Hypothesis Testing in R

Divit Vasu

7 August 2018

Hypothesis Testing

We are going to perform following hypothesis tests in this practical: * One Sample t-test * Two Sample t-test * Paired t-test

One sample t-test

One sample t-test is used to test whether the population mean is equal to the specified value or not.

Assumptions:

- The population from which, the sample drawn is assumed as Normal distribution.
- The population variance σ^2 is unknown.

Problem - One sample t-test

The specimen of copper wires drawn form a large lot have the following breaking strength (in kg. weight):

```r
strength <- c(578, 572, 570, 568, 572, 578, 570, 572, 596, 544)
wire_data <- data.frame(strength)
summary(wire_data$strength)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   544.0   570.0   572.0   572.0   576.5   596.0
```

Test (using Student's t-statistic) whether the mean breaking strength of the lot may be taken to be 578 kg. weight (Test at 5 per cent level of significance).

Let μ be the mean breaking strength of copper wires. The hypothesis testing problem is:

$H_0 : \mu = 578$ against $H_1 : \mu \neq 578$

Check the normality

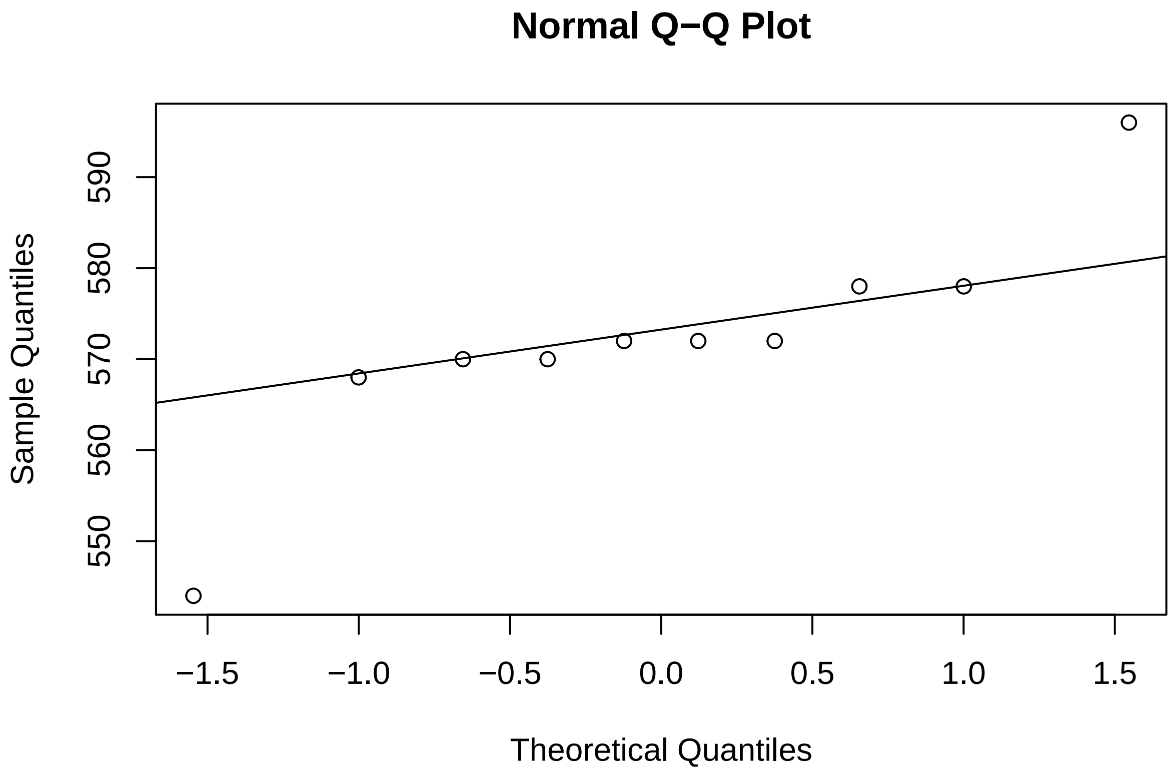
Shapiro-Wilks test is used to check the normality of the data.

```r
shapiro.test(wire_data$strength)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  wire_data$strength
## W = 0.84535, p-value = 0.05113
```

