Hypothesis Testing Assignment

1. A F&B manager wants to determine whether there is any significant difference in the diameter of the cutlet between two units. A randomly selected sample of cutlets was collected from both units and measured? Analyze the data and draw inferences at 5% significance level. Please state the assumptions and tests that you carried out to check validity of the assumptions.

Answer:-

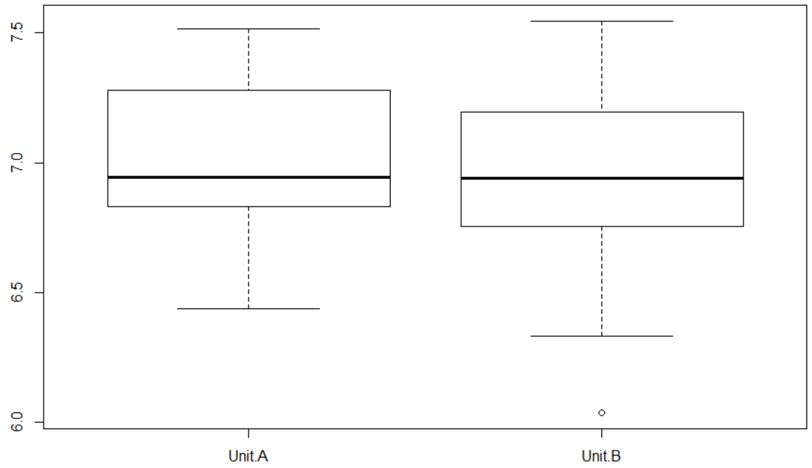
Step1: Business Problem: Two check whether the diameter of two units are similar or not?

Step2: y and x: So here is y is continuous and x is discrete

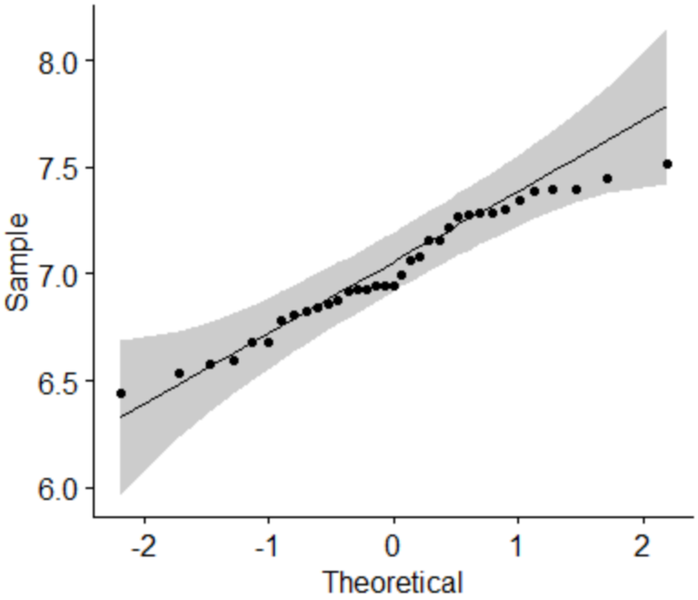
Step3: Here we will use 2-sample t test

Step4: Find normality of this data

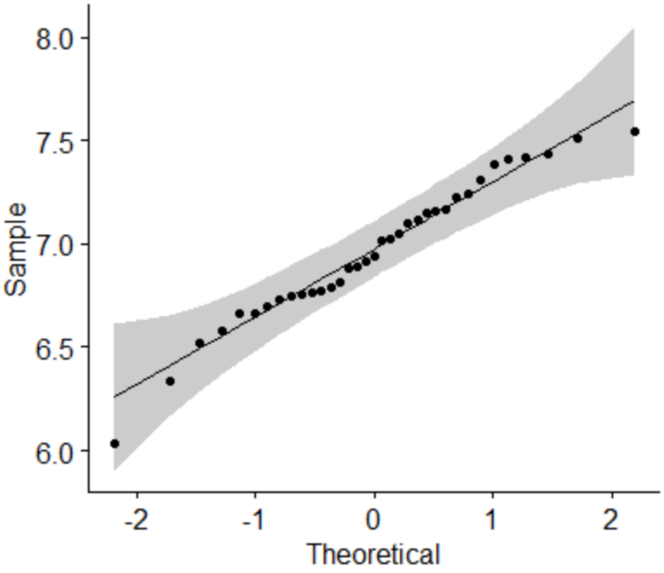
```
> Cutlets <- read.csv("C:/PRATIK/Data Science/Assignment/Hypothesis Testing/Cutlets.csv")
> View (Cutlets)
> attach (Cutlets)
> Boxplot (Cutlets)
> # H0 = There is no significant difference in the diameter of the cutlets bet 2 units
> # Ha = There is a difference in the diameter of the cutlets bet 2 units
> # Here we will use 2 sample t test and
> # if p-value is > 0.05 => Accept the Null Hypothesis
> # if p-value is < 0.05 => Reject Null Hypothesis
```



```
> library("dplyr")
> library("ggpubr")
> data <- read.csv(file.choose())
> set.seed(1234)
> dplyr::sample_n(data, 10)
   Unit.A Unit.B
1  6.6801 6.9182
2  6.8394 7.0240
3  7.1560 7.4220
4  7.4488 7.1522
5  7.5169 7.4059
6  6.5797 7.1581
7  6.6840 7.2402
8  7.2783 7.1180
9  7.3871 6.8110
10 7.3943 6.5780
> library(ggpubr)
> ggqqplot(data$Unit.A)
```



```
> ggqqplot(data$Unit.B)
```



```
t.test(Unit.A, Unit.B, alternative = "two.sided", var.equal = FALSE)

        Welch Two Sample t-test

data:  Unit.A and Unit.B
t = 0.72287, df = 66.029, p-value = 0.4723
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09654633  0.20613490
sample estimates:
mean of x mean of y
 7.019091  6.964297
```

Here our p-value is > 0.05, so data are normal

H0: variance of Unit.A = variance of Unit.B

Ha: variance of Unit.A NOT= variance of Unit.B

We can go for further test which is **variance test**

```
> var(data$Unit.A)
[1] 0.08317945
```

```
> var(data$Unit.B)
[1] 0.117924
```

```
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 0.45428, df = 34, p-value = 1
```

