Hypothesis Testing using R and t-Test

Anthoniraj Amalanathan

t-test

- A t-test is a statistical test that is used to compare the means of two groups.
- It is often used in hypothesis testing to determine whether two groups are different from one another.

Hypothesis

A t-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population.

- The null hypothesis (H0) is that there is no difference or significance relationship between certain characteristics of a population.
- The alternate hypothesis (Ha) is that the true difference is different from zero.

Five Steps in Hypothesis Testing:

- Pick the random sample from the population
- Specify the Null Hypothesis.
- Specify the Alternative Hypothesis.
- Set the Significance Level (alpha)
- Calculate the Test Statistic and Corresponding p-Value.
- Drawing a Conclusion.

Significance level Vs Confidence level Vs Confidence interval

- Significance level (alpha): Measure of the evidence against the null hypothesis.
- Confidence level: Tells us more about how certain (or uncertain) we are about the true figure in the population. A confidence level = 1 - alpha.
- Confidence interval: A confidence interval displays the probability that a parameter will fall between a pair of values around the mean.

Real Example

- Arun : Will I get my promotion within 1 year?
- Manager : I am absolutely positive that you will get in 1 year
- Confidence level : Better than 95%
- Confidence Interval : Would be 1 year
- Keywords : certain (<99%), positively (<95%), unlikely (<5%)

p-value

Set the significance level— 0.01, 0.05, or 0.10. Compare the P-value to . If the P-value is less than (or equal to significance level), reject the null hypothesis in favor of the alternative hypothesis. If the P-value is greater than significance level, do not reject the null hypothesis.

One-sample, two-sample, or paired t-test?

- If the groups come from a single population (e.g. measuring before and after an experiment), perform a paired t-test.

- If the groups come from two different populations (e.g. two different marks of the students), perform a two-sample t-test (a.k.a. independent t-test).
- If there is one group being compared against a standard value (mathematics mark of students), perform a one-sample t-test.

Data Set

Marks secured by the students in high school Students from the United States. This data set consists of the marks secured by the students in various subjects.

```
df <- read.csv('StudentsPerformance.csv')
head(df[c(1,2,6,7,8)])
```

```
##   gender race.ethnicity math.score reading.score writing.score
## 1 female        group B         72            72            74
## 2 female        group C         69            90            88
## 3 female        group B         90            95            93
## 4   male        group A         47            57            44
## 5   male        group C         76            78            75
## 6 female        group B         71            83            78
```

Data set summary

```
str(df)
```

```
## 'data.frame':    1000 obs. of  8 variables:
##  $ gender                     : chr  "female" "female" "female" "male" ...
##  $ race.ethnicity             : chr  "group B" "group C" "group B" "group A" ...
##  $ parental.level.of.education: chr  "bachelor's degree" "some college" "master's degree"
"associate's degree" ...
##  $ lunch                      : chr  "standard" "standard" "standard" "free/reduced" ...
##  $ test.preparation.course    : chr  "none" "completed" "none" "none" ...
##  $ math.score                 : int  72 69 90 47 76 71 88 40 64 38 ...
##  $ reading.score              : int  72 90 95 57 78 83 95 43 64 60 ...
##  $ writing.score              : int  74 88 93 44 75 78 92 39 67 50 ...
```

Read male and female math score

```
dfm <- df[df$gender == 'male',c(1,6)]
dff <- df[df$gender == 'female', c(1,6)]
head(dff)
```

```
##    gender math.score
## 1  female         72
## 2  female         69
## 3  female         90
## 6  female         71
## 7  female         88
## 10 female         38
```

Null Hypothesis

H0: There is no significance difference between male and female students with respect to Mathematics marks.

