Data Analysis: HWA06 (Group Exercise - solutions)

MANOS Thanos

Tue 23 Apr 2024

General HWAs instructions:

1. Load your packages and modify your paths and define your chucks as we have seen in previous examples and HWAs (according to your needs).
2. Read **carefully** all the text and answer the questions.
3. Provide a short phrases to accompany your answers. For example: "**The mean value of the male subjects age is 13.4 years**" instead of simply "**13.4**" (just the output value).
4. Use **bold color** text for your answers. In this way, I can inspect faster and locate immediately **your reply**.
5. Always keep the questions or problems statements in the template.
6. **CHECK your PDF files** for missing answers and chopped text at the boarders of the chunks (hit enter and use extra lines if this occurs).
7. If you are unable to give an answer, write down -> "**No answer**" after the respective question in **bold red**.
8. **DO NOT modify the HWA template** (sometimes **xelatex** struggles to knit Rmd files with long names. In such a case, try to rename it and re-knit.)

One-Sample t-test: Implementing the One-Sample t-test in R

In a one-sample t-test, we compare the average (or mean) of one group against the set average (or mean). This set average can be any theoretical value (or it can be the population mean).

Background: assumptions for performing a t-test

There are certain assumptions we need to heed before performing a t-test:

- The data should follow a continuous or ordinal scale.
- The observations in the data should be randomly selected.
- The data should resemble a bell-shaped curve when we plot it, i.e., it should be normally distributed.
- Large sample size should be taken for the data to approach a normal distribution (although t-test is essential for small samples as their distributions are non-normal).
- Variances among the groups should be equal (for independent two-sample t-test).

Exercise 1

A mobile manufacturing company has taken a sample of mobiles of the same model from the previous month's data. They want to check whether the average screen size of the sample differs from the desired length of 10 cm. You can download the data "**screen_size-data.csv**" from Microsoft Teams.

Answer:

- Step 1: First, import the data.
- Step 2: Validate it for correctness in R (check dimensions & view first rows).

```
#Step 1 - Importing Data
#-------------------------------------------------------

#Importing the csv data
data<-read.csv(file = "./screen_size-data.csv")

#data<-read.csv(file.choose()) # screen_size-data.csv

#Step 2 - Validate data for correctness
#-------------------------------------------------------

#Count of Rows and columns
dim(data)
```

```
## [1] 1000    1
```

```
#View top 10 rows of the dataset
head(data,10)
```

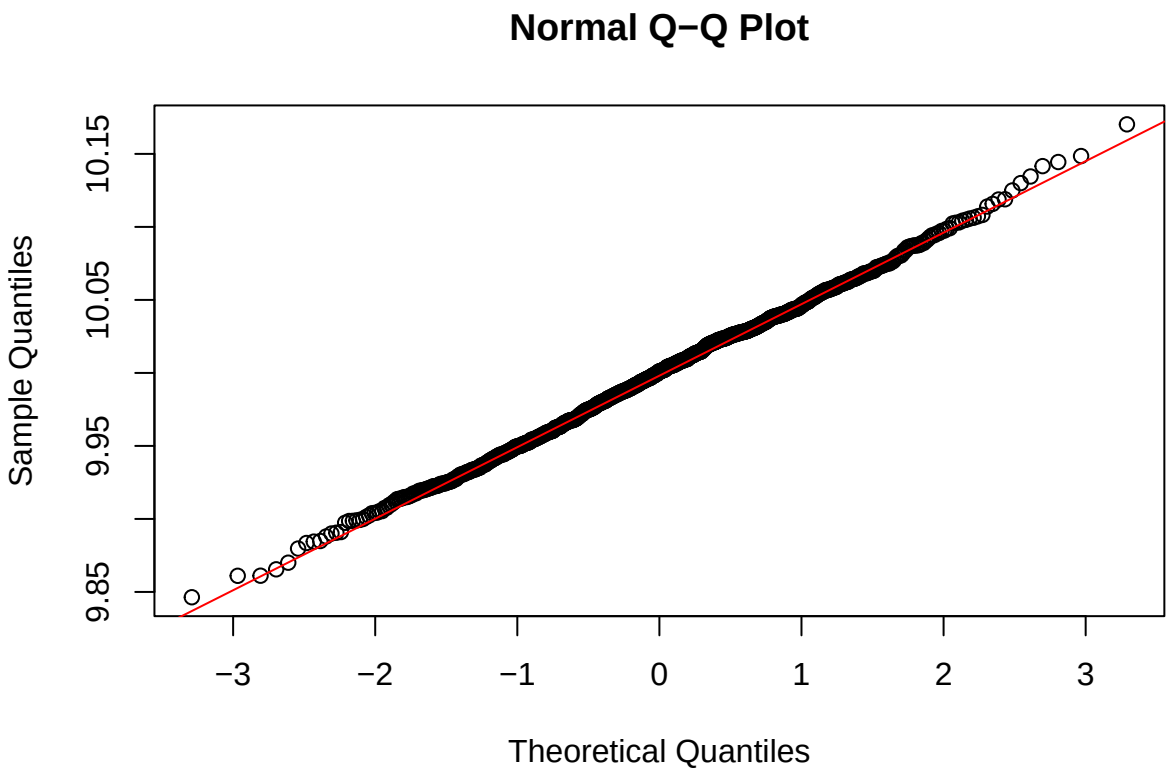
```
##    Screen_size.in.cm.
## 1           10.006692
## 2           10.081624
## 3           10.072873
## 4            9.954496
## 5            9.994093
## 6            9.952208
## 7            9.947936
## 8            9.988184
## 9            9.993365
## 10          10.016660
```

