Test - 3

Abhay_Vaidian

2022-12-07

Problem 1 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.

```
## Problem 1:  Does Confidence Interval work?  .

# Use a 4-digit number (nnnn)  of your choice to set the seed using this command: set.seed=nnnn
set.seed(1215)

# Normal distribution problem with N = 2,500, Mean =180,  and Std dev = 30. Round to 1 decimal place
p1_nm <- round(rnorm(2500,180,30),1)

# Find the mean of this population.
p1_mean1 <- mean(p1_nm)
```

Mean of population = 179.76356

```
# Now, from this population, after setting the same seed again, draw a random sample of size n = 30.
set.seed(1215)

p1_s_nm <- sample(p1_nm, size=30)

# a.  Find the mean and the std error of this sample.
p1_s_mean <- mean(p1_s_nm)

p1_s_se <- sd(p1_s_nm)/sqrt(length(p1_s_nm))
```

Mean = 178.7866667
Std error = 4.7730367

```
# b. Get the proper t-score for a Confidence level of 84.65%.
t.test(p1_s_nm,conf.level=.8465)
```

```
##
##  One Sample t-test
##
## data:  p1_s_nm
## t = 37.458, df = 29, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 84.65 percent confidence interval:
##  171.7910 185.7823
```

```
## sample estimates:
## mean of x
##  178.7867
```

```
tscore <- qt(0.07675,df=29,lower.tail = FALSE)
```

T score = 1.4656614

```
# c. Find the lower and upper limits of the Confidence interval.
lowerinterval <- p1_s_mean - tscore*p1_s_se
upperinterval <- p1_s_mean + tscore*p1_s_se
```

Upper limit = 185.7823225
Lower limit = 171.7910109

```
## [1] "Population mean falls within the Confidence interval"
```

Problem 2 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

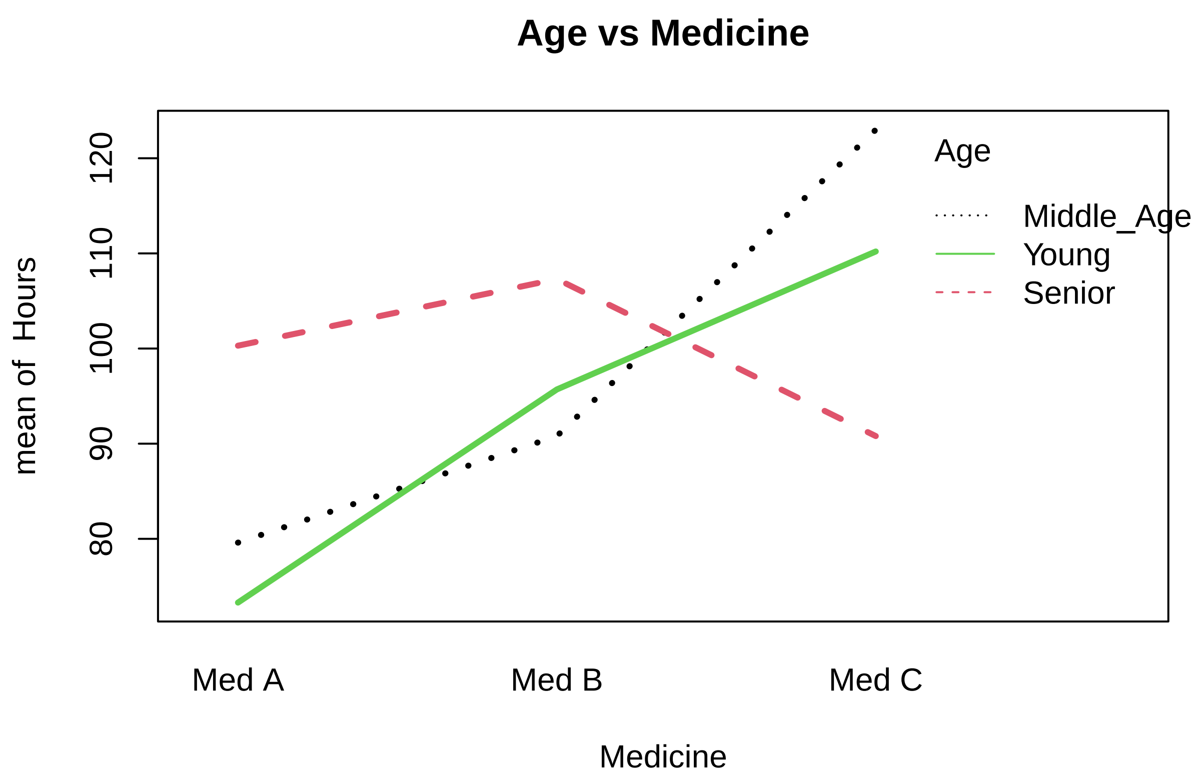
```
## Problem 2 (Set-1)

library(readxl) #import library
table1 <- read_excel("F22-6359-Test-3.xlsx", sheet="Set-1");     # reading excel sheet

# a.  Run this as an ANOVA 2-factor  R program.

# Create individual vectors.   rep command rep("Young",30) will repeat Young 30 times.
v1<-data.frame(Hours = table1[, 2], Medicine=table1[, 1], Age=rep("Young",30))
v2<-data.frame(Hours = table1[, 3], Medicine=table1[, 1], Age=rep("Middle_Age",30))
v3<-data.frame(Hours = table1[, 4], Medicine=table1[, 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);

# run the anova function
a1<-aov(Hours ~ Medicine + Age + Medicine:Age, data = data1)
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
```