The p-value of the Shapiro-Wilk Normality test is 0.05113 which is not less than 0.05, we fail to reject the null hypothesis about the nomality. That is the strength of wires are normally distributed.

```
qqnorm(wire_data$strength) # draw qq plot
qqline(wire_data$strength) # add reference line
```



```
Result1<-t.test(wire_data$strength,mu=578,alternative="two.sided")
Result1 # display the result of t-test
```

```
##
##  One Sample t-test
##
## data:  wire_data$strength
## t = -1.4917, df = 9, p-value = 0.17
## alternative hypothesis: true mean is not equal to 578
## 95 percent confidence interval:
##  562.9012 581.0988
## sample estimates:
## mean of x
##       572
```

The test statistic is -1.4917 and the p-value is 0.17. As the p-value (0.17) is greater than 0.05 (level of significance) we accept the null hypothesis at 0.05 level of significance. There is enough evidence to support the null hypothesis that the mean strength of wires may be taken as 578 kg. weight.

Independent Sample t-test

Independent sample t-test is used to check whether there is statistically significant difference between the means in two independent groups.

Assumptions:

Assumptions for two sample t-test are as follows:

- The two samples are independently distributed
- The population from which, the two samples drawn are Normally distributed.
- The two population variances are unknown (equal or unequal).

Problem - Two sample t-test

Blood pressure data from two different group of patient which are using old drug and new drug respectively are given below:

```
drug_old <- c(90, 95, 67, 120, 89, 92, 100, 82, 79, 85)
drug_new <- c(71, 79, 69, 98, 91, 85, 89, 75, 78, 80)
summary(drug_old)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   67.00   82.75   89.50   89.90   94.25  120.00
```

```
summary(drug_new)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   69.00   75.75   79.50   81.50   88.00   98.00
```

Test at 5 per cent level whether the average blood pressure is the same between the drug_old and drug_new.

Let μ_1 and μ_2 be the mean weight of two groups. The hypothesis testing problem is: $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

Check the normality

Shapiro-Wilks test is used to check the normality of the data.

```
shapiro.test(drug_old)
```

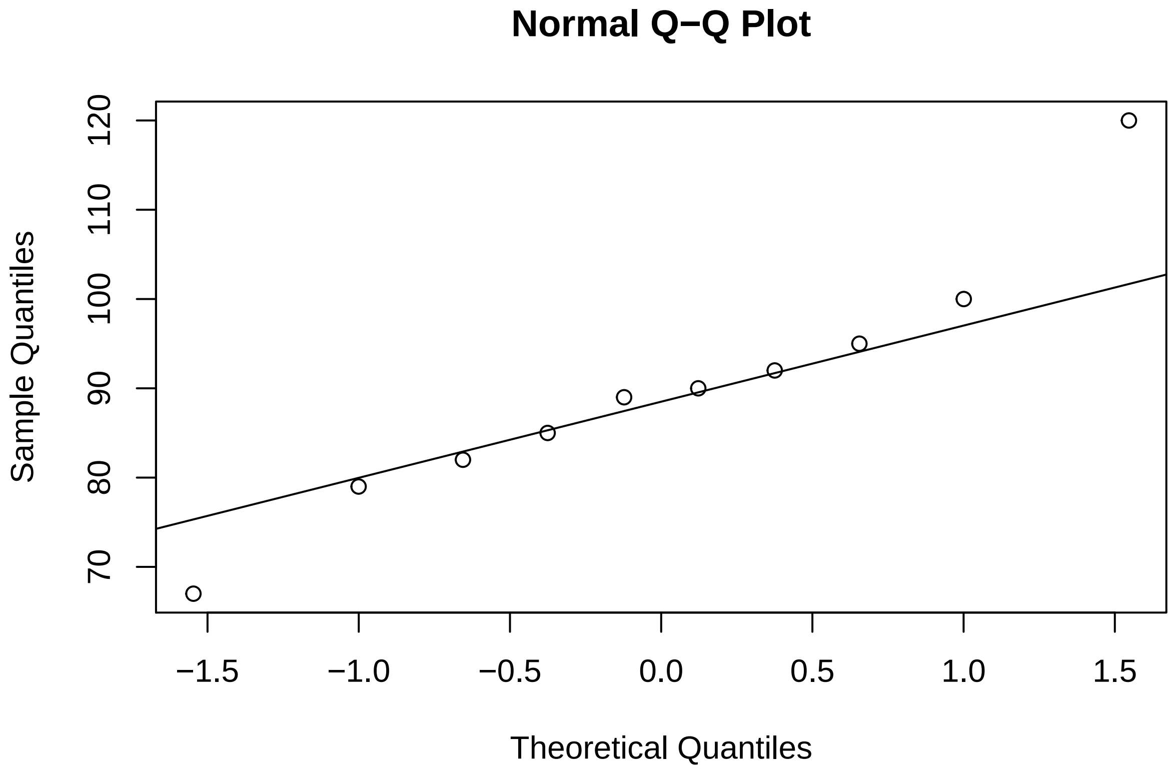
```
## 
##  Shapiro-Wilk normality test
## 
## data:  drug_old
## W = 0.95037, p-value = 0.6729
```

```
shapiro.test(drug_new)
```

