BUAN 6359.003 - Advanced Statistics for Data Science - Final Test

Prriyamvradha Parthasarathi - pxp220005

```
library(readxl)
library(tidyr)
library(moments)
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
```

##

Problem 1: Does Confidence Interval work?
We'll generate a population, get a sample from it, create a Confidence Interval using this sample, and then check if our Confidence interval has the Population parameter. Use a 4-digit number (nnnn) of your choice to set the seed using this command: set.seed=nnnn Generate a Normal distribution problem with N = 2,500, Mean =180, and Std dev = 30. Round it off to 1 decimal place Find the mean of this population.Now, from this population, after setting the same seed again, draw a random sample of size n = 30.

```
rm(list=ls()) ;

# Set Seed
set.seed(1600)

# Generating a Normal distribution
ND <- round(rnorm(2500,180,30),1);

# Mean of the Population.
Population_Mean <- mean(ND); cat ("Population_Mean:", Population_Mean )
```

```
## Population_Mean: 180.4429
```

```
# Setting same seed again,
set.seed(1600)

# Random sample of size n = 30.
Sample <- sample(ND,30)

# Mean of the Sample
Sample_Mean <- mean(Sample)

# Standard Deviation of the Sample
```

```r
Sample_Sd <- sd(Sample)

# Standard Error of the Sample
Standard_Error <- Sample_Sd/sqrt(30)

# t-score for a Confidence level of 84.65%.
n=30
t <- qt(0.07675,n-1,lower.tail = FALSE)

# Lower and Upper limits of the Confidence interval
Lower_Tail <- Sample_Mean - t*Standard_Error
Upper_Tail <- Sample_Mean + t*Standard_Error

# If statement to see if the population mean falls within the Confidence interval
if(Population_Mean >= Lower_Tail & Population_Mean <= Upper_Tail)
    {cat("Population Mean is within the Confidence Interval")} else
    {cat("Population Mean is not within the Confidence Interval")}
```

```
## Population Mean is within the Confidence Interval
```

Problem 1 Answers

a. Find the mean and the std error of this sample: **Sample Mean: 185.3433333** and **Standard Error: 5.3748321**

b. Get the proper t-score for a Confidence level of 84.65%: **t-score: 1.4656614**

c. Find the lower and upper limits of the Confidence interval: **Lower limit: 177.4656491 Upper Limit: 193.2210175**

d. Use If statement to see if the population mean falls within the Confidence interval. Get the appropriate output from the R code: **Population Mean is within the Confidence Interval**

###

Problem 2: (Set-1)
Three anti-bacteria creams were used on three age groups.The number of hours before the medicines started to show a noticeable effect are recorded in the table.Assume Alpha = 0.05

```r
rm(list=ls()) ;

# Read Set-1 from the Excel
prob_2<-data.frame(read_excel("F22-6359-Test-3.xlsx", sheet="Set-1"))

# Create individual vectors.   rep command rep("Young",30) will repeat Young 30 times.
v1<-data.frame(Hours = prob_2[, 2], Medicine = prob_2[, 1], Age=rep("Young",30))
v2<-data.frame(Hours = prob_2[, 3], Medicine = prob_2[, 1], Age=rep("Middle_Age",30))
v3<-data.frame(Hours = prob_2[, 4], Medicine = prob_2[, 1], Age=rep("Senior",30))

# Rename columns
names(v1)[1] <- 'Hours'
names(v2)[1] <- names(v1)[1]
names(v3)[1] <- names(v1)[1]

# Combine everything and create a new dataset
data1=rbind(v1, v2, v3)
```

