitter` and `Facebook`, the medians are approximately equal, however, by looking at the lower halves of the bar plots, it is evident that Twitter has more users captured than `Facebook` in the lower half, thus `Twitter` has relatively the youngest users in general.

6c) Create a pie chart for the age distribution of Twitter and LinkedIn. (10 points)

Solution:

```
par(mfrow = c(1,2))
pie(socialnt$TwitterPct, socialnt$Age, col = heat.colors(length(socialnt$Age)),
    main = "Twitter Age Demographics", clockwise = F)
pie(socialnt$LinkedInPct, socialnt$Age, col = heat.colors(length(socialnt$Age)),
    main = "LinkedIn Age Demographics", clockwise = F)
```



Is it easy to compare the age distributions of the three social networking sites? Why? (2 points) Yes, by using a bar graph, it is far easier to tell that `LinkedIn` has more older users, based on the amount of the data that is captured on the right half of the bar plot. In contrast, it is easy to tell that `Twitter` and `Facebook` have more younger users, by sheer comparison of the left halves of the bar plots to the left half of the bar plot of `LinkedIn` users.

However, it is difficult to tell between `Facebook` and `LinkedIn` which has the youngest users in general, by closely looking at the left-halves of the bar plots, and noting that `Twitter` tends to have more of a percentage captured than `Facebook` users. Additionally, note that by using a pie chart, it is easier to discern between `Twitter` and `Facebook` which has the youngest users.

Problem 7: Numerical Summary and Graphical Display of Two Categorical Variables (25 points)

7a) Create a two-way table based on the table. You may use the `matrix` function or `cbind` or `rbind` function. Make sure you add appropriate row names and column names. (6 points)

Solution:

```
gaming <- read.table("C:/Users/gordo/Desktop/gaming.txt", header = TRUE) #read in gaming.txt
M <- matrix(gaming$Count, nrow = 2, ncol = 3)
colnames(M) <- c("A&B", "C", "D&F")
rownames(M) <- c("Yes", "No")
```

7b) Calculate the marginal tables. (6 points)

Solution:

```
total.row <- margin.table(M, 1)
total.row #Calculate row totals
```

```
##  Yes   No
## 1379  429
```

```
total.col <- margin.table(M, 2)
total.col #Calculate column totals
```

```
## A&B   C D&F
## 941 594 273
```

7c) Create a conditional proportion table to examine the effect of playing games on grades. (5 points)

Solution:

```
#Row-wise, so condition is played games, and response is grade
prop.table(M, 1) #Given played games = ??, what is grade?
```

```
##           A&B         C       D&F
## Yes 0.5337201 0.3263234 0.1399565
## No  0.4778555 0.3356643 0.1864802
```

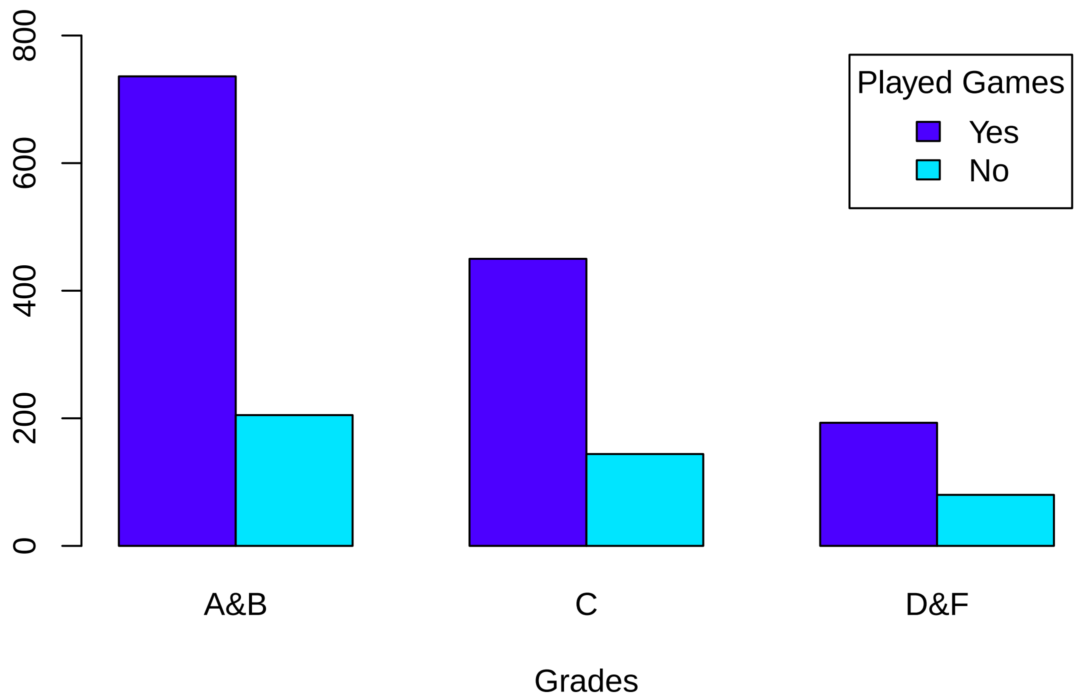
Which is higher? `P("A&B" given "Played games")` or `P("A&B" given "Never played games")`?