- Step 3: Check the assumptions mentioned earlier, namely whether the data is continuous, randomly selected and normally distributed (use *qqnorm* and *qqline* for the latter):

```
#Step 3 - Check for assumptions
#-------------------------------------------------------

#1. Data is continuous.
#2. Observations are randomly selected.
#3. To check the data is normally distributed, we will use the following commands:

qqnorm(data$Screen_size.in.cm.)
qqline(data$Screen_size.in.cm.,col="red")
```



Almost all the values lie on the red line. We can confidently say that the data follows a normal distribution.

- Step 4: Conduct a one-sample t-test and report your answer.

```
#Step 5 - Conduct one-sample t-test
#Null Hypothesis: Mean screensize of sample does not differ from 10 cm
#Alternate Hypothesis: Mean screensize of sample differ from 10 cm

t.test(data$Screen_size.in.cm.,mu=10)
```

```
##
##  One Sample t-test
##
## data:  data$Screen_size.in.cm.
## t = -0.39548, df = 999, p-value = 0.6926
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.996361 10.002418
## sample estimates:
## mean of x
##   9.99939
```

The t-statistic comes out to be -0.39548. Note that we can treat negative values as their positive counterpart here. Now, refer to the table mentioned earlier for the t-critical value. The degree of freedom here is 999 and the confidence interval is 95%.

The t-critical value is 1.962. Since the t-statistic is less than the t-critical value, we fail to reject the null hypothesis and can conclude that the average screen size of the sample does not differ from 10 cm.

We can also verify this from the p-value, which is greater than 0.05. Therefore, we fail to reject the null hypothesis at a 95% confidence interval.

Independent Two-Sample t-test: Implementing the Two-Sample t-test in R

The two-sample t-test is used to compare the means of two different samples.

Exercise 2

A mobile manufacturing company has taken a sample of mobiles of the same model from the previous month's data. They want to check whether the average screen size of the sample differs from the desired length of 10 cm. You can download the data "**ind.csv**" from Microsoft Teams.

Answer:

- Step 1: Again, first import the data.
- Step 2: Validate it for correctness in R (check dimensions & view first rows).

```
#Step 1 - Importing Data
#-----------------------------------------------------------

#Importing the csv data
data<-read.csv(file = "./ind.csv")

#data<-read.csv(file.choose()) # ind.csv

#Step 2 - Validate data for correctness
#-----------------------------------------------------------

#Count of Rows and columns
dim(data)
```

```
## [1] 1000    2
```

```
#View top 10 rows of the dataset
head(data,10)
```

