**Module Assignment**

**Module 3**

**QMB-6304 Analytical Methods for Business**

Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

**Preprocessing**

1. Load the file “6304 Module 3 Assignment Data.xlsx” into R. This file contains information on the times required for each of 337,145 transactions at tax collector facilities in four disguised cities of a county in Florida.
2. Split the data by facility. You can use any of several approaches to this, including the “split” command which will create four data objects nested inside a list. The command is of the form:

*xx=split(master.data.frame,master.data.frame$split.variable)*

1. Pull out data on the facilities in the cities of Hooterville and Lapataganj. To do this use the R command:

*facility.data.object=xx[[order.in.xx.list]]*

1. Using the numerical portion of your U number as a random number seed and the method demonstrated in class, take a random sample of 40 cases from each facility. Store the sampled observations from each facility in separate data frames.

**Analysis**

1. Using your sample, construct a 90% confidence interval on the population mean transaction time for Hooterville.
2. Assuming the data in the Hooterville data frame represents the population, does your 90% confidence interval include the true population mean on the transaction time variable?
3. Use R and your reduced 40-case data set for Lapataganj. Can you say (α = .05) that the population mean transaction time is greater than 8 minutes? How about greater than 9 minutes and 15 seconds?
4. Referencing Part 3 above, what “test against” (mu) value in a two-tailed hypothesis test would yield p = .05 in a two-tailed hypothesis test on the Lapataganj transaction time?
5. Using R and your reduced 40-case data sets, show comparative notched boxplots of the two facilities’ transaction time variable. Your boxplots should be displayed side by side in a single graphic with an appropriate title and x-axis labels. Do these plots indicate a possible difference between the transaction times for the two facilities? Do these plots indicate a difference in skewness or number of potential outliers between Hooterville and Lapataganj?
6. Using R and your reduced 40-case data sets, does there appear to be a statistically significant difference (α = .05) between the mean transaction times for Hooterville and Lapataganj?

Your deliverable will be a single MS-Word file created using R Markdown. Your file will show 1) the R script which executes the above instructions and 2) the results of those instructions. The first two lines of your deliverable will state this is “Module 3 Assignment” of our course and your name as it appears in Canvas. Your code chunks and analysis results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.

**Preprocessing:**

#Varun Teja Kolluru

#Module 3 Assignment

rm(list=ls())

#import required libraries

library(rio)

library(moments)

#Preprocessing

#import the data from the excel

my\_data=import('6304 Module 3 Assignment Data.xlsx')

#split the data based on facility

df<-split(my\_data,f=my\_data$facility)

df\_hl<-rbind(df$Lapataganj,df$Hooterville)

#set seed and get 40 random samples for each facility

set.seed(97)

fd\_c=df$Cecily[sample(1:nrow(df$Cecily),40,replace=FALSE),]

fd\_h=df$Hooterville[sample(1:nrow(df$Hooterville),40,replace=FALSE),]

fd\_l=df$Lapataganj[sample(1:nrow(df$Lapataganj),40,replace=FALSE),]

fd\_p=df$Pixley[sample(1:nrow(df$Pixley),40,replace=FALSE),]

Using rm, we can clear all the variables in the environment window. All the useful libraries are imported and using the import statement the data is imported to R. As per the given steps data is split and a seed is set with my last 2 digits of U number and a random sample of 40 are taken for the analysis part.

**Analysis:**

1. Using your sample, construct a 90% confidence interval on the population mean transaction time for Hooterville.

R code:

#1 question 90% CI

#mean+\_z\*(sd/sqrt(n))- to find the CI

#h\_data=subset(my\_data,facility=="Hooterville")

mn=mean(df$Hooterville$transaction.time)#Mean of the Hooterville sample

stn=sd(df$Hooterville$transaction.time)#Standard deviation of the sample

n=nrow(df$Hooterville)#number of rows in the sample

z=qnorm(.95)

me<-z\*(stn/sqrt(n))

mn-me#Lower Value

mn+me#Upper value

Result in console:

> mn=mean(df$Hooterville$transaction.time)#Mean of the Hooterville sample

> stn=sd(df$Hooterville$transaction.time)#Standard deviation of the sample

> n=nrow(df$Hooterville)#number of rows in the sample

> z=qnorm(.95)

> me<-z\*(stn/sqrt(n))

> mn-me#Lower Value

[1] 7.227569

> mn+me#Upper value

[1] 7.335198

90% Confidence Interval for Hooterville sample transaction time is

Lower Value- ‘7.227569’

Upper Value- ‘7.335198’

1. Assuming the data in the Hooterville data frame represents the population, does your 90% confidence interval include the true population mean on the transaction time variable?