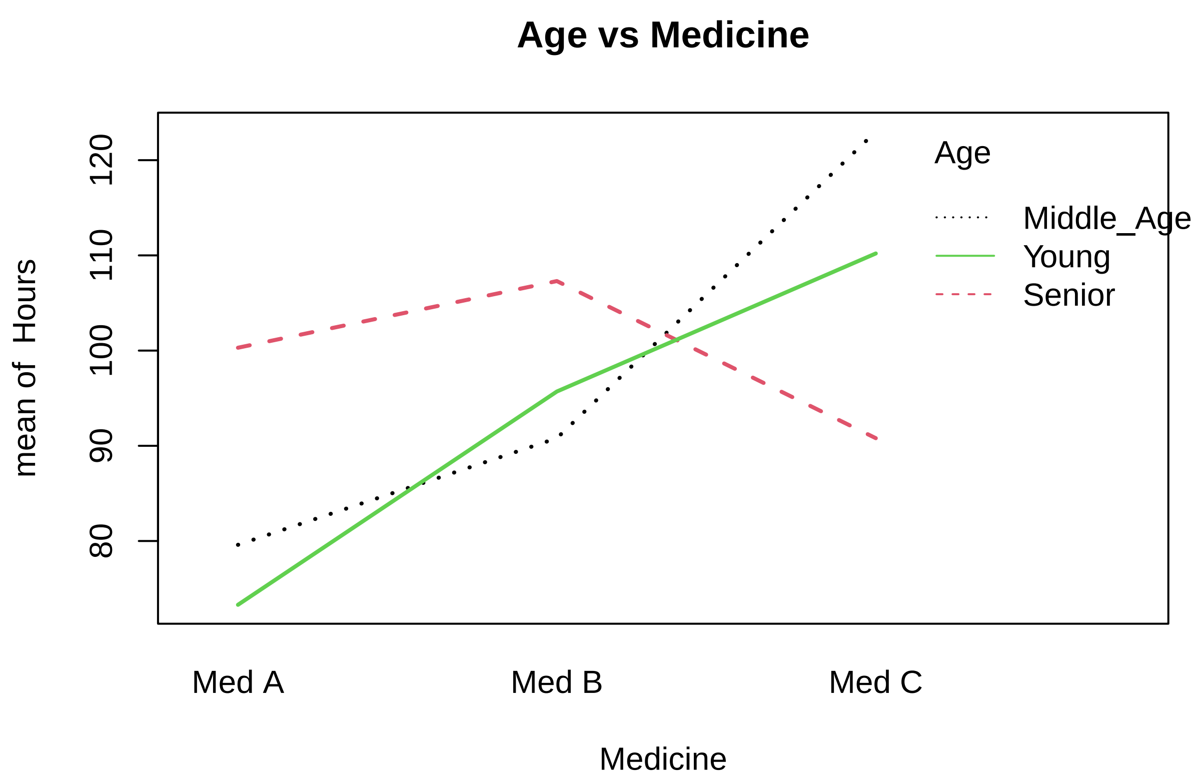
```
attach(data1)

# a.  Run this as an ANOVA 2-factor  R program.
a1 <- aov(Hours~Medicine+Age+Medicine:Age)
summary(a1)
```