As per chi-square test p-value is 1.00 > 0.05 = Accept Ho

H0: variance of Unit.A = variance of Unit.B

2 Sample t test for compare mean

H0: Average of Unit.A = Average of Unit.B

Ha: variance of Unit.A NOT = variance of Unit.B

```
> t.test(data$Unit.A,data$Unit.B)

	Welch Two Sample t-test

data:  data$Unit.A and data$Unit.B
t = 0.72287, df = 66.029, p-value = 0.4723
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09654633  0.20613490
sample estimates:
mean of x mean of y 
 7.019091  6.964297 
```

P-value is 0.4723 > 0.05=> Accept Ho, hence Average of unit A = Average of unit B

There is no significant difference in the diameter of the cutlets bet 2 units

2. A hospital wants to determine whether there is any difference in the average Turnaround Time (TAT) of reports of the laboratories on their preferred list. They collected a random sample and recorded TAT for reports of 4 laboratories. TAT is defined as sample collected to report dispatch.

Analyze the data and determine whether there is any difference in average TAT among the different laboratories at 5% significance level.

Answer:

Business Problem: TAT for all 4 laboratories are same or different

H0: Data are normal
Ha: Data are not normal

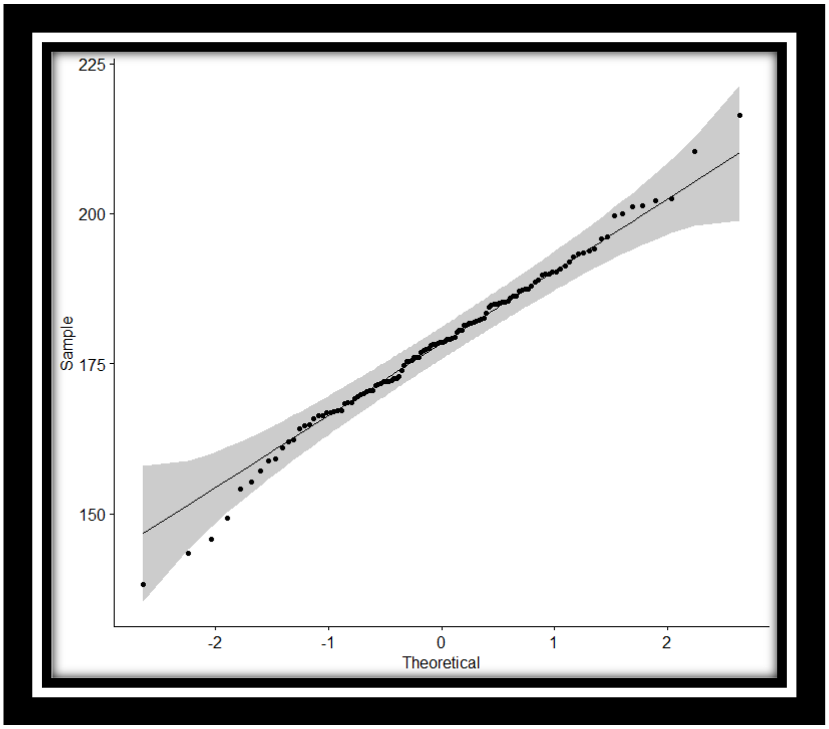
H0: There is no difference in average TAT among those laboratories
Ha: There is a difference in average TAT among those laboratories

P-value <=5% then accept the Ha/Reject H0   (There is difference in average TAT)
P-value >5% then accept the H0/Fail to reject H0   (There is no difference in average TAT)

Normality Test:

Lab 1:

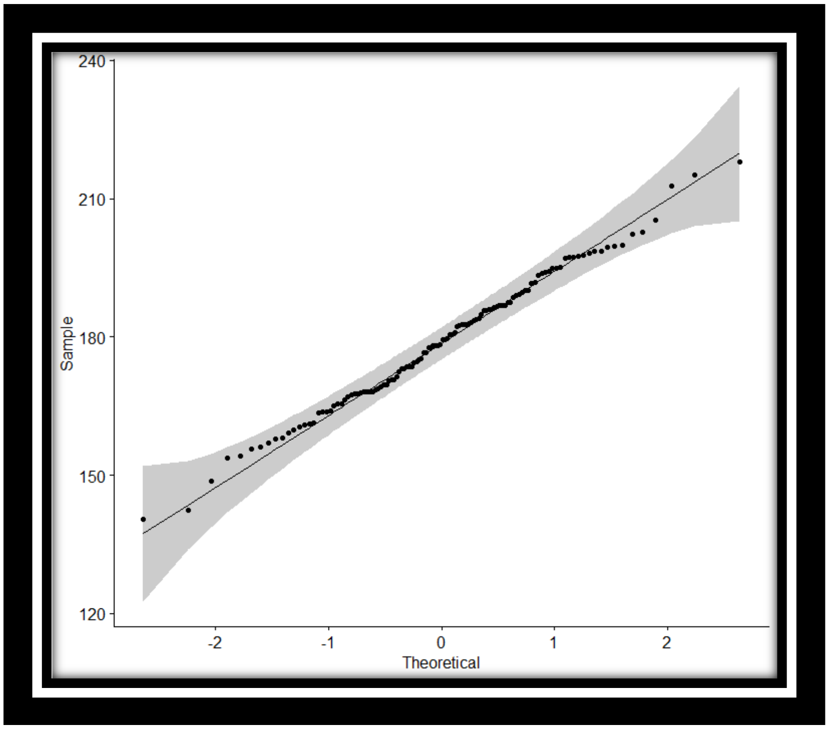
```
> shapiro.test(my_data$Laboratory.1)

        Shapiro-Wilk normality test

data:  my_data$Laboratory.1
W = 0.99018, p-value = 0.5508
```



```
P-value = 0.5508 > 0.05 => Accept H0
```

Lab 2:

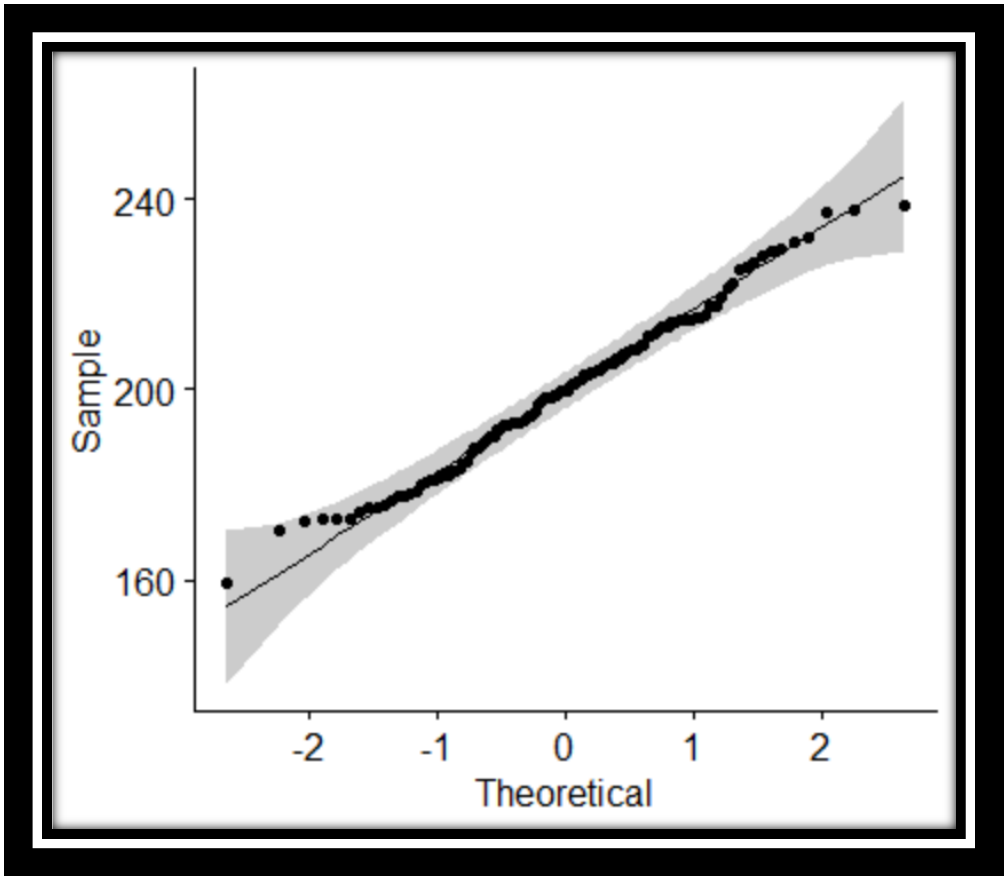
```
> shapiro.test(my_data$Laboratory.2)

        Shapiro-Wilk normality test

data:  my_data$Laboratory.2
W = 0.99363, p-value = 0.8637
```



P-value = 0.8637 > 0.05 => Accept H0

Lab 3:

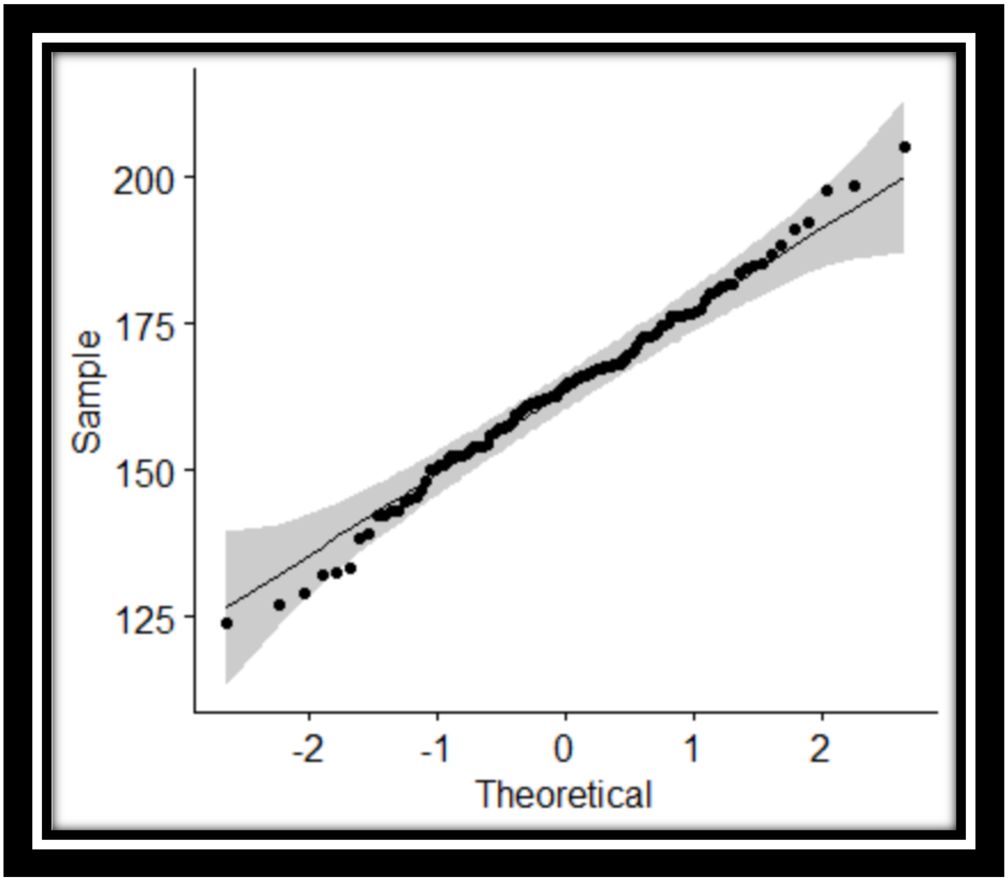
```
> shapiro.test(my_data$Laboratory.3)

        Shapiro-Wilk normality test

data:  my_data$Laboratory.3
W = 0.98863, p-value = 0.4205
```



```
P-value = 0.4205 > 0.05 => Accept H0
```