It appears that `P("A&B" given "Played games")` is higher with `0.5337201`, while `P("A&B" given "Never played games")` is lower with `0.4778555`.

7d) Create a bargaph to display the relationship between playing games and grades. (8 points)

Solution:

```
barplot(M, beside = TRUE, legend.text = TRUE,
        args.legend= list(title = "Played Games"),
        col = topo.colors(2),
        ylim = c(0, 800), xlab = "Grades")
```

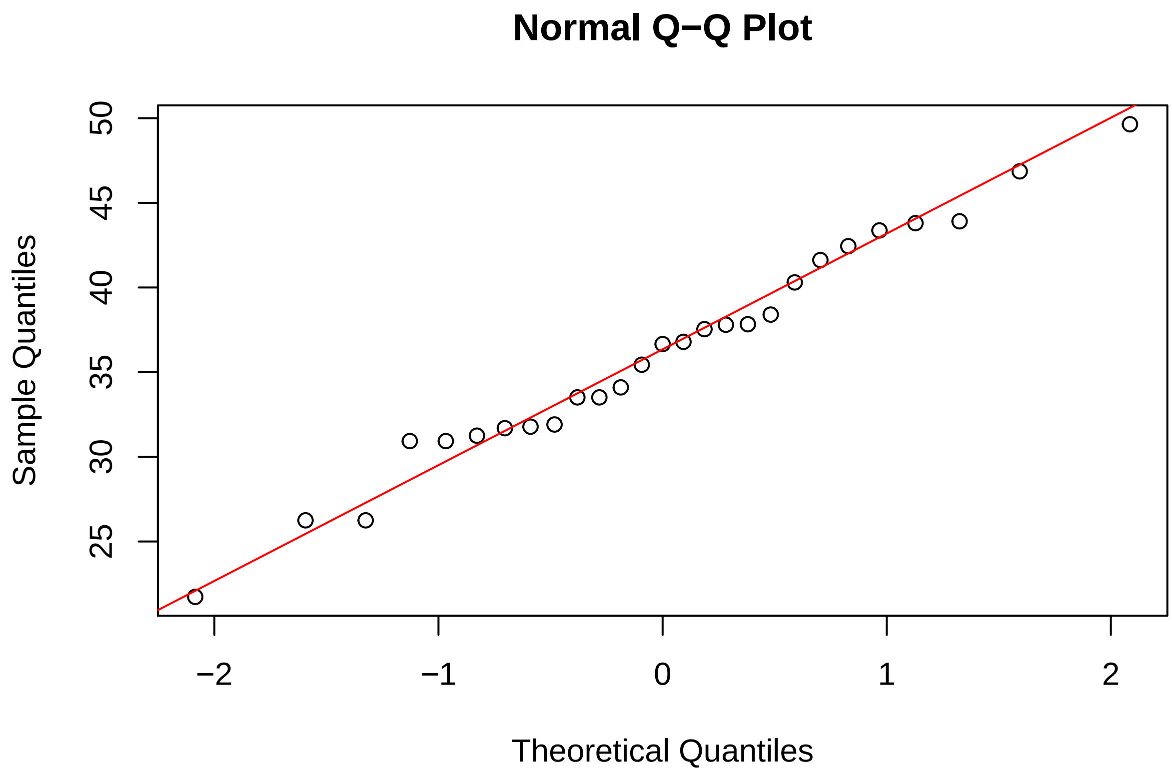


Problem 8: Inference about a single mean (24 points)