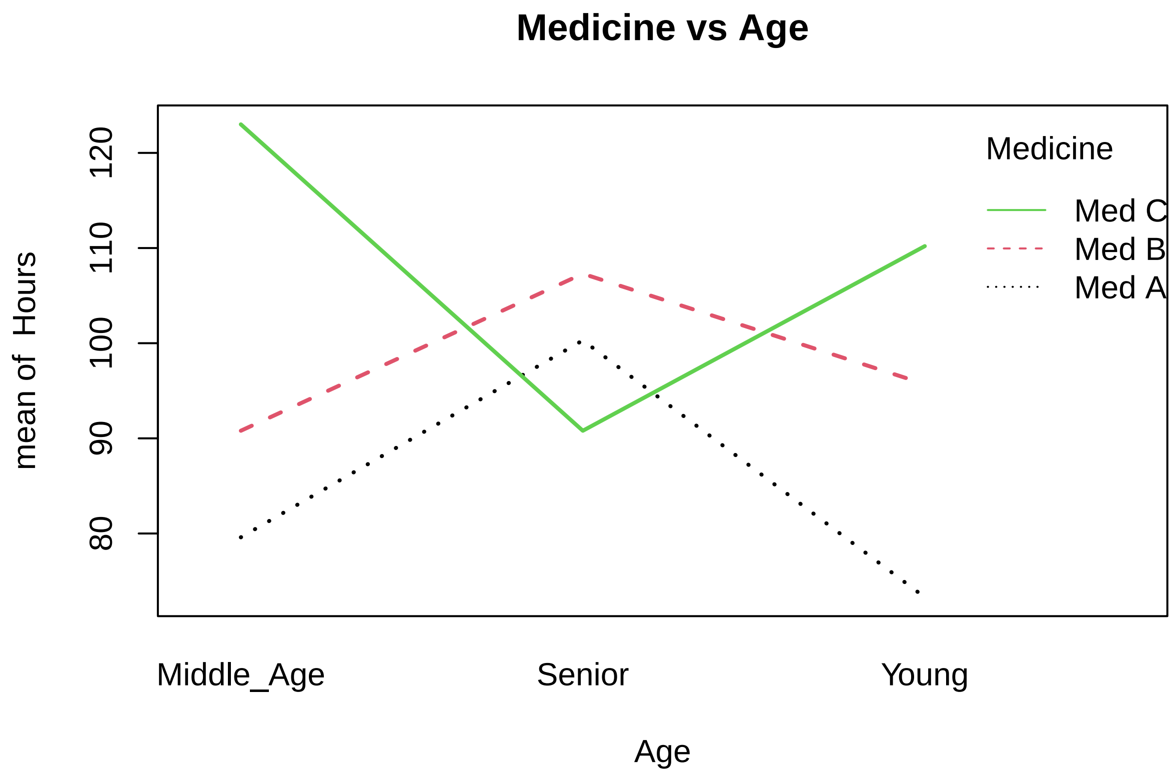
```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Medicine      2   8414    4207   5.950 0.00388 **
## Age           2    661     331   0.468 0.62814
## Medicine:Age  4  10022    2506   3.543 0.01026 *
## Residuals    81  57280     707
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
P_Age <- summary(a1)[[1]][[2,"Pr(>F)"]]
P_Medicine <- summary(a1)[[1]][[1,"Pr(>F)"]]
F_Medicine <- summary(a1)[[1]][[1,"F value"]]

# c.  Also draw the interaction graph to show the interaction between the two factors.
interaction.plot (Medicine, Age, Hours, lwd = 2,  col=1:3, main="Age vs Medicine")
```




```
interaction.plot (Age, Medicine, Hours, lwd = 2,  col=1:3, main="Medicine vs Age")
```



Problem 2 Answers

QUESTION 1: For Set 1, the P-value for Age is: **P-value for Age: 0.6281449**

QUESTION 2: For Set 1, what is the P-value for Medicine?: **P-value for Medicine: 0.0038831**

QUESTION 3: For Set 1, is there an interaction between Medicine and Age: **There is interaction**

QUESTION 4: For Set 1, what is the F-stat for medicine ?: **F-stat: 5.9495318**

QUESTION 5: Does Age have any effect on the number of hours before the medicines start work?: **No**

QUESTION 6: For Set 1, can you say that the medicines behave differently in regards to the time it takes to show an effect?: **Yes because we reject the Null hypothesis**

##

Problem 3: (Set-2) Two sample t-test Automobile Insurance companies consider many factors including the miles driven by a driver and the gender.
The dataset consists of the reported miles (in thousands) driven by young drivers (25 years or less) in the previous year. One insurance company wants to know if there are any difference between the two genders.

```
rm(list=ls()) ;

# Read Set-2 from the Excel
prob_3 <-read_excel("F22-6359-Test-3.xlsx", sheet="Set-2")

# a.  Do a variance test to see if the two variances are equal.
var.test(Distance~Gender, data = prob_3)
```

```
##
##  F test to compare two variances
##
## data:  Distance by Gender
## F = 1.0246, num df = 99, denom df = 99, p-value = 0.904
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6893942 1.5227966
## sample estimates:
## ratio of variances
##           1.024601
```

```
var_test_pval <- var.test(Distance~Gender, data = prob_3)$p.value

# b.  Do the appropriate t-test at Alpha = 5%.
t.test(Distance~Gender, var.equal= TRUE, data = prob_3)
```

```
##
##  Two Sample t-test
##
## data:  Distance by Gender
## t = -1.4193, df = 198, p-value = 0.1574
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -1.3810709  0.2250709
## sample estimates:
## mean in group Female   mean in group Male
##                9.677               10.255
```

```
t_test_pval <- t.test(Distance~Gender, var.equal= TRUE, data = prob_3)$p.value
t_test_tstat <- t.test(Distance~Gender, var.equal= TRUE, data = prob_3)$statistic
```