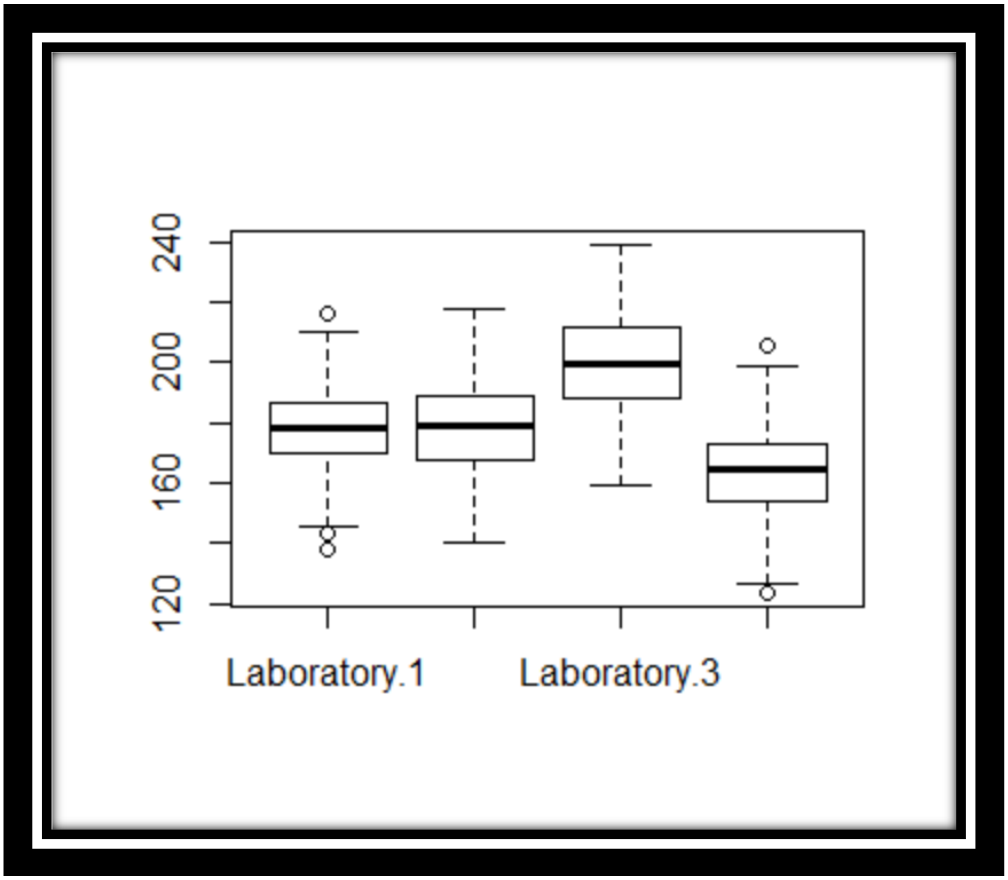
Lab 4:

```
> shapiro.test(my_data$Laboratory.4)

        Shapiro-Wilk normality test

data:  my_data$Laboratory.4
W = 0.99138, p-value = 0.6619
```



P-value = 0.6619 > 0.05 => Accept H0

```
> boxplot(my_data)
```



Here data are normal so we will do the **Variance Test**:

H0 = σ^2 (L1) = σ^2 (L2) = σ^2 (L3)..................... (All variances are equal)

Ha = σ^2 (L1) ≠ σ^2 (L2) ≠ σ^2 (L3)..................... (At least 1 variance is different)

H0 = Variance TAT of all 4 Labs are same
Ha = Variance TAT of at least 1 Lab is different

By using F-Test

```
> res.ftest <- var.test(lab$Laboratory.1,lab$Laboratory.2,data = lab)
> res.ftest

        F test to compare two variances

data:  lab$Laboratory.1 and lab$Laboratory.2
F = 0.77573, num df = 119, denom df = 119, p-value = 0.1675
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5406345 1.1130690
sample estimates:
ratio of variances
```

```
         0.7757342

> res.ftest <- var.test(lab$Laboratory.2,lab$Laboratory.3,data = lab)
> res.ftest

	F test to compare two variances

data:  lab$Laboratory.2 and lab$Laboratory.3
F = 0.81785, num df = 119, denom df = 119, p-value = 0.2742
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5699887 1.1735038
sample estimates:
ratio of variances
         0.8178532

> res.ftest <- var.test(lab$Laboratory.3,lab$Laboratory.4,data = lab)
> res.ftest

	F test to compare two variances

data:  lab$Laboratory.3 and lab$Laboratory.4
F = 1.2021, num df = 119, denom df = 119, p-value = 0.3168
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8377527 1.7247817
sample estimates:
ratio of variances
          1.202057
```

Here p-value is > 0.05 => Accept Ho, hence we prove variance of all laboratory are same

3. Sales of products in four different regions is tabulated for males and females. Find if male-female buy er rations are similar across regions.

Answer:

Step 1: Business Problem: Male-female buyer rations are similar across regions

Step 2: y and x : x is more than 2 discrete and y is discrete

Step 3: Here we will use Chi-square test

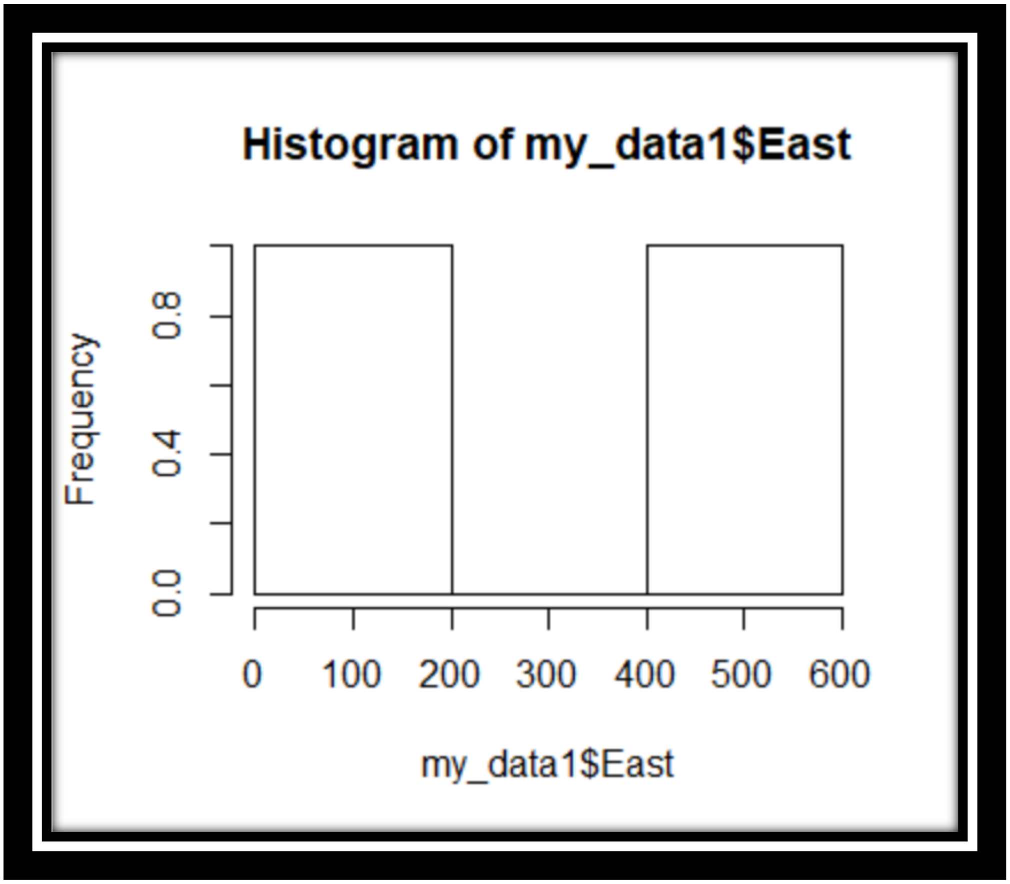
H0: Data are normal
Ha: Data are not normal

H0: Male-female buyer rations are similar across regions
Ha: Male-female buyer rations are not similar across regions

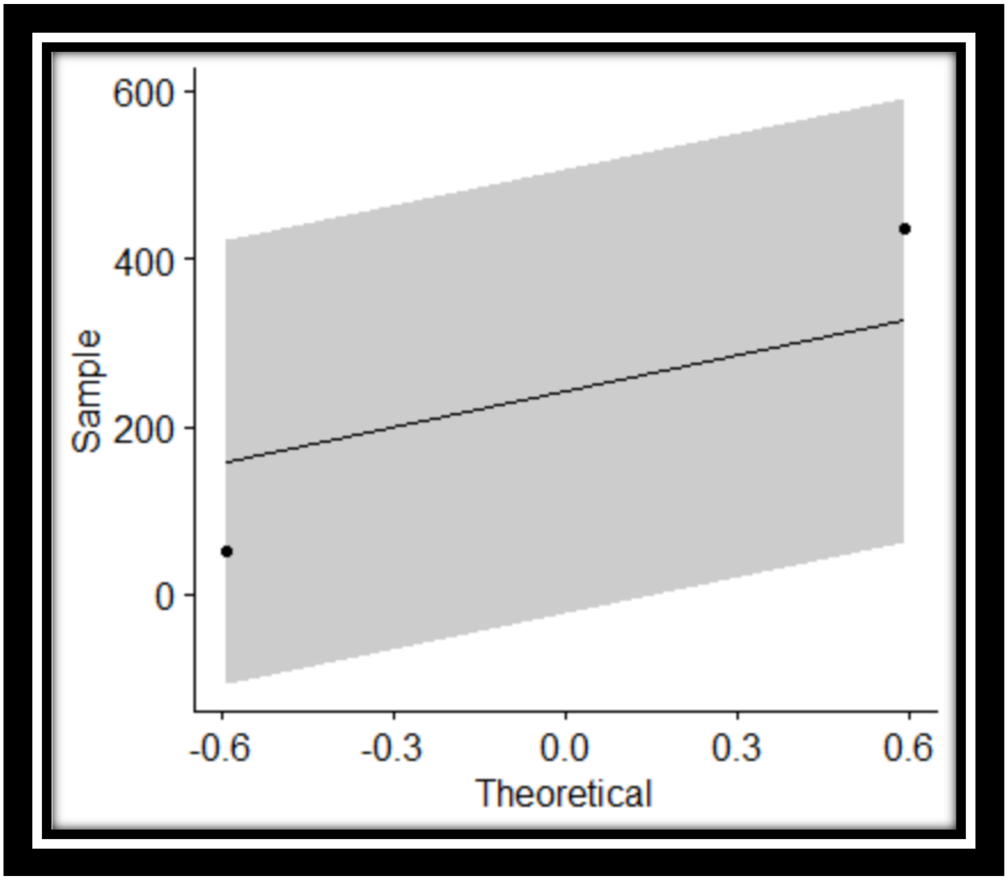
P-value <=5% then accept the Ha/Reject H0 (male-female buyer rations are similar)
P-value >5% then accept the H0/Fail to reject H0 (male-female buyer rations are not similar)

Normality Test:
East Region:

```
>hist(my_data1$East)
```


Histogram of my_data1$East
Frequency
0.0
0.4
0.8
0
100
200
300
400
500
600
my_data1$East