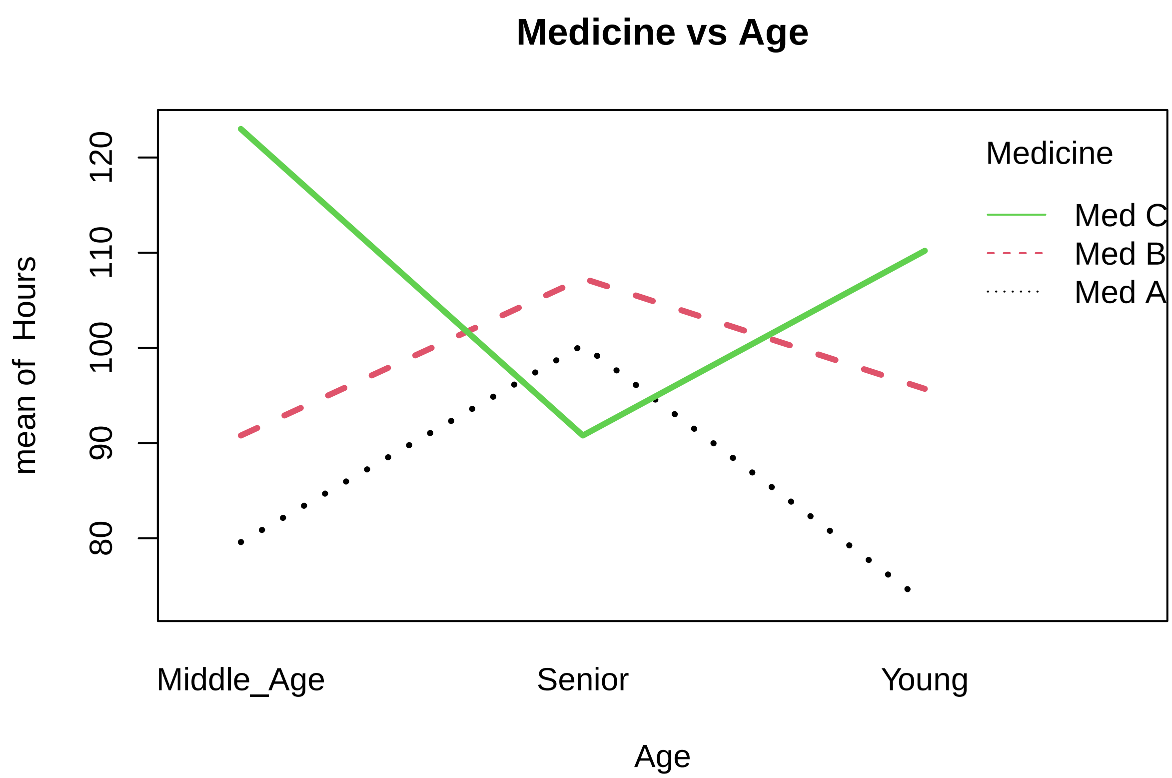
```
# b.  Look at the online test and get the relevant output from your R code (with proper labels)
```