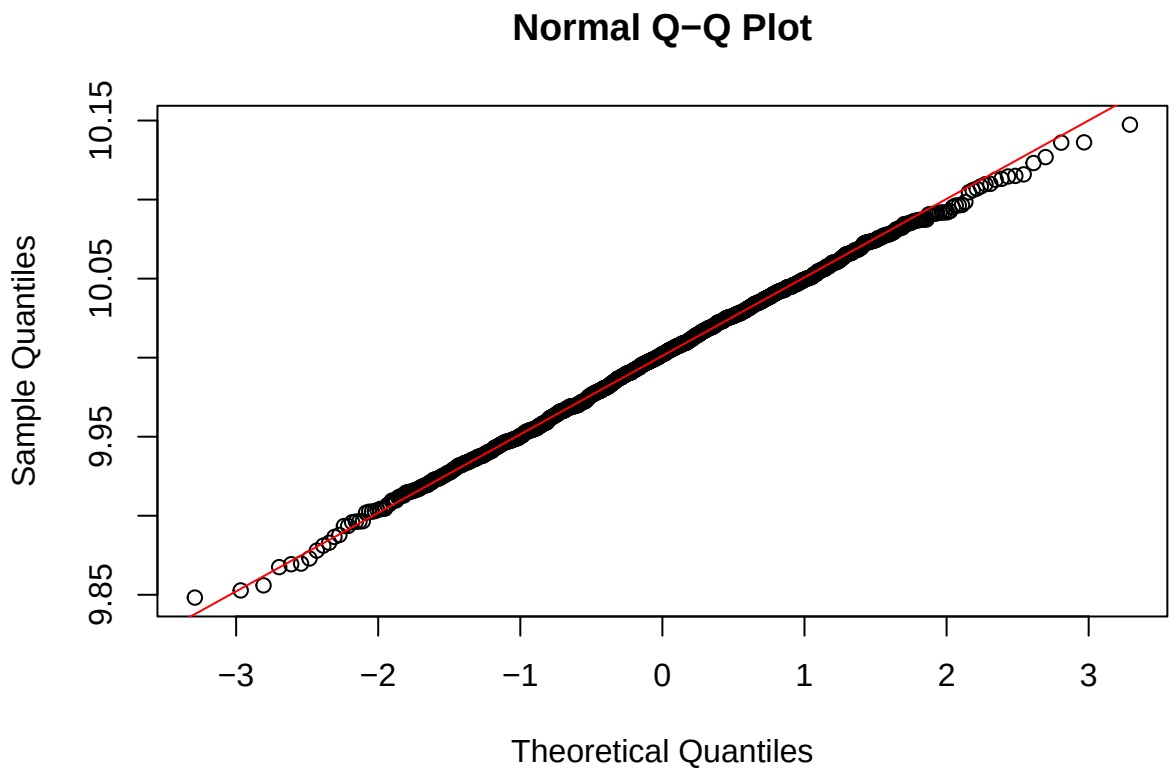
```
##    screensize_sample1 screensize_sample2
## 1            9.995571          10.018528
## 2            9.989146           9.973087
## 3           10.019307          10.027963
## 4           10.047096           9.964532
## 5            9.958848          10.063205
## 6            9.949121          10.020565
## 7           10.060778           9.995917
## 8            9.922951          10.001295
## 9            9.943103           9.932117
## 10          10.096234           9.967647
```

- Step 3: Check the assumptions mentioned earlier and the homogeneity of variance in addition.