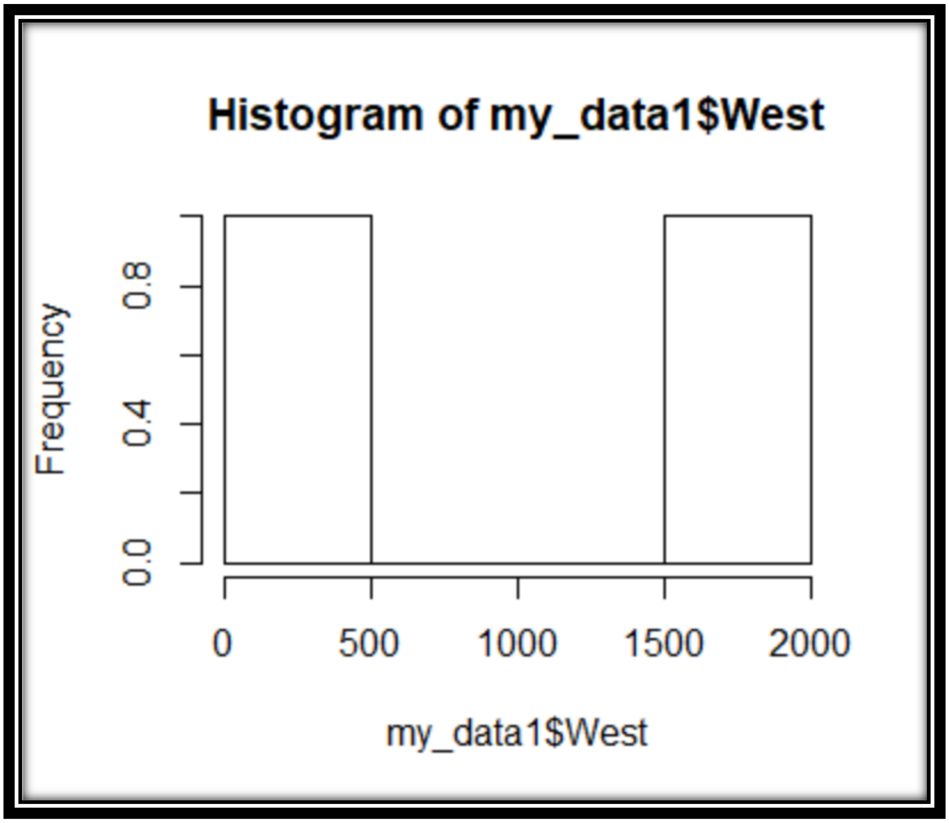
```
> ggqqplot(my_data1$East)
```



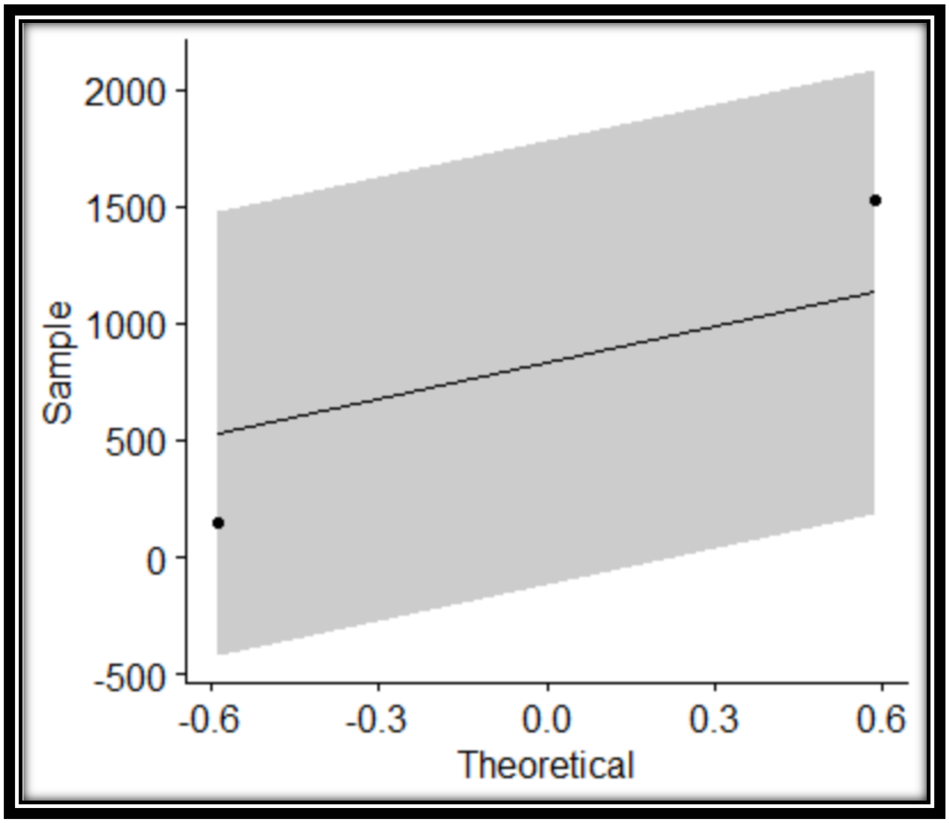
```
> t.test(my_data1$East, alternative = "two.sided", var.equal = FALSE)

        One Sample t-test

data:  my_data1$East
t = 1.2597, df = 1, p-value = 0.4271
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2203.444  2688.444
sample estimates:
mean of x
    242.5
```

