# Statistics using R

Haseeb Raza

2024-07-11

## T-test in statistics

### One sample t-test & two sample t-test

#### To check the normal distribution ;

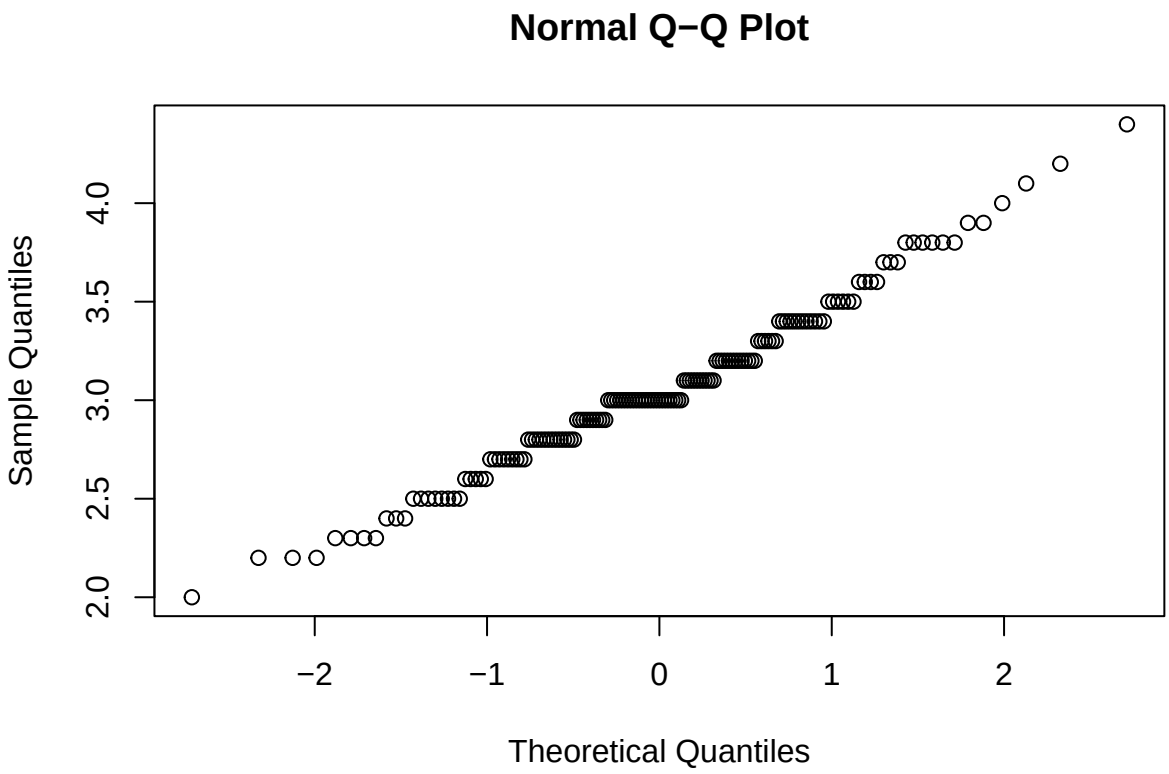
to check the normality distribution we can perform Shapiro Wilk test and Kolmogrov Simirnov Test, or we can draw a histogram and qqnorm plot,

```
df <- iris
shapiro.test(df$Sepal.Width)
```