Problem 3 Answers

QUESTION 7: For Set 2, what is the p-value from the Variance test?: **P-value for Variance test: 0.9040073**

QUESTION 8: For Set 2, after calculating Variance what are you observations?: **Variances are equal**

QUESTION 9: What is the Null Hypothesis for Set-2?: **Mean male = Mean female**

QUESTION 10: For Set 2, what is the p-value for the t-test?: **P-value for t-test: 0.1573739**

QUESTION 11: For Set 2, what is the t-statistics?: **t-statistics: -1.4193343**

QUESTION 12: For Set 2, what decision is made after calculating T-test?: **There is no difference between the male and the female drivers**

###

Problem 4 (Set-3) A bank has collected a sample and is trying to see how various factors impact it's loan approvals. Divide Credit Scores by 10 and incomes by 1000 (in R) and perform Logistics Regression.

```
rm(list=ls()) ;
```

```r
# Read Set-3 from the Excel
prob_4<-read_excel("F22-6359-Test-3.xlsx", sheet = "Set-3")

# Rename the variables
names(prob_4)[1] <- 'Loan.Approved'
names(prob_4)[2] <- 'Credit.scores'
names(prob_4)[3] <- 'Income'
names(prob_4)[4] <- 'Neighborhood.income'

# Divide Credit Scores by 10 and incomes by 1000
prob_4$Credit.scores<- prob_4$Credit.scores/10
prob_4$Income<- prob_4$Income/1000
prob_4$Neighborhood.income<- prob_4$Neighborhood.income/1000

attach(prob_4)

# Logistic Regression
Loan <-glm(Loan.Approved ~ Credit.scores + Income + Neighborhood.income, family="binomial")
summary(Loan)
```

```
##
## Call:
## glm(formula = Loan.Approved ~ Credit.scores + Income + Neighborhood.income,
##     family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5276  -0.9104  -0.6602   1.1599   2.1914
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -9.23814    1.54545  -5.978 2.26e-09 ***
## Credit.scores        0.03141    0.01088   2.888  0.00387 **
## Income               0.04892    0.02028   2.412  0.01589 *
## Neighborhood.income  0.09551    0.02615   3.653  0.00026 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 510.13  on 399  degrees of freedom
## Residual deviance: 468.78  on 396  degrees of freedom
## AIC: 476.78
##
## Number of Fisher Scoring iterations: 4
```

```r
# Maximum likelihood estimates of the parameters
RegOut<-c(coef(Loan))

# Neighborhood with income of 40192
odd40192<-exp(RegOut[1]+RegOut[2]+RegOut[3]+RegOut[4]*40.192)
# Neighborhood income of 46569
Odd46569<-exp(RegOut[1]+RegOut[2]+RegOut[3]+RegOut[4]*46.569)
```

```r
# odds of loan approval for Neighborhood with income of 40192 vs Neighborhood income of 46569
odds_ratio <- odd40192/Odd46569

# Loan approval for a person whose credit score is 728, income is 61653 and lives in a neighborhood whos
odds<-exp(RegOut[1]+RegOut[2]*72.8+RegOut[3]*61.653+RegOut[4]*35.436)
#Probability
probabbility <- odds/(odds+1)

# Loan approval for a person whose credit score is 716, income is 56759 and whose income is 40746
odds1<-exp(RegOut[1]+RegOut[2]*71.6+RegOut[3]*56.759+RegOut[4]*40.746)
```

Problem 4 Answers

QUESTION 13: For Set 3, what are the odds of loan approval for a person who lives in a neighborhood with income of 40192 vs someone with the neighborhood income of 46569 assuming everything else being equal?: **Odds ratio: 0.5438447**

QUESTION 14: For Set 3, what is the probability of loan approval for a person whose credit score is 728, income is 61653 and lives in a neighborhood whose income is 35436?: **Probability: 0.3656949**

QUESTION 15: For Set 3, what are the Odds of loan approval for a person whose credit score is 716, income is 56759 and lives in a neighborhood whose income is 40746: **Odds: 0.7256806**