8a) Create a Q-Q plot for variable "Earnings" and visually check if the normality assumption is satisfied. Add a red reference (diagonal) line. (4 points)

Solution:

```
uber <- read.table("C:/Users/gordo/Desktop/uber.txt", header = TRUE) #read in uber.txt
attach(uber)
qqnorm(Earnings)
qqline(Earnings, col = 2)
```



Yes, it appears that the normality assumption is satisfied.

8b) Use Shapiro-Wilk test to check normality. (4 points)

Solution:

```
shapiro.test(Earnings)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Earnings
## W = 0.98508, p-value = 0.9547
```

Yes, based on the Shapiro-Wilk normality test, it appears that the normality assumption is satisfied.

8c) Report a 95% confidence interval for the average earnings per hour of New York City Uber drivers. (Use `t.test`) (4 points)

Solution:

```r
t.test(Earnings, conf.level = 0.95)$conf.int
```

```
## [1] 33.55426 38.75981
## attr(,"conf.level")
## [1] 0.95
```

8d) Does the data provide sufficient evidence that the average hourly pay is higher than $30? Use ‘t.test to do the testing. Based on the test result, write down the 4 steps of the hypothesis test. (12 points)

Solution:

```r
t.test(Earnings, alternative = "greater", mu = 30)
```

```
##
##  One Sample t-test
##
## data:  Earnings
## t = 4.8625, df = 26, p-value = 2.415e-05
## alternative hypothesis: true mean is greater than 30
## 95 percent confidence interval:
##  33.99733      Inf
## sample estimates:
## mean of x
##  36.15704
```

```r
detach(uber)
```

4-Step H.T.

Hypothesis: $H_0 : \mu = 30$ vs. $H_a : \mu > 30$

Test statistic: $t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} = \texttt{4.8625}$

P-value: $P(T > t_0) = 1 - P(T < t_0) = \texttt{2.415e-05}$

Conclusion: Reject H_0, since p-value < 0.05.

Problem 9: Matched-Pairs Design (12 points)

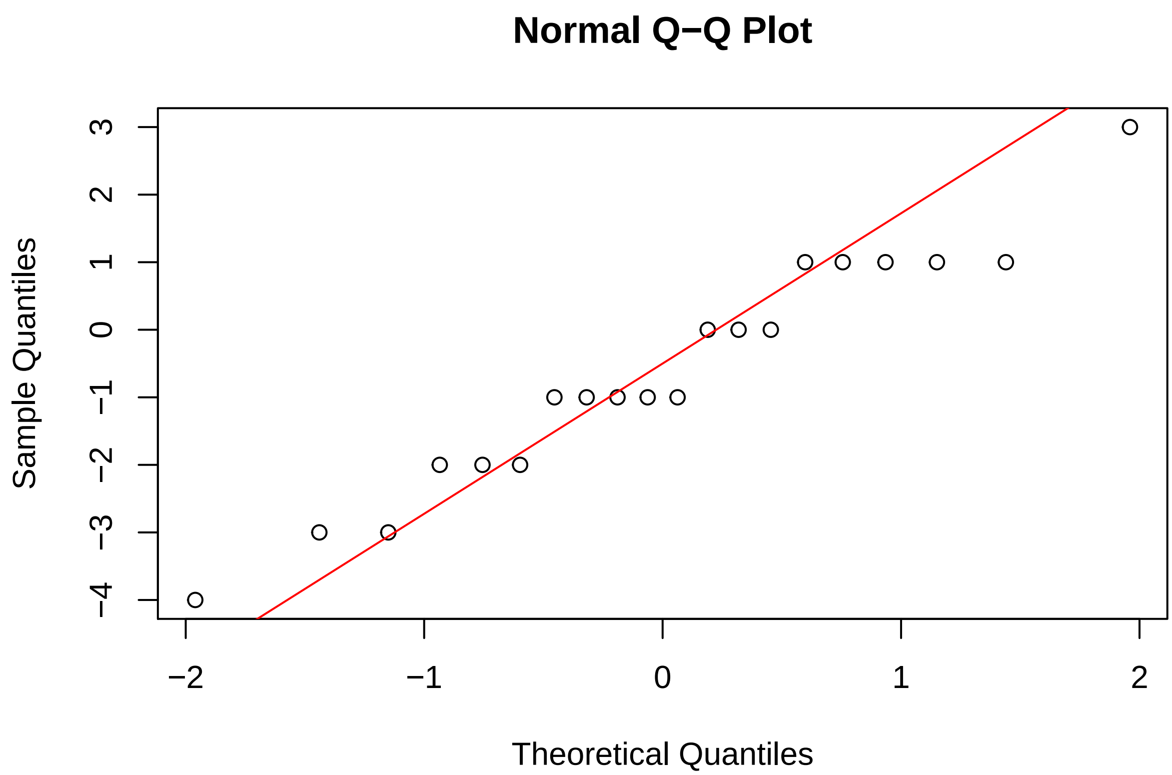
9a) Check normality of the variable `Diff` using a Q-Q plot and the Shapiro-Wilk test. (4 points)