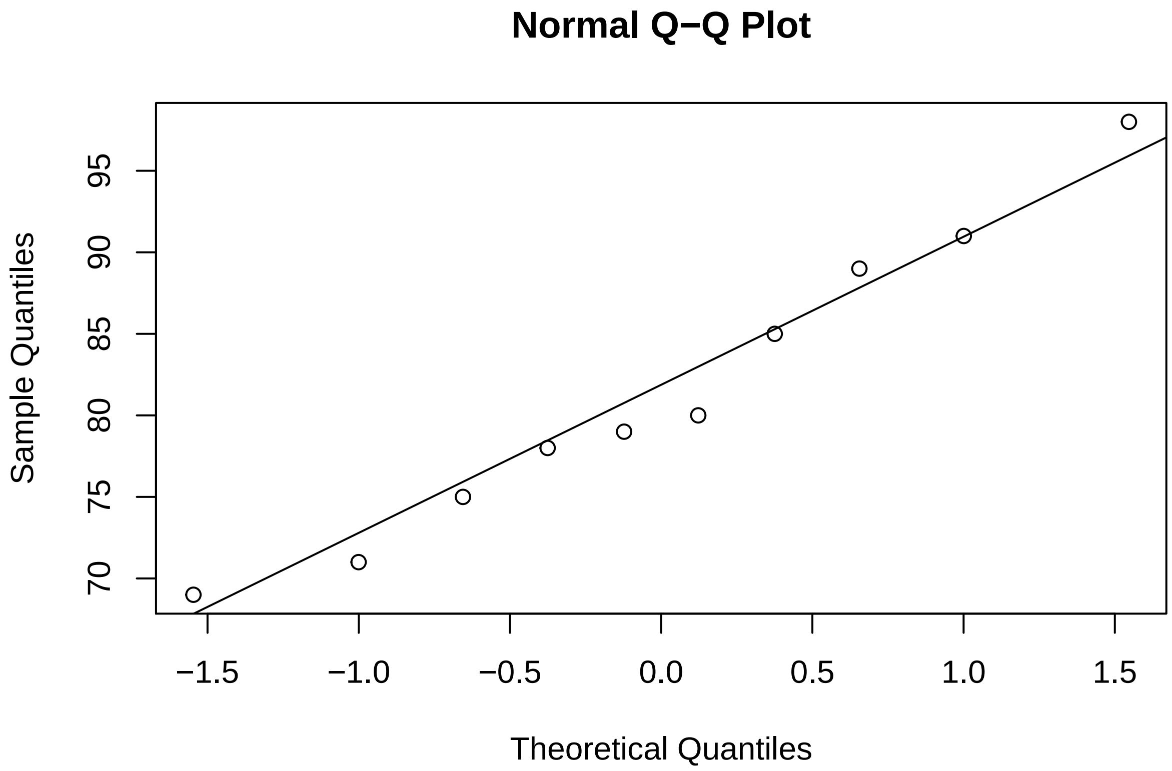
```
## 
##  Shapiro-Wilk normality test
## 
## data:  drug_new
## W = 0.96695, p-value = 0.8612
```