QUESTION 16: For the Logistics Problem in Set 3, what is the coefficient of Income as calculated by R? : **Coefficient of Income: 0.0489162**

QUESTION 17: For the Logistics Problem in Set 3, what is the coefficient of Credit Scores as calculated by R?: **Coefficient of Credit Scores: 0.0314141**

QUESTION 18: For the Logistics Problem in Set 3, what is the Intercept calculated by R? : **Intercept: -9.2381363**

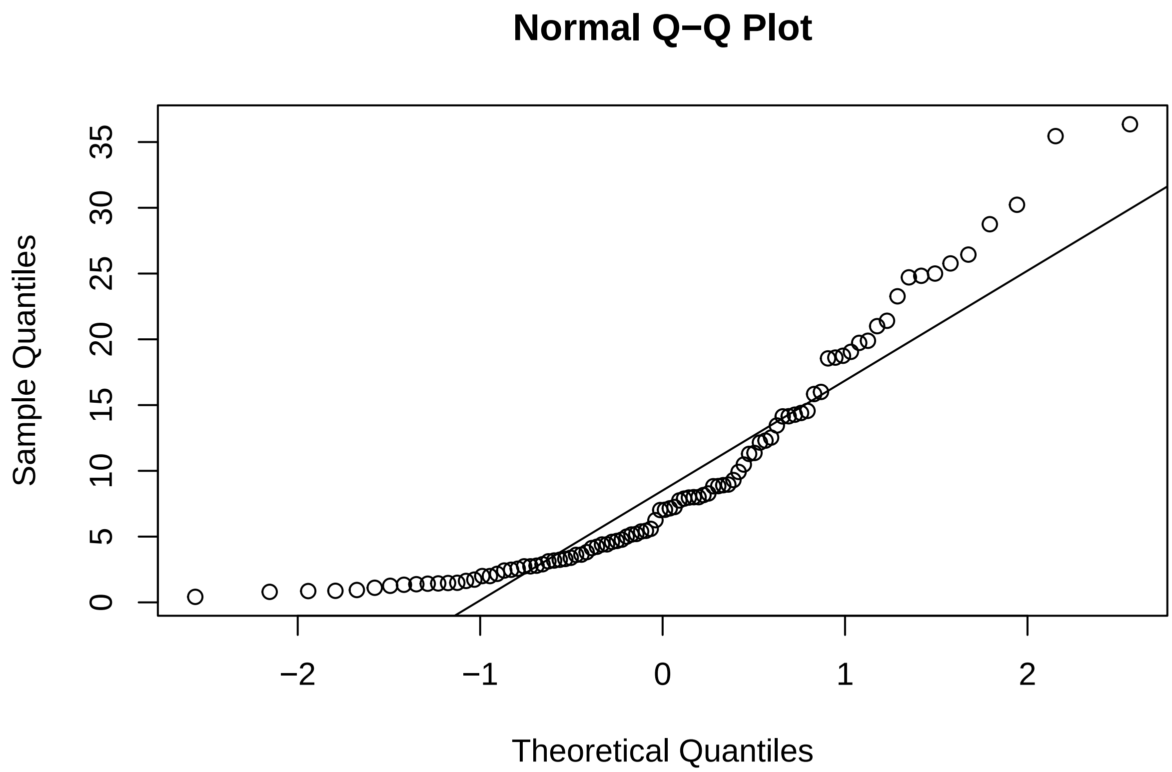
###

Problem 5 (Set-4) You've picked up a bunch of rocks from a rocky beach and want to estimate the weight of all the rocks at the beach with a Confidence level of 93.47%.