Solution:

```r
equiv <- read.table("C:/Users/gordo/Desktop/equiv.txt", header = TRUE) #read in equiv.txt
attach(equiv)
summary(equiv)
```

```
##     Subject          Paper        Computer          Diff
##  Min.   : 1.00   Min.   :2.0   Min.   : 2.00   Min.   :-4.00
##  1st Qu.: 5.75   1st Qu.:4.0   1st Qu.: 3.75   1st Qu.:-2.00
##  Median :10.50   Median :4.5   Median : 5.00   Median :-1.00
##  Mean   :10.50   Mean   :4.8   Mean   : 5.45   Mean   :-0.65
##  3rd Qu.:15.25   3rd Qu.:6.0   3rd Qu.: 7.00   3rd Qu.: 1.00
##  Max.   :20.00   Max.   :8.0   Max.   :10.00   Max.   : 3.00
```

```r
qqnorm(Diff)
qqline(Diff, col = 2)
```




```r
shapiro.test(Diff)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Diff
## W = 0.96048, p-value = 0.5534
```

9b) Conduct a one-sample t test about the difference, `Diff`. What conclusion can you make? (4 points)

Solution:

```r
t.test(Diff, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  Diff
## t = -1.685, df = 19, p-value = 0.1084
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.4574186  0.1574186
## sample estimates:
## mean of x
##     -0.65
```

Conclusion: Fail to reject H_0, since p-value > 0.05.

9c) Conduct a matched-pairs t test. What conclusion can you make? (4 points)

Solution:

```r
t.test(Paper, Computer, paired = TRUE, conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  Paper and Computer
## t = -1.685, df = 19, p-value = 0.1084
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.4574186  0.1574186
## sample estimates:
## mean of the differences
##                   -0.65
```

```r
detach(equiv)
```

Conclusion: Fail to reject H_0, since p-value > 0.05. The same result as part (b).

Problem 10: Inference about two means (24 points)