Rcode:

t.test(df$Hooterville$transaction.time,conf.level=0.90,

alternative=c("two.sided"))

One Sample t-test

data: df$Hooterville$transaction.time

t = 222.56, df = 58780, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

90 percent confidence interval:

7.227568 7.335199

sample estimates:

mean of x

7.281383

t.test(df$Hooterville$transaction.time,conf.level=0.95,

alternative=c("two.sided"))

One Sample t-test

data: df$Hooterville$transaction.time

t = 222.56, df = 58780, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

7.217258 7.345509

sample estimates:

mean of x

7.281383

The True population mean on the transaction variable falls in between the 95% confidence interval. The confidence interval for 95% ranges from 7.217258 to 7.345509. The true mean on the transaction variable is in between this range. The range for 90% confidence interval is from 7.227568 to 7.335199. Clearly we can see that the 90% confidence interval includes the true population mean on the transaction variable.

1. Use R and your reduced 40-case data set for Lapataganj. Can you say (α = .05) that the population mean transaction time is greater than 8 minutes? How about greater than 9 minutes and 15 seconds?

R Code:

lapa=t.test(fd\_l$transaction.time,mu=8,alternative = "greater")

lapa

Result in console:

data: fd\_l$transaction.time

t = 2.4173, df = 39, p-value = 0.01021

alternative hypothesis: true mean is greater than 8

95 percent confidence interval:

8.939273 Inf

sample estimates:

mean of x

11.1

Here, null hypothesis says our population mean transaction time, Ho, is < 8 and the alternate,

Ha, is > 8. Below results reflect probability, aka p value is telling us the chance of being wrong or committing a type 1 error is 0.01021 there is a less than 1 percent chance you are wrong by rejecting the null and accepting the alternate. We are concerned with p values that are greater than 5 percent or .05000

In this case, the p value is less than 5 percent. **So that, we are rejecting the null and accepting the alternate.**

For testing with greater than 9 min 15 sec.

R Code:

lapa\_gt\_9=t.test(fd\_l$transaction.time,mu=9.25,alternative = 'greater')

lapa\_gt\_9

Result in console:

One Sample t-test

data: fd\_l$transaction.time

t = 1.4426, df = 39, p-value = 0.07856

alternative hypothesis: true mean is greater than 9.25

95 percent confidence interval:

8.939273 Inf

sample estimates:

mean of x

11.1

Here, null hypothesis says our population mean transaction time, Ho, is < 9.25 and the alternate, Ha, is > 9.25. Below results reflect the probability, aka p value is telling us the chance of being wrong or committing a type 1 error is 0.07856

There is a less than 7.8 percent chance you are wrong by rejecting the null and accepting the alternate.

**In this case we would fail to reject the null as the p value is greater than 5 percent.**

1. Referencing Part 3 above, what “test against” (mu) value in a two-tailed hypothesis test would yield p = .05 in a two-tailed hypothesis test on the Lapataganj transaction time?