For Set 1, the P-value for Age is 0.62814
For Set 1, the P-value for Medicine is 0.00388
For Set 1, the F-stat for medicine is 5.950

```
# c.  Also draw the interaction graph to show the interaction between the two factors.

# Plot
attach(data1) # attaching data1

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




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




```r
detach(data1) # detaching data1
```

Problem 3 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.

```r
## Problem 3 (Set-2)
## Two sample t-test

# a.  Do a variance test to see if the two variances are equal.

table2 <- read_excel("F22-6359-Test-3.xlsx", sheet="Set-2") # reading excel sheet
var.test(table2$Distance~table2$Gender)
```

```
##
##  F test to compare two variances
##
## data:  table2$Distance by table2$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
```

Variances are approximately equal

```r
#b.  Do the appropriate t-test at alpha = 5%.

t.test(table2$Distance~table2$Gender, var.equal = TRUE, alternative = "two.sided")
```

```
##
##  Two Sample t-test
##
## data:  table2$Distance by table2$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
```

```r
#c.  Look at the online test and get the relevant output from your R code (with proper labels)
```

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

```r
## Problem 4 (Set-3)

table3<-read_excel("F22-6359-Test-3.xlsx", sheet="Set-3") # reading excel sheet

credit_scores <- table3$`Credit scores`/10 # dividing credit score by 10
incomes <- table3$Income/1000 # dividing income by 1000


#Logistic regression

log_reg<- glm(table3$`Loan Approved`~incomes + credit_scores + table3$`Neighborhood income` , family =
summary(log_reg)
```

```
##
## Call:
## glm(formula = table3$`Loan Approved` ~ incomes + credit_scores +
##     table3$`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)                  -9.238e+00  1.545e+00  -5.978 2.26e-09 ***
## incomes                       4.892e-02  2.028e-02   2.412  0.01589 *
## credit_scores                 3.141e-02  1.088e-02   2.888  0.00387 **
## table3$`Neighborhood income`  9.551e-05  2.615e-05   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
```

```
# a.  Look at the online test and get the relevant output from your R code (with proper labels)

RegOut<-c(coef(log_reg));  RegOut # take coeff of estimates
```

```
##                  (Intercept)                        incomes
##                -9.238136e+00                   4.891621e-02
##                credit_scores table3$`Neighborhood income`
##                 3.141410e-02                   9.551381e-05
```

```
names(RegOut) <- NULL # removing column names

#For Set 3, what is the probability of loan approval for a person whose credit score is 837, income is 
q13 <- exp(RegOut[1]+RegOut[2]*60.899+RegOut[3]*83.7+RegOut[4]*40.158)
q13
```

```
## [1] 0.02662268
```

```
#For Set 3, what are the odds of loan approval for a person who lives in a neighborhood with income of 
q14 <- exp(RegOut[4]*(48.726-45.110))
q14
```