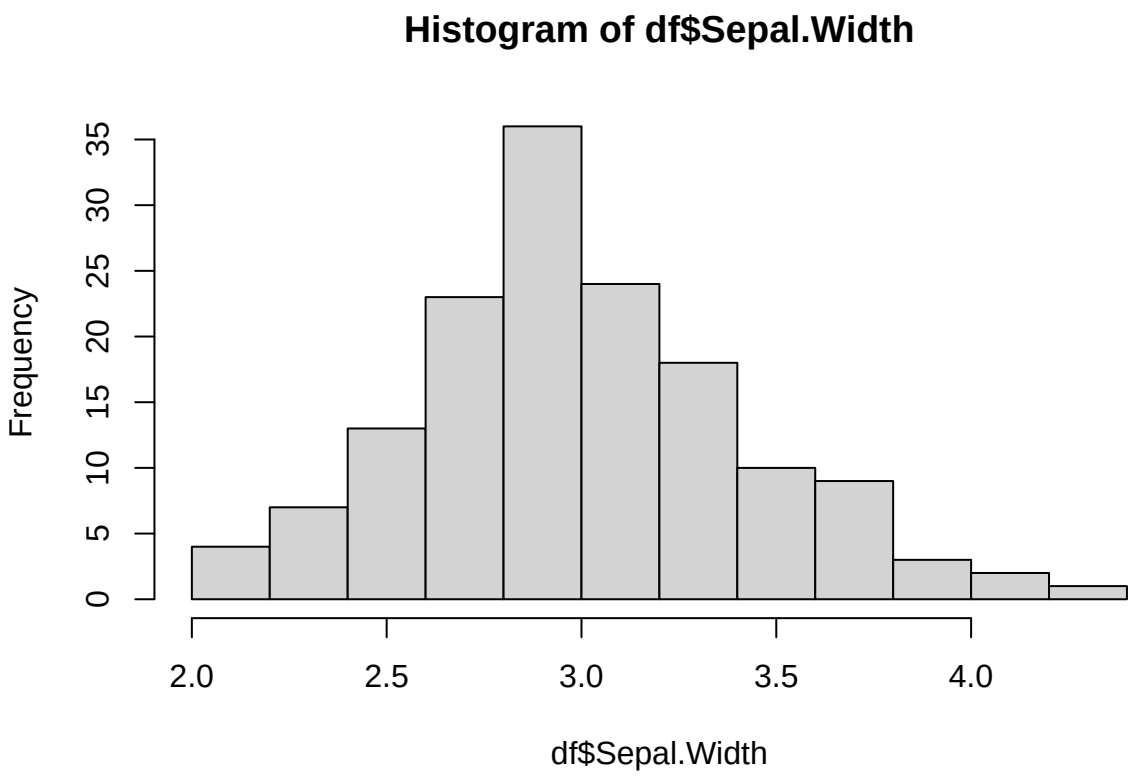
```
##
##  Shapiro-Wilk normality test
##
## data:  df$Sepal.Width
## W = 0.98492, p-value = 0.1012
```

so based upon the above result we can say that Sepal.Width in iris is normal.

```
qqnorm(df$Sepal.Width)
```




```r
hist(df$Sepal.Width)
```




**To check the composition of data;**

we can perform Leven's Test to check the composition of variable

```
library(car)
```

```
## Loading required package: carData
```

```
leveneTest(df$Sepal.Width ~ df$Species , data=df)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.5902 0.5555
##       147
```

## One sample T-test on Sepal Width Data

```
t.test(df$Sepal.Width , mu=5)
```

```
##
##  One Sample t-test
```

```
##
## data:  df$Sepal.Width
## t = -54.587, df = 149, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  2.987010 3.127656
## sample estimates:
## mean of x
##  3.057333
```

## Two sample T-test

There two types of two sample t-test - Unpaired T-test - Paired T-test

### Unpaired T-test

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
t.test(df$Sepal.Width ~ df$Species=="setosa" |df$Species=="versicolor")
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Sepal.Width by df$Species == "setosa" | df$Species == "versicolor"
## t = -1.8905, df = 135.22, p-value = 0.06083
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.255765152  0.005765152
## sample estimates:
## mean in group FALSE  mean in group TRUE
##               2.974               3.099
```

The above result indicates that these two variable has no significant difference.

```
library(tidyverse)
t.test(df$Sepal.Width ~ df$Species=="versicolor" |df$Species=="virginica")
```

```
##
##  Welch Two Sample t-test
##
## data:  df$Sepal.Width by df$Species == "versicolor" | df$Species == "virginica"
## t = 8.8121, df = 87.596, p-value = 1.035e-13
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  0.4306031 0.6813969
## sample estimates:
## mean in group FALSE  mean in group TRUE
##               3.428               2.872
```

According to above the result the variables are significantly different from each other.

### Paired T-test

For this test we have randomly created a data set having two thousand observations, with a standard deviation of ten.

```
pre_treatment <- c(rnorm(2000, mean = 140, sd=10 ))
tibble(pre_treatment)
```

```
## # A tibble: 2,000 x 1
##    pre_treatment
##            <dbl>
##  1          142.
##  2          132.
##  3          145.
##  4          147.
##  5          133.
##  6          142.
##  7          157.
##  8          144.
##  9          144.
## 10          126.
## # i 1,990 more rows
```

```
post_treatment <- c(rnorm(2000, mean = 170, sd=10))
tibble(post_treatment)
```

```
## # A tibble: 2,000 x 1
##    post_treatment
##             <dbl>
##  1           161.
##  2           159.
##  3           166.
##  4           169.
##  5           182.
##  6           187.
##  7           153.
##  8           175.
##  9           169.
```

```
## 10           163.
## # i 1,990 more rows
```

```r
t.test(pre_treatment, post_treatment, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  pre_treatment and post_treatment
## t = -95.981, df = 1999, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -30.28099 -29.06832
## sample estimates:
## mean difference
##       -29.67466
```

The result indicates that the averages of pre_treatment and post_treatment data differ significantly from each other.