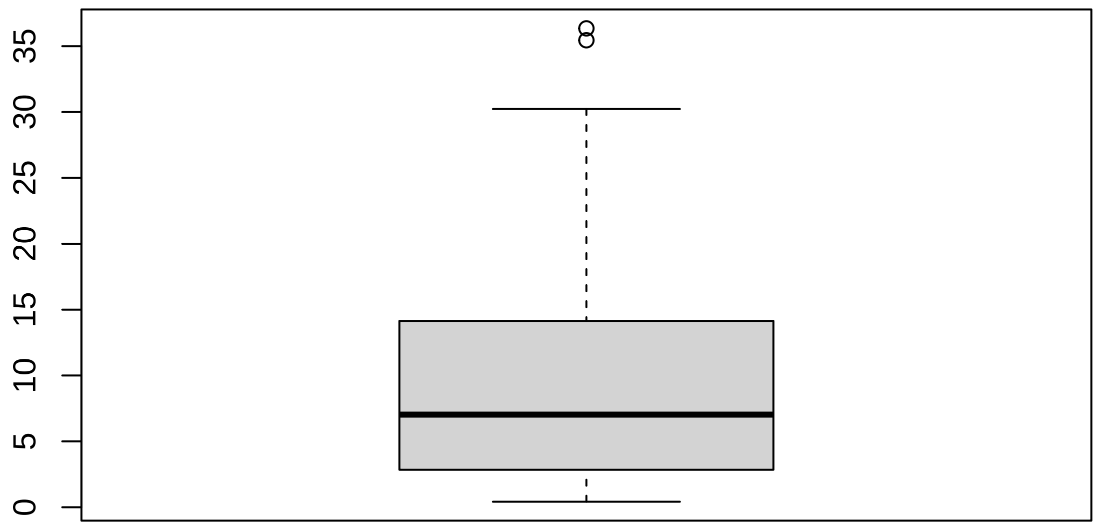
```r
rm(list=ls()) ;

# Read Set-4 from the Excel
prob_5 <-read_excel("F22-6359-Test-3.xlsx", sheet="Set-4")
attach(prob_5)

# a.  Plot the qqline and box plot of the data. Also get the skewness.
qqnorm(Weight)
qqline(Weight)
```




```
boxplot(Weight)
```




```r
# skewness Before log Transformation?
sk1 <- skewness(Weight); cat("Skewness before log Transfprmation:",sk1)
```

```
## Skewness before log Transfprmation: 1.233937
```

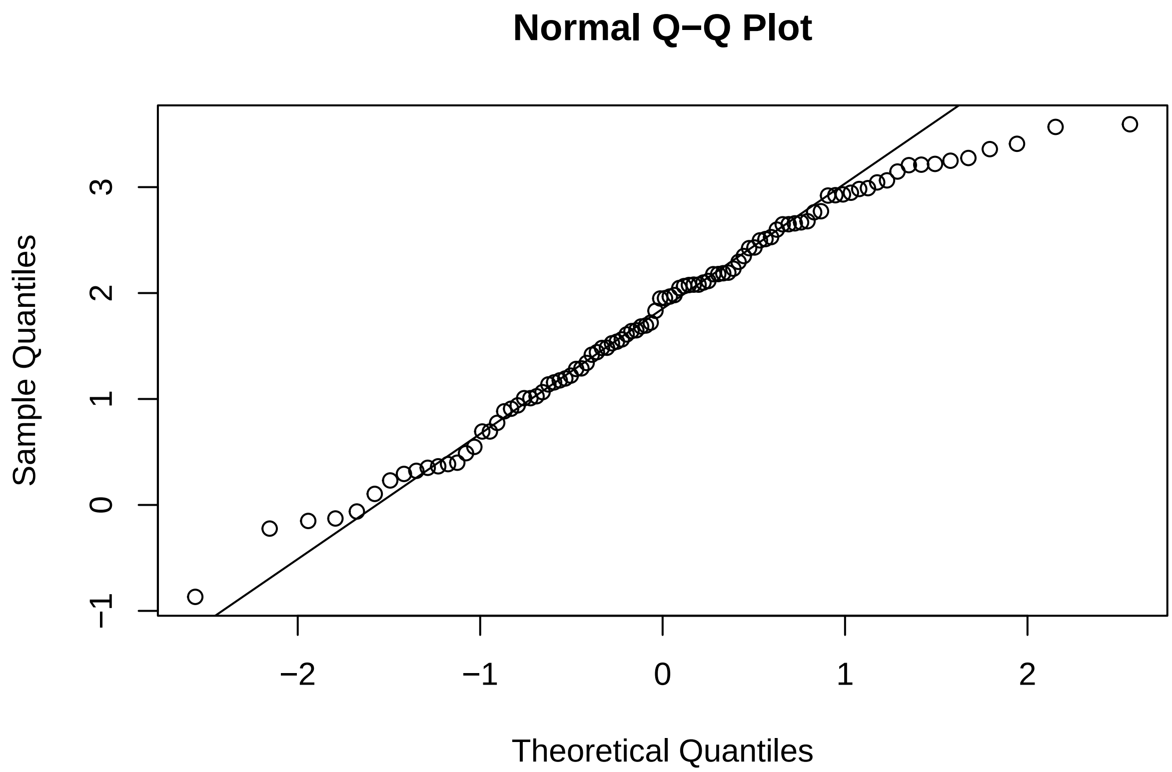
```r
# What is your conclusion about the distribution being normal?
print("The data is nnot Normally Distributed")
```

```
## [1] "The data is nnot Normally Distributed"
```