```
## [1] 1.000345
```

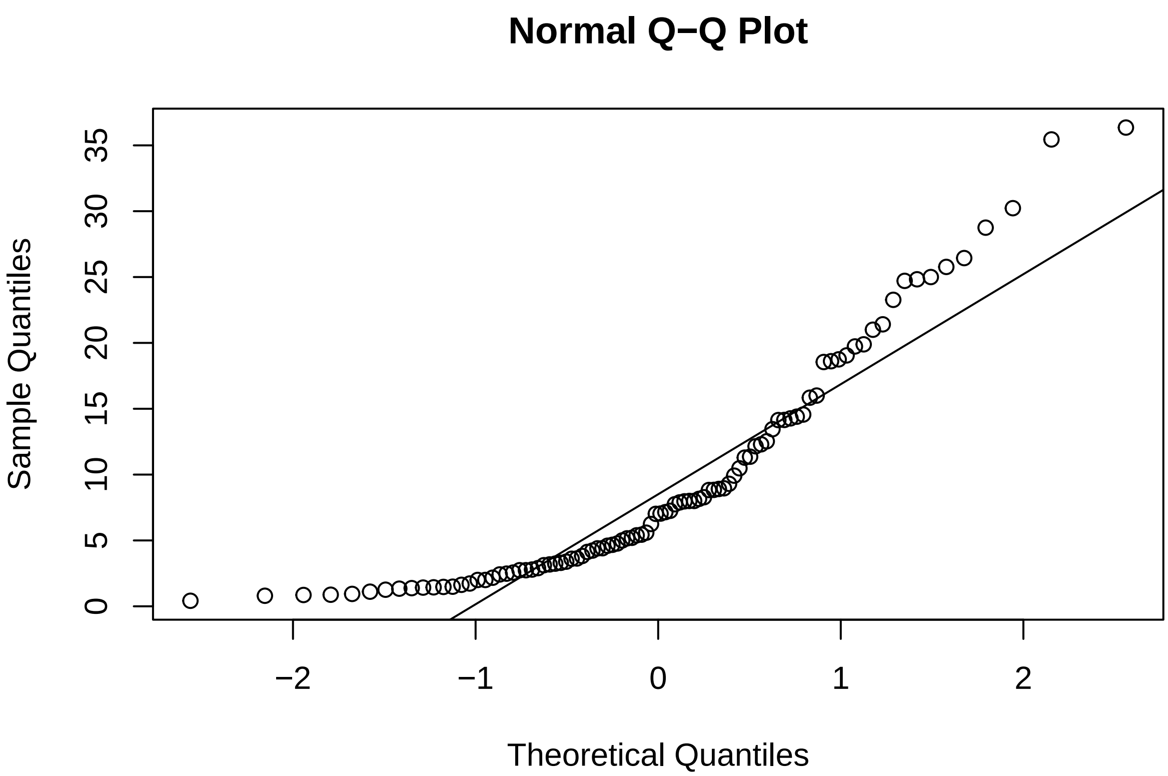
```
#For Set 3, what are the Odds of loan approval for a person whose credit score is 826, income is 56217 
q15 <- exp(RegOut[1]+RegOut[2]*56.217+RegOut[3]*82.6+RegOut[4]*42.744)
q15
```

```
## [1] 0.02045913
```

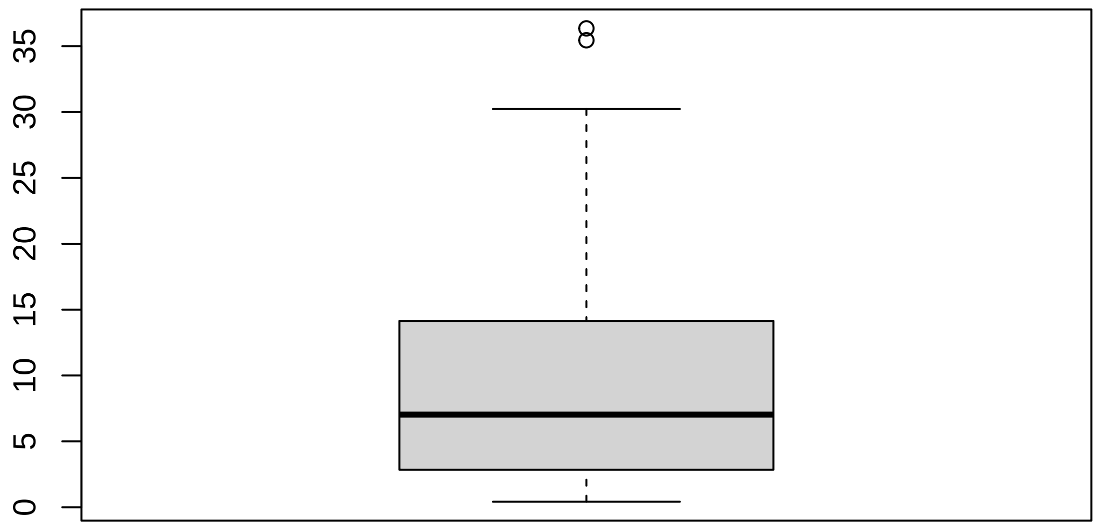
Problem 5 You've picked up a bunch of rocks from a rocky beach and want to estimate the weight of all the rocksat the beach with a Confidence level of 93.47%.