The p-value of the Shapiro-Wilk Normality test is 0.6729 for drug_old and 0.8612 for drug_new which are both not less than 0.05, we fail to reject the null hypothesis about the nomality. That is the weights of chickens are normally distributed.

```
qqnorm(drug_old) # draw qq plot
qqline(drug_old) # add reference line
```



```
qqnorm(drug_new) # draw qq plot
qqline(drug_new) # add reference line
```



```
vtest <- var.test(drug_old,drug_new,alternative = "two.sided")
vtest
```

```
##
##  F test to compare two variances
##
## data:  drug_old and drug_new
## F = 2.326, num df = 9, denom df = 9, p-value = 0.2246
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.577738 9.364341
## sample estimates:
## ratio of variances
##            2.32597
```

The p-value of the test is 0.2246, which is not less than 0.05, we fail to reject the null hypothesis about the equality of variances. There is enough evidence to support the null hypothesis that the mean blood pressure of both groups of patients may be considered equal. That is new drug has no significant impact.

Apply t test now

Paired t-test

The paired t-test is used when we have dependent samples. It is used to compare two population means in the case of dependent. Paired t-test for dependent samples is used in ‘before-after’ studies, or when the samples are the matched pairs.

Assumptions:

- The two samples are dependent.
- The difference between the two samples are independently distributed.
- The difference between the two samples are normally distributed.

Problem - Paired t-test

In a test given to two groups of students, the marks obtained were as follows:

```
first_group <- c(18, 20, 36, 50, 49, 36, 34, 49, 41, 30)
second_group <- c(29, 28, 26, 35, 30, 44, 46, 25, 32, 29)
```

Examine the significance of difference between mean marks obtained by students of the above two groups. Test at five per cent level of significance. Let μ_1 be the mean marks of first group of students and μ_2 be the mean mean marks of second group of students. The hypothesis testing problem is: $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 \neq 0$

```
summary(first_group - second_group)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  -12.00   -8.00    5.00    3.90   13.75   24.00
```

```
sd(first_group - second_group)
```

```
## [1] 13.27027
```

Check the normality

We wish to test H_0 : Distribution of the differece is normal against H_1 : Distribution of the difference is not normal.

```
resks <- ks.test(first_group, second_group, "pnorm")
```

```
## Warning in ks.test(first_group, second_group, "pnorm"): cannot compute
## exact p-value with ties
```

```
resks
```

```
## 
##  Two-sample Kolmogorov-Smirnov test
## 
## data:  first_group and second_group
## D = 0.4, p-value = 0.4005
## alternative hypothesis: two-sided
```

The p-value of the Kolmogorov-Smirnov Test is 0.4005 which is not less than 0.05. We fail to reject the null hypothesis about the nomality.

We conclude that the distribution of the difference between the marks of students of two different groups of same class is normally distributed.

```
res01 <- t.test(first_group, second_group, paired=T, alternative="two.sided")
res01
```

```
## 
##  Paired t-test
## 
```

```
## data:  first_group and second_group
## t = 0.92936, df = 9, p-value = 0.377
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.592978 13.392978
## sample estimates:
## mean of the differences
##                     3.9
```

The p-value of the test is 0.377, which is not less than 0.05, we fail to reject the null hypothesis about the equality of means.

There is enough evidence to conclude that there is no difference in the performance of students of both groups in same exam.

Conclusion

In this practical we learnt how to perform hypothesis testing in R. Specifically, we performed One Sample t-test, Two Samplet-test and paired t-test on various real life problems.