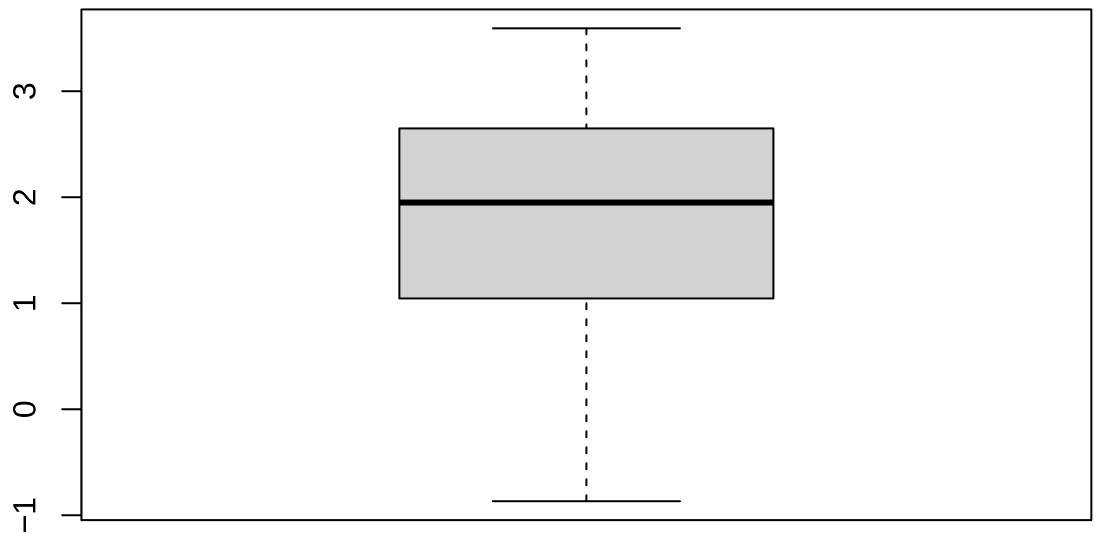
```r
print("Additionally the box plot consists of outliers which is misleading")
```

```
## [1] "Additionally the box plot consists of outliers which is misleading"
```

```r
# b.  Do a log transformation (base e) and perform the steps in a. Use Log transformed data for the fol
log_data <- log(prob_5$Weight,base = exp(1))
qqnorm(log_data)
qqline(log_data)
```




```
boxplot(log_data)
```




```r
sk2 <- skewness(log_data); cat("Skewness after log Transfprmation:",sk2)
```

```
## Skewness after log Transfprmation: -0.2867998
```

```r
# What's your conclusion?
print("After Log Transformation the data has been Normally Distributed")
```

```
## [1] "After Log Transformation the data has been Normally Distributed"
```

```r
print("and from the Box plot we can see that outliers have been removed")
```

```
## [1] "and from the Box plot we can see that outliers have been removed"
```

```r
# c.  What is the mean, Std dev, and the sample size?
mean_prob5 <- mean(log_data); cat("Mean:",mean_prob5)
```

```
## Mean: 1.79195
```

```r
sd_prob5 <- sd(log_data); cat("Standard Deviation:", sd_prob5)
```

```
## Standard Deviation: 1.026756
```

```r
sample_size_prob5 <- length(log_data); cat("Sample Size:", sample_size_prob5)
```

```
## Sample Size: 96
```

```r
# d.  Find std error using the std error formula we've discussed.
std_error_prob5 <- (sd_prob5/sqrt(sample_size_prob5)); cat ("Standard Error:", std_error_prob5)
```

```
## Standard Error: 0.1047928
```

```r
# e.  Find the t-score for the 93.47% confidence interval.
t_score_prob5  <- qt(0.03265,sample_size_prob5-1,lower.tail = FALSE); cat ("t-score:", t_score_prob5)
```

```
## t-score: 1.864775
```

```r
# f.  Use this t-score, sample mean, std error to get the upper and lower limit of the Confidence Inter
Lower_Tail_prob5 <- mean_prob5 - t_score_prob5*std_error_prob5; cat("Lower Tail:", Lower_Tail_prob5)
```

```
## Lower Tail: 1.596535
```

```r
Upper_Tail_prob5 <- mean_prob5 + t_score_prob5*std_error_prob5; cat("Upper Tail:", Upper_Tail_prob5)
```

```
## Upper Tail: 1.987366
```

```r
# g.  Do reverse transformation to get the Confidence Interval in Ounces.
Upper_Tail_rev <- exp(Upper_Tail_prob5); cat("Reverse transformed Upper Tail:", Upper_Tail_rev)
```

```
## Reverse transformed Upper Tail: 7.296286
```

```r
Lower_Tail_rev <- exp(Lower_Tail_prob5); cat("Reverse transformed Lower Tail:", Lower_Tail_rev)
```

```
## Reverse transformed Lower Tail: 4.935902
```