```
## Problem 5 (Set-4)
library(moments)

# Reading the data
table4<-read_excel("F22-6359-Test-3.xlsx", sheet="Set-4") # reading excel sheet
```

```r
# a.  Plot the qqline and boxplot of the data.  Also get the skewness.
# What is your conclusion about the distribution being normal?

#  qqline
qqnorm(table4$Weight)
qqline(table4$Weight)
```




```r
# boxplot
boxplot(table4$Weight)
```

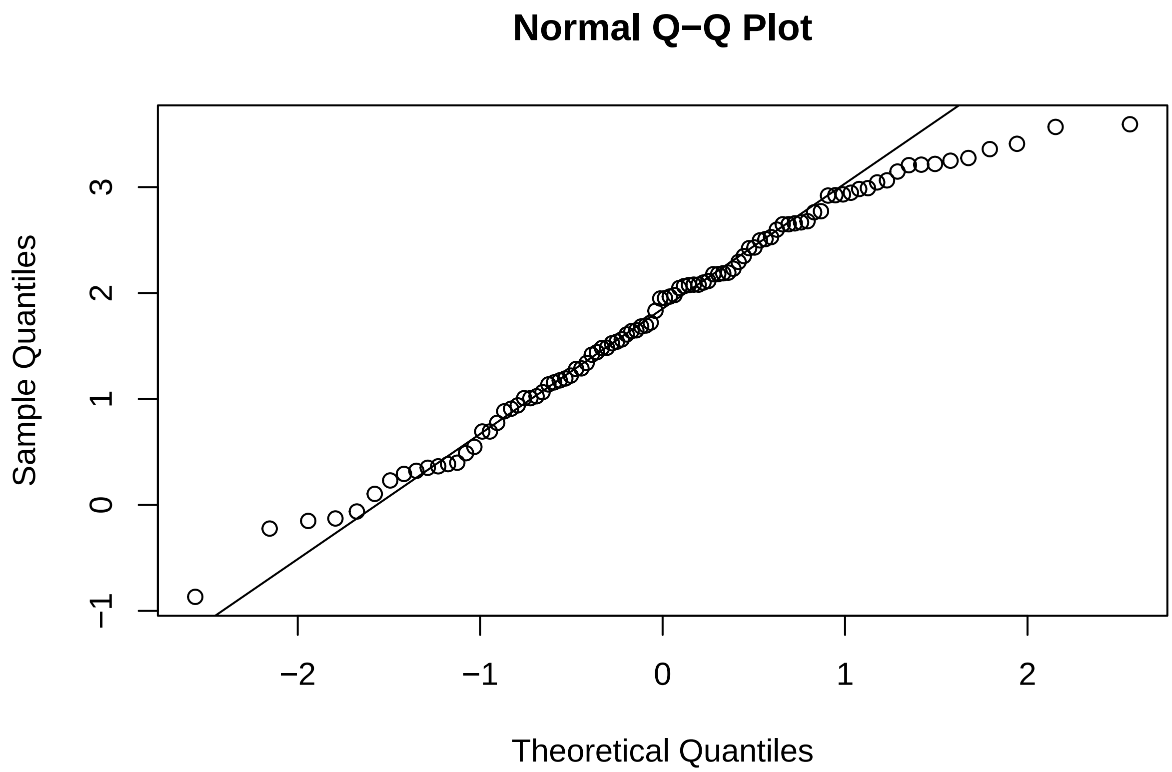