Apply t-Test

```
result = t.test(sample(dfm$math.score,100), sample(dff$math.score,100), var.equal = T)
result
```

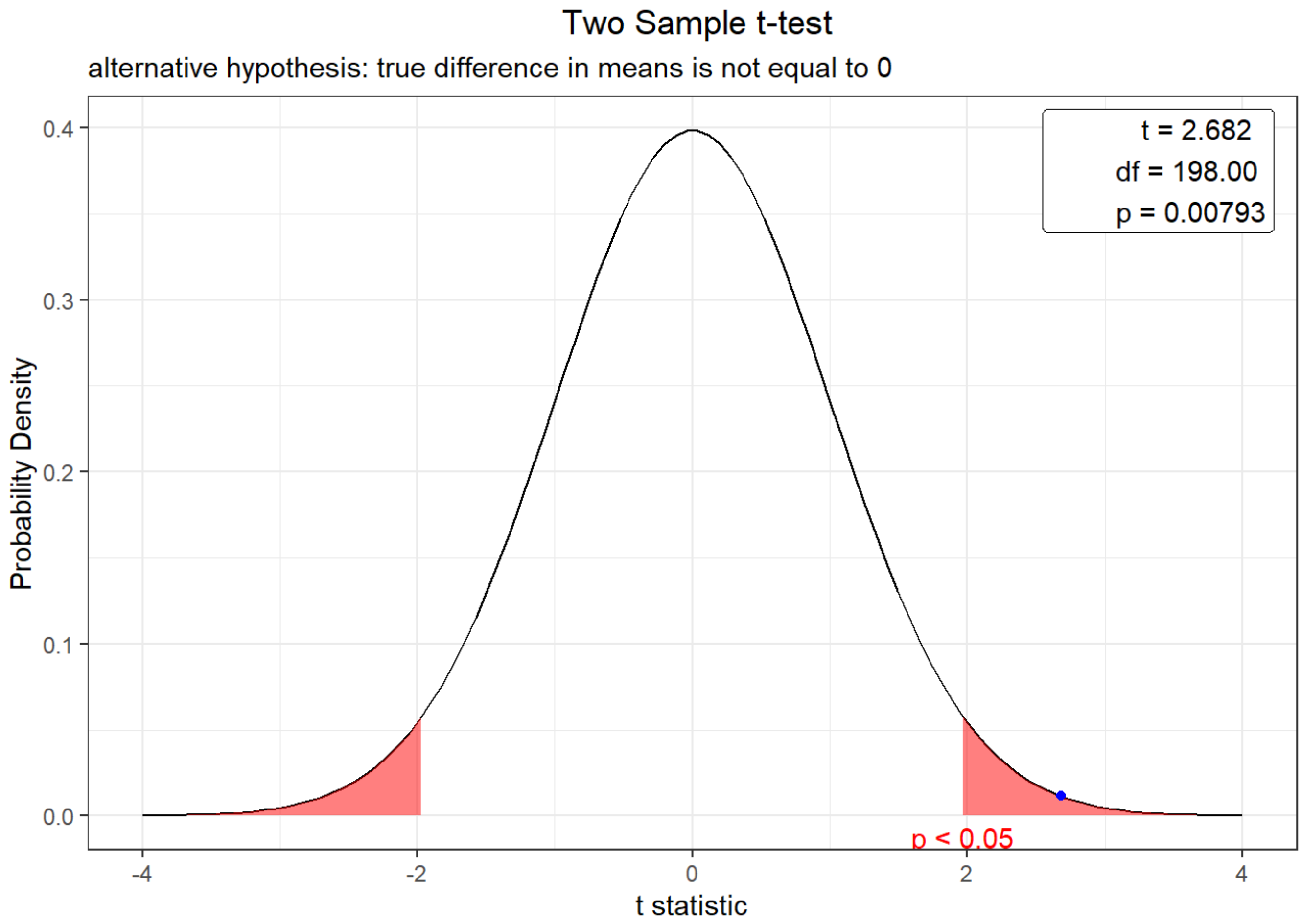
```
##
##  Two Sample t-test
##
## data:  sample(dfm$math.score, 100) and sample(dff$math.score, 100)
## t = 2.6824, df = 198, p-value = 0.007928
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.710784 11.209216
## sample estimates:
## mean of x mean of y
##     69.20     62.74
```

Ploting t-test

```
library(webr)
```

```
## Warning: package 'webr' was built under R version 4.1.1
```

```
plot(result)
```



### Observation * df - Degrees of Freedom (n-1 per sample) * p-vale * Alpha = 0.05

Conclusion

- Ha (Alternative Hypothesis): There is a significance difference between male and female students with respect to Mathematics marks.
- The male students have got higher marks than the female students.