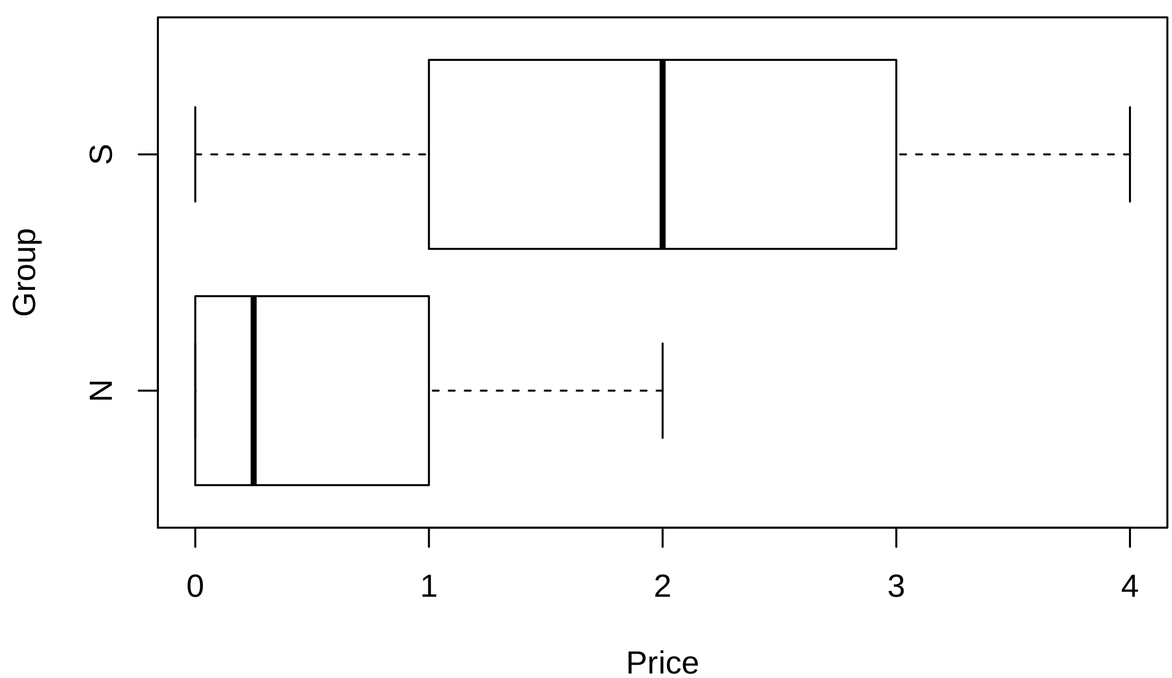
10a) Create a horizontal side-by-side boxplot for the two samples. Do you think normality assumption is roughly satisfied? (4 points)

Solution:

```r
sad <- read.table("C:/Users/gordo/Desktop/sad.txt", header = TRUE) #read in sad.txt
attach(sad)
summary(sad)
```

```
##      Price        Group
##  Min.   :0.000   N:14
##  1st Qu.:0.250   S:17
##  Median :1.000
##  Mean   :1.419
##  3rd Qu.:2.250
##  Max.   :4.000
```

```r
boxplot(Price ~ Group, horizontal = TRUE, outline = TRUE)
```



No, I do not think that the normality assumption is satisfied. Note that the Boxplot for Group `S` looks fairly symmetric, however for Group `N`, the data looks skewed. Based on the Boxplots, it would imply that the data is right-skewed.

10b) State the appropriate null and alternative hypotheses for comparing the two groups. (2 points)

Solution:

Hypothesis: $H_0 : (\mu_S - \mu_N) = 0$ vs. $H_a : (\mu_S - \mu_N) \neq 0$

10c) Perform the significance test at the $\alpha = 0.05$ level, make sure to report the test statistic and the p-value. What is your conclusion? (8 points)

Solution:

```
#on inspection of boxplots, variances are not equal:
Price.N <- Price[Group == "N"]
Price.S <- Price[Group == "S"]
t.test(Price.N, Price.S, conf.level = 0.95, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Price.N and Price.S
## t = -4.3031, df = 26.48, p-value = 0.0002046
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.2841749 -0.8082621
## sample estimates:
## mean of x mean of y
## 0.5714286 2.1176471
```

Conclusion: Reject H_0, since p-value < 0.05.

10d) Use an F test to check equal-variance assumption at $\alpha = 0.05$ level. (4 points)

Solution:

```
var.test(Price.N, Price.S, conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  Price.N and Price.S
## F = 0.34434, num df = 13, denom df = 16, p-value = 0.05873
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1207973 1.0422853
## sample estimates:
## ratio of variances
##          0.3443397
```

10d) Perform the pooled t-test at the $\alpha = 0.05$ level. Compare the test result with that of Part c. Any difference? (6 points)

Solution:

```
t.test(Price.N, Price.S, conf.level = 0.95, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Price.N and Price.S
## t = -4.0982, df = 29, p-value = 0.0003062
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.3178661 -0.7745709
## sample estimates:
## mean of x mean of y
## 0.5714286 2.1176471
```

Conclusion: Reject H_0, since p-value < 0.05. Note that the p-value, t-statistic, and degrees of freedom are different, however we reach the same conclusion as Part c.