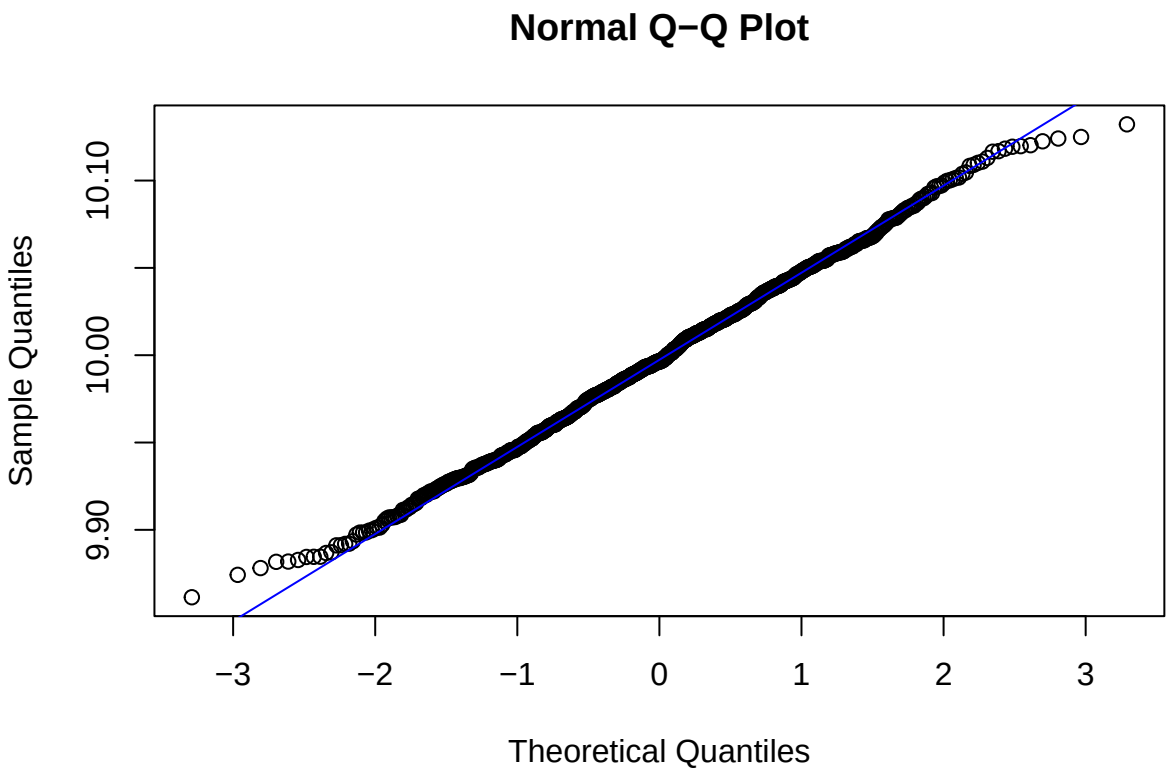
```
#Step 3 - Check for assumptions
#-------------------------------------------------------

#1. Data is continuous.
#2. Observations are randomly selected.
#3. To check the data is normally distributed, we will use the following commands:

qqnorm(data$screensize_sample1)
qqline(data$screensize_sample1,col="red")
```




```
qqnorm(data$screensize_sample2)
qqline(data$screensize_sample2,col="blue")
```



Also, in this case, we will check the homogeneity of variance:

```
#Homogeneity of variance
var(data$screensize_sample1)
```

```
## [1] 0.00238283
```

```
var(data$screensize_sample2)
```

```
## [1] 0.002353585
```

Great, the variances are equal. We can move ahead.

- Step 4: Conduct an independent two-sample t-test and report your answer.

```
#Step 4 - Conduct two-sample t-test

#Null Hypothesis: There is no difference between the mean of two samples
#Alternate Hypothesis: There is difference between the men of two samples

t.test(data$screensize_sample1,data$screensize_sample2,var.equal = T)
```

```
##
##  Two Sample t-test
```

```
##
## data:  data$screensize_sample1 and data$screensize_sample2
## t = 1.3072, df = 1998, p-value = 0.1913
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001423145  0.007113085
## sample estimates:
## mean of x mean of y
## 10.000976  9.998131
```

Note: Rewrite the above code with "var.equal = F" if you get unequal or unknown variances. This will be a case of **Welch's t-test** (we haven't seen it in detail in our course) which is used to **compare the means of two samples with unequal variances**.

```
#Step 4b - Conduct two-sample t-test

#Null Hypothesis: There is no difference between the mean of two samples
#Alternate Hypothesis: There is difference between the men of two samples

t.test(data$screensize_sample1,data$screensize_sample2,var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  data$screensize_sample1 and data$screensize_sample2
## t = 1.3072, df = 1997.9, p-value = 0.1913
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001423145  0.007113086
## sample estimates:
## mean of x mean of y
## 10.000976  9.998131
```

What can you infer from the above output?

We can confirm that the t-statistic is again less than the t-critical value so we fail to reject the null hypothesis. Hence, we can conclude that there is no difference between the mean screen size of both samples.

We can verify this again using the p-value. It comes out to be greater than 0.05, therefore we fail to reject the null hypothesis at a 95% confidence interval. There is no difference between the mean of the two samples.