Rcode:

lapa\_ts=t.test(fd\_l$transaction,mu=9,alternative = "two.sided")

lapa\_ts

lapa\_ts=t.test(fd\_l$transaction,mu=8,alternative = "two.sided")

lapa\_ts

lapa\_ts=t.test(fd\_l$transaction,mu=8.5,alternative = "two.sided")

lapa\_ts

Result in console:

> #4

> lapa\_ts=t.test(fd\_l$transaction,mu=9,alternative = "two.sided")

> lapa\_ts

One Sample t-test

data: fd\_l$transaction

t = 1.6375, df = 39, p-value = 0.1096

alternative hypothesis: true mean is not equal to 9

95 percent confidence interval:

8.506049 13.693951

sample estimates:

mean of x

11.1

> lapa\_ts=t.test(fd\_l$transaction,mu=8,alternative = "two.sided")

> lapa\_ts

One Sample t-test

data: fd\_l$transaction

t = 2.4173, df = 39, p-value = 0.02041

alternative hypothesis: true mean is not equal to 8

95 percent confidence interval:

8.506049 13.693951

sample estimates:

mean of x

11.1

> lapa\_ts=t.test(fd\_l$transaction,mu=8.5,alternative = "two.sided")

> lapa\_ts

One Sample t-test

data: fd\_l$transaction

t = 2.0274, df = 39, p-value = 0.04949

alternative hypothesis: true mean is not equal to 8.5

95 percent confidence interval:

8.506049 13.693951

sample estimates:

mean of x

11.1

We are only concerned with the p value, the p value should be less than 5 percent. For the two tailed hypotheses test, first I have taken mu= 9 , where the p value is nearly 11 percent, which is more than 5 percent. **So in this case, we would fail to reject the null as the p value is greater than 5 percent.**

So next I have taken the mu=8, where the p value is nearly 2 percent, where the p value is less than 5 percent. **So that, we are rejecting the null and accepting the alternate**.

Next, I have taken the mu=8.5, where the p value is nearly equaling 5 percent. **So that, we are rejecting the null and accepting the alternate**.

1. Using R and your reduced 40-case data sets, show comparative notched boxplots of the two facilities’ transaction time variable. Your boxplots should be displayed side by side in a single graphic with an appropriate title and x-axis labels. Do these plots indicate a possible difference between the transaction times for the two facilities? Do these plots indicate a difference in skewness or number of potential outliers between Hooterville and Lapataganj?

R Code:

skewness(fd\_h$transaction.time)

skewness(fd\_l$transaction.time)

boxplot(fd\_h$transaction.time,fd\_l$transaction.time,

col=c("red","blue"),names=c("Hooterville","Lapataganj"),

main="Transaction time Notched Boxplot",

notch = TRUE)

Result from console:

> skewness(fd\_h$transaction.time)

[1] 3.188138

> skewness(fd\_l$transaction.time)

[1] 1.173407

> boxplot(fd\_h$transaction.time,fd\_l$transaction.time,

+ col=c("red","blue"),names=c("Hooterville","Lapataganj"),

+ main="Transaction time Notched Boxplot",

+ notch = TRUE)

Chart

Description automatically generated

The graph displays the transaction time for two facilities, Hooterville and Lapataganj. The notch in the graph is the 95% confidence interval and we can say that the confidence interval for Hooterville is tighter than Lapataganj, which is good. And we can see that the confidence interval for both the facilities are in same range.

In the graph, for Hooterville, there are several outliers when compared to Lapataganj and from the skewness values both the facilities have left skewness and Hooterville is more skewed than Lapataganj, because of more outliers.

1. Using R and your reduced 40-case data sets, does there appear to be a statistically significant difference (α = .05) between the mean transaction times for Hooterville and Lapataganj?

R Code:

results=t.test(fd\_h$transaction.time,fd\_l$transaction.time,mu=0,

alternative=c("two.sided"),paired=TRUE)

results

Result in console:

> results

Paired t-test

data: fd\_h$transaction.time and fd\_l$transaction.time

t = -0.55185, df = 39, p-value = 0.5842

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

-5.598347 3.198347

sample estimates:

mean difference

-1.2

Here, null hypothesis Ho, mean transaction time of Hooterville and mean transaction time of Lapatagunj is = 0, and the alternate, Ha, Htt – Ltt is not = 0

Below results reflect the probability, aka p value is telling us the chance of being wrong or committing a type 1 error is 0.5842.

There is a less than 58 percent chance you are wrong by rejecting the null and accepting the alternate.

**In this case we would fail to reject the null as the p value is greater than 5 percent.**

I would conclude there is a difference between the population mean transaction time of 95 percent confident the true population transaction time mean difference is between -5.598347 and 3.198347