Problem 5 Answers

QUESTION 19: For Set 4, what is the mean of the log-transformed data? : **Mean: 1.7919505**

QUESTION 20: For Set 4, what is the skewness Before log Transformation? : **Skewness: 1.2339372**

QUESTION 21: For Set 4, calculate Skewness after log transformation? : **Skewness: -0.2867998**

QUESTION 22: For Set 4, calculate the standard Deviation after log transformation :**Standard Deviation: 1.0267557**

QUESTION 23: For Set 4, what is the standard error? : **Standard Error: 0.1047928**

QUESTION 24: For Set 4, what is the lower limit of Confidence Interval for a Confidence level of LCL? : **Lower Limit of CI: 1.5965354**

QUESTION 25: For Set 4, calculate Upper Limit after reverse Transformation? : **Upper Limit: 7.2962865**

QUESTION 26: For Set 4, what is the Lower Limit after reverse Transformation? : **Lower Limit: 4.935902**

##

Problem 6 (Set-5) A random sample of 1100 U.S. adults were questioned regarding their political affiliation and opinion on a tax reform bill Perform a test to see if the political affiliation and their opinion on a tax reform bill are independent

```
rm(list=ls()) ;

# Read Set-5 from the Excel
prob_6 <- data.frame(read_excel("F22-6359-Test-3.xlsx", sheet="Set-5"))
```

```
## New names:
## * `` -> `...1`
```

```
# Get ChiSq Stats, P-values, etc. as required by the Online test.
chisq_prob_6 <- prob_6[1:3,2:4]
chisq.test(chisq_prob_6)
```

```
##
##  Pearson's Chi-squared test
##
## data:  chisq_prob_6
## X-squared = 8.9437, df = 4, p-value = 0.06252
```

```
df <- chisq.test(chisq_prob_6)$parameter
Chi_Sq_Critical <- qchisq(p =0.05, df = 4, lower.tail =FALSE)
p_val <- chisq.test(chisq_prob_6)$p.value
```

Problem 6 Answers

QUESTION 27: For Set-5, what are the degrees of Freedom? : **Degrees of Freedom: 4**

QUESTION 28: For Set-5, what is the ChiSq Critical? Assume Alpha = 5% : **ChiSq Critical: 9.487729**

QUESTION 29: For Set-5, what is the P-Value? : **P-value: 0.0625224**

QUESTION 30: For Set-5, what is the correct outcome? Assume Alpha = 5% : **The opinion on the Tax form doesn't depend on the political affliation**