Paired Sample t-test: Implementing the Paired t-test in R

We measure one group at two different times. We compare separate means for a group at two different times or under two different conditions.

Exercise 3

The manager of a tyre manufacturing company wants to compare the rubber material for two lots of tyres. One way to do this – check the difference between average kilometers covered by one lot of tyres until they wear out. You can download the data "**paired1.csv**" from Microsoft Teams.

Answer:

- Step 1: First, import the data.
- Step 2: Validate it for correctness in R (check dimensions & view first rows).

```
#Step 1 - Importing Data
#_____________________________________________________

#Importing the csv data
data<-read.csv(file = "./paired1.csv")

#data<-read.csv(file.choose()) # paired1.csv

#Step 2 - Validate data for correctness
#_____________________________________________________

#Count of Rows and columns
dim(data)
```

```
## [1] 25  2
```

```
#View top 10 rows of the dataset
head(data,10)
```

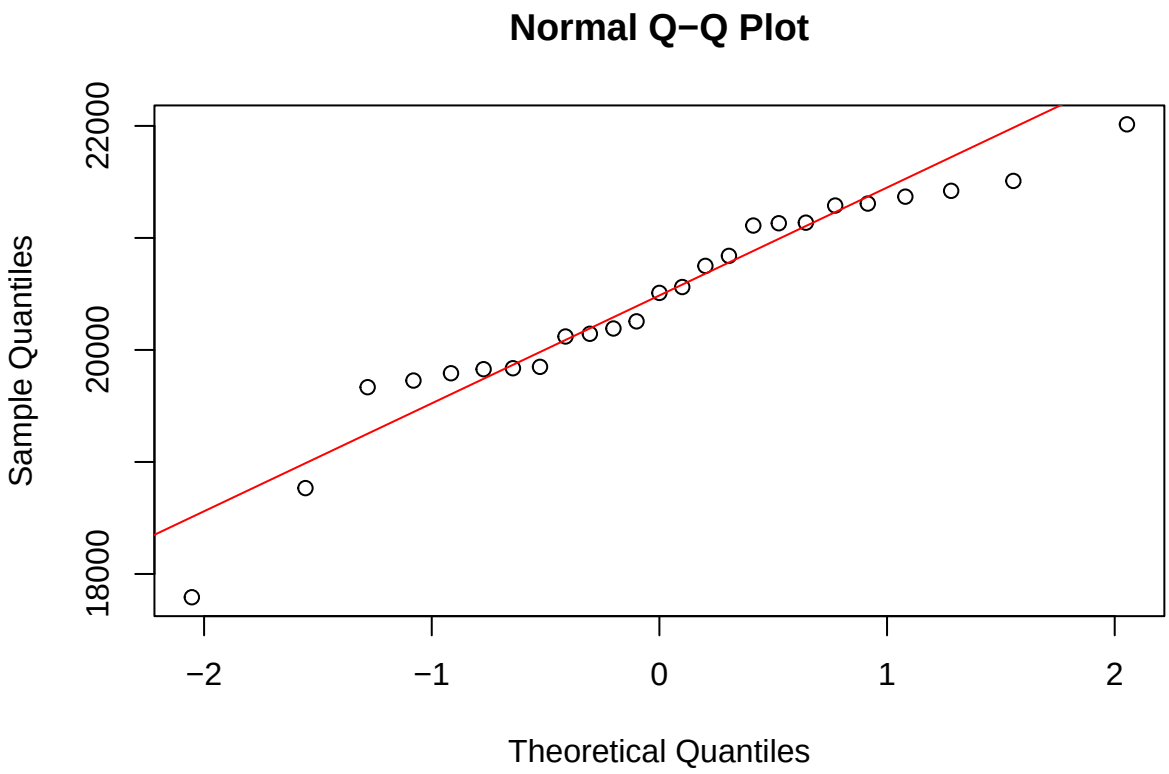
```
##      tyre_1   tyre_2
## 1  19849.13 21269.83
## 2  19836.81 23076.43
## 3  20750.42 21859.52
## 4  20191.55 24227.20
## 5  21131.22 21940.67
## 6  20120.14 21313.80
## 7  19726.70 22086.81
## 8  20561.69 21859.07
## 9  21289.19 19275.01
## 10 21420.56 23060.54
```

- Step 3: We now check the assumptions just as we did in a one-sample t-test.