West Region:

```
> hist(my_data1$West)
```



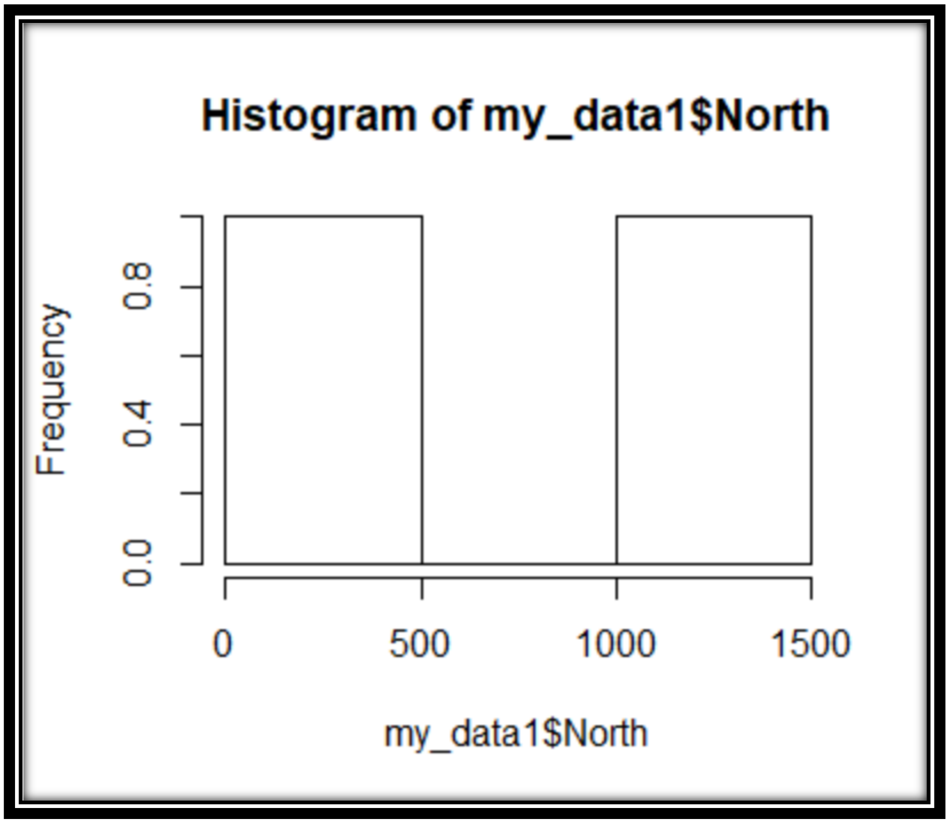
```
> ggqqplot(my_data1$West)
```



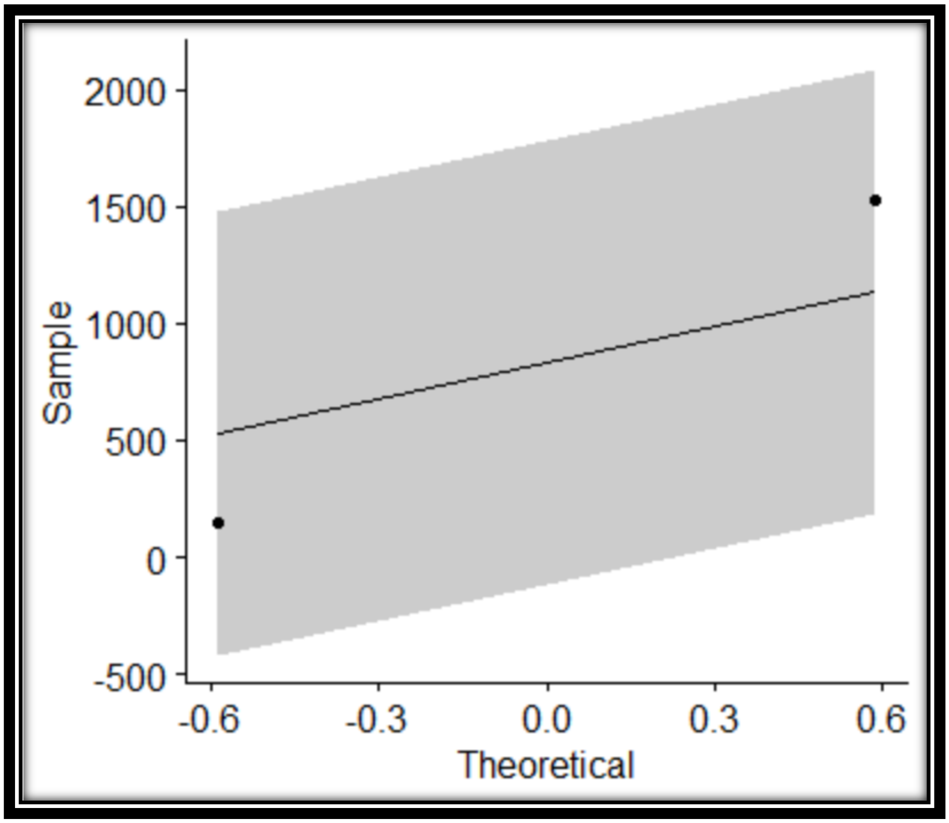
```
> t.test(my_data1$West, alternative = "two.sided", var.equal = FALSE)

        One Sample t-test

data:  my_data1$West
t = 1.2056, df = 1, p-value = 0.4408
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -7941.134  9606.134
sample estimates:
mean of x 
    832.5 
```