```r
# Skewness
skewness<-skewness(table4$Weight)
skewness
```

```
## [1] 1.233937
```

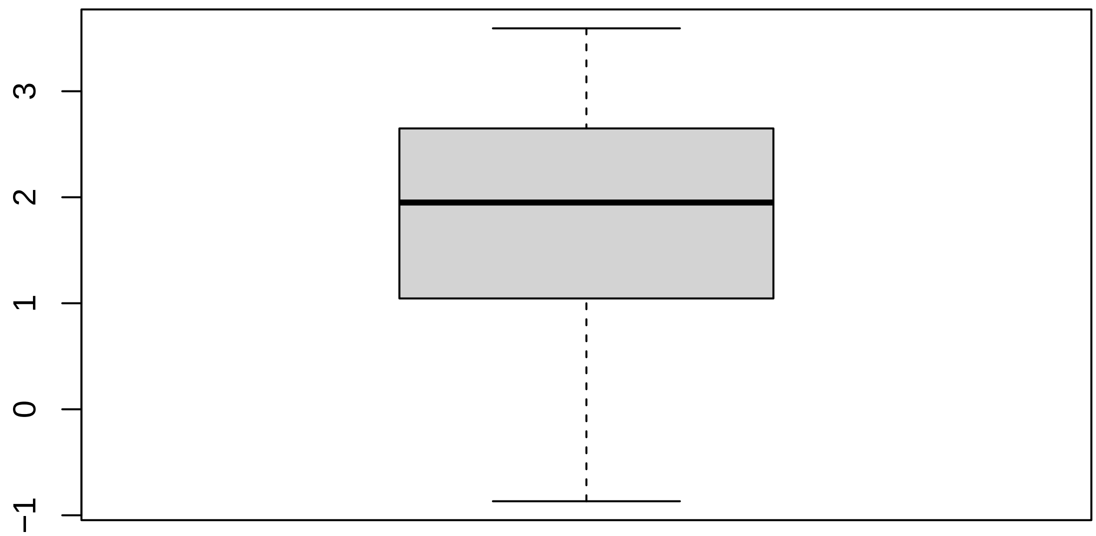
Conclusion: not normally distributed, skewness > 1

```r
#Log Transformation with base e
z<-log(table4$Weight)

qqnorm(z)
qqline(z)
```



Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−1
0
1
2
3
−2
−1
0
1
2


```
boxplot(z)
```




```
skewnwss.transformation<-skewness(z)
skewnwss.transformation
```

```
## [1] -0.2867998
```

Conclusion: normally distributed , skewness is approximately 0

```
#c.  What is the mean, Std dev, and the sample size?

# mean
mean_p5 <- mean(z)

# std deviation
sd_p5<-sd(z)

# sample size
sample_size_p5<- length(table4$Weight)
```

Mean = 1.7919505
Standard Deviation = 1.0267557
sample size 96

```
#d.  Find std error using the std error formula we've discussed.

# std error
se_p5<- sd_p5/sqrt(sample_size_p5)
```

Standard error = 0.1047928

```r
# e.  Find the t-score for the 93.47% confidence interval.

# T-score
tscore_p5 <- qt(0.03265, df=sample_size_p5, lower.tail = FALSE)
```

T-score = 1.8645479

```r
# f.  Use this t-score, sample mean, std error to get the upper and lower limit of the Confidence Inter

# Upperlimit
upperlimit_p5<- mean_p5 + tscore_p5*se_p5

# Lowerlimit
lowerlimit_p5<- mean_p5 - tscore_p5*se_p5
```

Upper limit = 1.9873417
Lower limit = 1.5965592

```r
# g.  Do reverse transformation to get the Confidence Interval in Ounces.

# Reverse transformation
u_p5<- exp(upperlimit_p5)
l_p5<- exp(lowerlimit_p5)
```

Upper limit = 7.2961127 Ounces
Lower limit = 4.9360195 Ounces

Problem 6 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.

```r
## Problem 6 (Set-5)

table5<- read_excel("F22-6359-Test-3.xlsx", sheet="Set-5") #reading excel sheet
```

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

```r
data5 <- data.frame(table5) # convert to dataframe

d <- data5
d$...1 <- NULL # remove first column
d <- d[-4,1:3] # remove totals

rownames(d) <- data5$...1[-4] # setting row names
```

```r
d # new dataset
```

```
##             Favor Indifferent Opposed
## Democrat      169         185     158
## Republican    164         145     157
## Independent    28          50      44
```

```r
chisq.test(d) # running chisq-test function
```

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

```r
# Determine Chi-Square critical value
qchisq(p = .05, df = 4, lower.tail = FALSE)
```

```
## [1] 9.487729
```