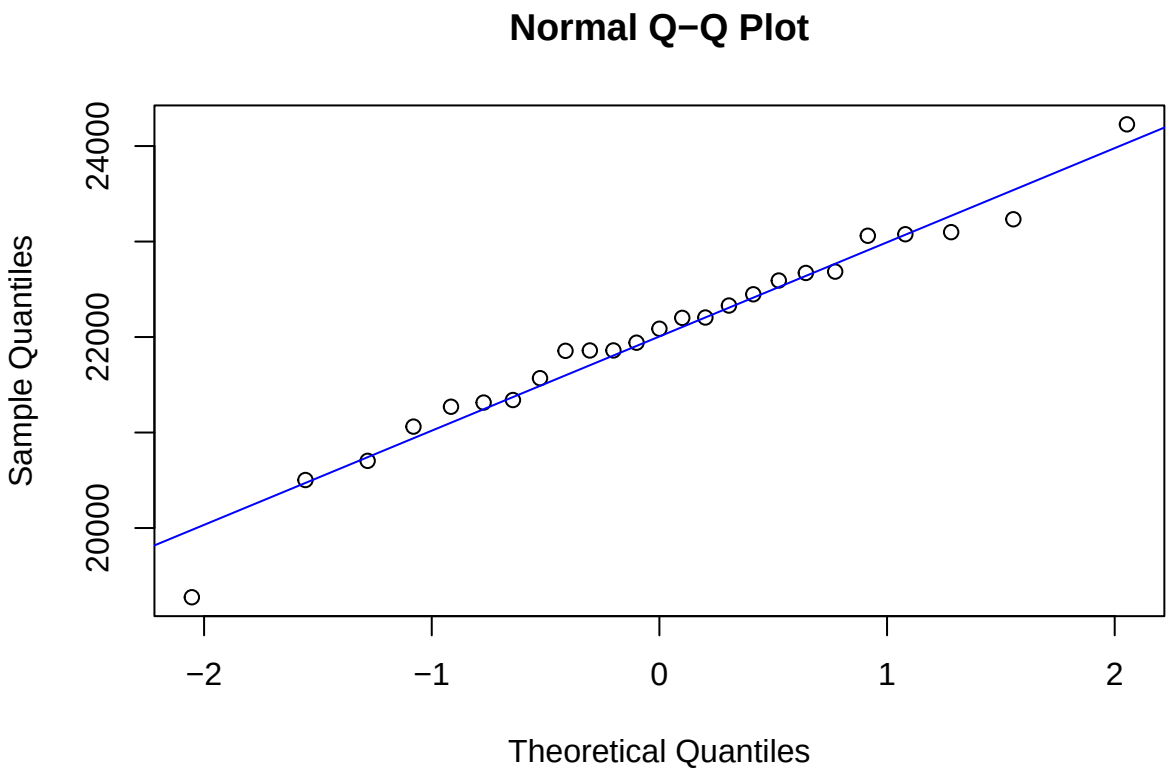
```
#Step 3 - Check for assumptions
#_____________________________________________________

#1. Data is continuous.
#2. Observations are randomly selected.
#3. To check the data is normally distributed, we will use the following codes:

qqnorm(data$tyre_1)
qqline(data$tyre_1,col="red")
```




```
qqnorm(data$tyre_2)
qqline(data$tyre_2,col="blue")
```



Step 4: Conduct the paired t-test and report your answer.

```
#Step 4 - Conduct two-sample t-test

#Null Hypothesis: There is no difference between the means of
# tyres before and after changing the rubber material.
#Alternate Hypothesis: There is a difference between the means
# of tyres before and after changing the rubber material.

t.test(data$tyre_1,data$tyre_2,paired = T)
```

```
##
##  Paired t-test
##
## data:  data$tyre_1 and data$tyre_2
## t = -5.2662, df = 24, p-value = 2.121e-05
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -2201.6929  -961.8515
## sample estimates:
## mean difference
##       -1581.772
```

The p-value is less than 0.05. We can reject the null hypothesis at a 95% confidence interval and conclude that there is a significant difference between the means of tyres before and after the rubber material replacement.

The negative mean in the difference depicts that the average kilometers covered by tyre 2 are more than the average kilometers covered by tyre 1.