North Region:

```
> hist(my_data1$North)
```



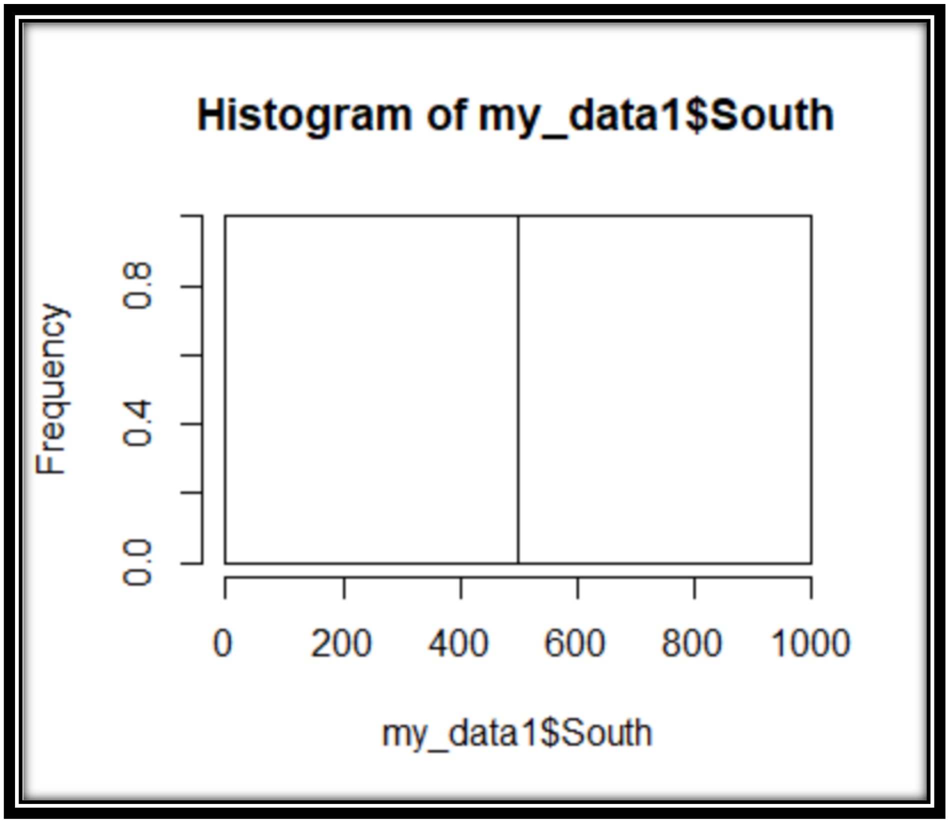
```
> ggqqplot(my_data1$West
```



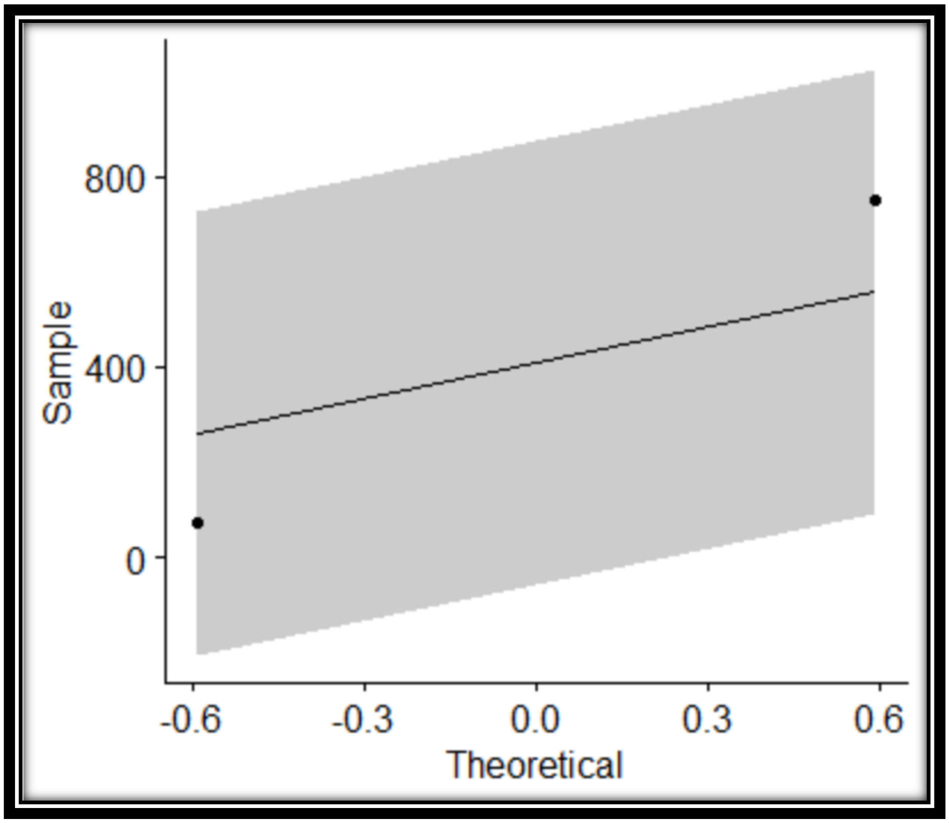
```
> t.test(my_data1$North, alternative = "two.sided", var.equal = FALSE)

        One Sample t-test

data:  my_data1$North
t = 1.2139, df = 1, p-value = 0.4387
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -7039.05  8526.05
sample estimates:
mean of x
    743.5
```

South Region:

```
> hist(my_data1$South)
```



```
> ggqqplot(my_data1$South)
```



```
> t.test(my_data1$South, alternative = "two.sided", var.equal = FALSE)

        One Sample t-test

data:  my_data1$South
t = 1.2059, df = 1, p-value = 0.4408
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -3910.11  4730.11
sample estimates:
mean of x 
      410 
```

For all regions the P value is greater than 0.05 Hence we accept the H0.

4. Tele Call uses 4 centers around the globe to process customer order forms. They audit a certain % of the customer order forms. Any error in order form renders it defective and has to be reworked before processing. The manager wants to check whether the defective % varies by center. Please analyze the data at *5%* significance level and help the manager draw appropriate inferences

Answer:

Step1: Business Problem: To check whether the defective % varies by center or not

Step2: **y and x**

x is more than 2 discrete and y is discrete

Step3: **Here we will use Chi-square test**

H0: All are same

Ha: at least 1 are different

```
> chisq.test(telecall$Phillippines, telecall$Indonesia, correct=FALSE)

        Pearson's Chi-squared test

data:  telecall$Phillippines and telecall$Indonesia
X-squared = 0.55216, df = 1, p-value = 0.4574
```

```
> chisq.test(telecall$Malta, telecall$India, correct=FALSE)

        Pearson's Chi-squared test

data:  telecall$Malta and telecall$India
X-squared = 2.4695, df = 1, p-value = 0.1161
```

```
> chisq.test(telecall$Malta, telecall$Phillippines, correct=FALSE)

        Pearson's Chi-squared test

data:  telecall$Malta and telecall$Phillippines
X-squared = 0.41474, df = 1, p-value = 0.5196
```

P-value is 0.5196 > 0.05=> Accept Ho, hence Average are same

As per results we can say that all the canters are equal.

5. Fantaloons Sales managers commented that *%* of males versus females walking in to the store differ based on day of the week. Analyze the data and determine whether there is evidence at *5 %* significance level to support this hypothesis.

Answer:

Step1: Business Problem: To find proportion male vs female differ from weekdays or weekends are equal or not

Step2: y and x

x is discrete with 2 categories and y is discrete

Step3: Here we will use 2-Proportion test

2-Proprotion Test

H0: Proportion of male vs female in weekdays = Proportion of male vs female in weekends

Ha: Proportion of male vs female in weekdays NOT = Proportion of male vs female in weekends

```
> faltoons <- read.csv(file.choose())
> chisq.test(faltoons$Weekdays, faltoons$Weekend, correct=FALSE)

	Pearson's Chi-squared test

data:  faltoons$Weekdays and faltoons$Weekend
X-squared = 0.0015979, df = 1, p-value = 0.9681
```

P-value is 0.968 > 0.05 => Accept Ho

Hence Proportion of male vs female in weekdays = Proportion of male vs female in weekends