# Shapiro-Wilk Test, F Test and t Test with sleep in R

Caio di Felice Cunha

## The Problem

Using the "sleep" dataset from the R package

Our aim in this study is to answer the following question:

"Is there a significant difference in the mean sleep of the 2 groups of patients, that is, is there a significant difference between the two drugs that help with the sleep disorder?

Since we have two samples (two groups), we can apply the t Test to answer the question. But to apply the t Test, we first need to validate its assumptions (explained in the previous item) and for that we need the Shapiro-Wilke Test of the F Test.

We define the hypotheses for our test as follows:

- H0 (Null Hypothesis) =There is no significant difference between the means of the 2 groups.
- There is (Alternative Hypothesis) =There is a significant difference between the means of the 2 groups.

The interpretation of the t-Test result will help define whether or not we should reject H0 and answer the business question of this case study.

## Stage 1 - Apllying the t Test

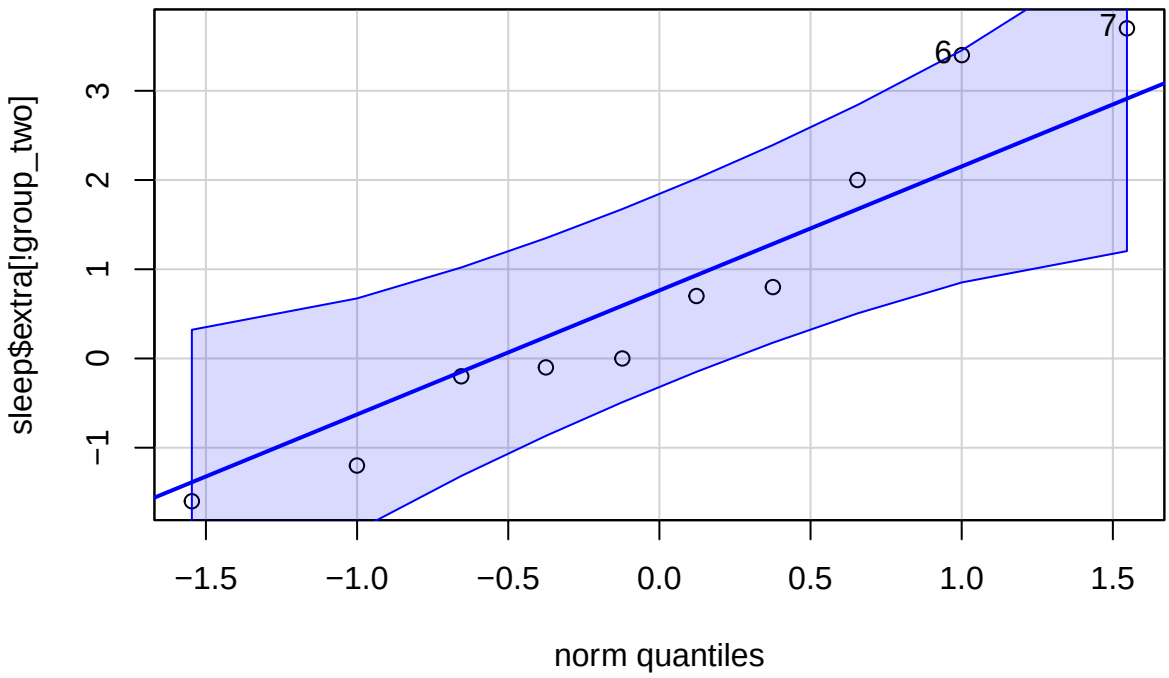
To apply the t Test first we need to validate the 5 test assumptions.

1- Data are random and representative of the population. 2- The dependent variable is continuous. 3- Both groups are independent (i.e. exhaustive and exclusionary groups). 4- The residuals of the model are normally distributed. 5- The residual variance is homogeneous (principle of homoscedasticity).

For our example in this case study, we will assume as true the assumptions 1 to 3 and we will validate assumptions 4 and 5. For assumption 4 we will use the Shapiro-Wilk Test and for assumption 5 we will use the F Test.

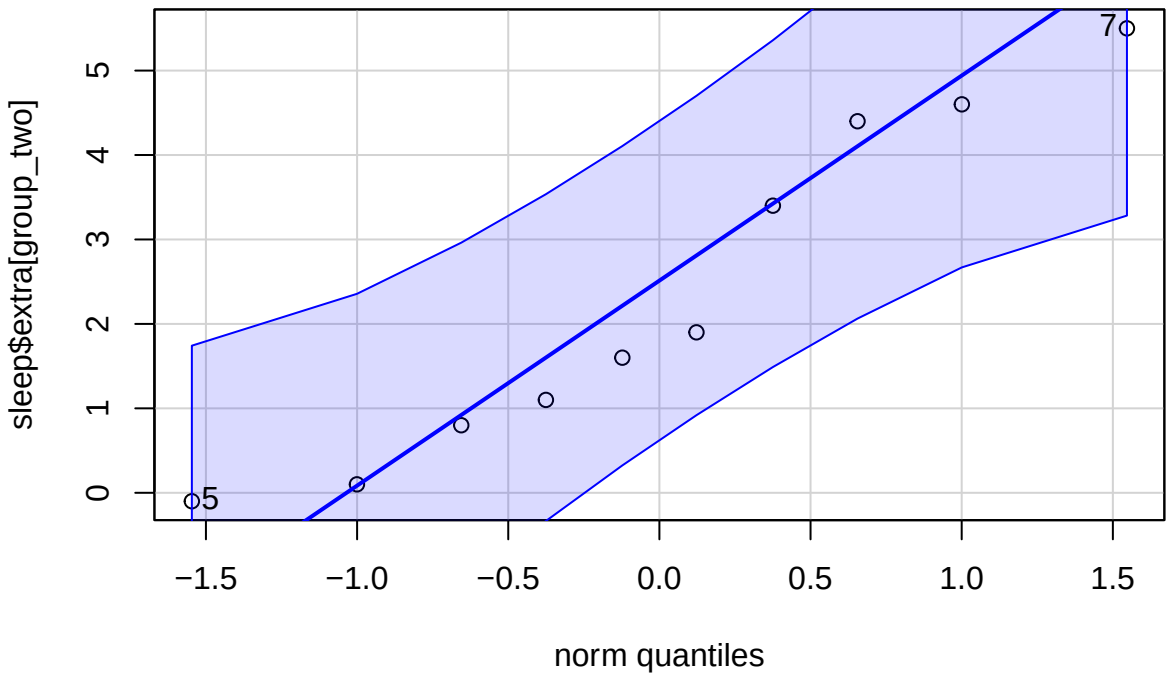
Let's extract data from one of the groups

```
# packages
library(car)
library(tidyverse)

group_two <- sleep$group == 2

# Validating Assumption 4 with qqPlot
qqPlot(sleep$extra[! group_two])
```




```
## [1] 7 6
```

```
# Validating Assumption 4 with qqPlot
qqPlot(sleep$extra[group_two])
```




```
## [1] 7 5
```

Analysis: The "extra" variable data points are within the confidence area, indicating that the data follow a normal distribution.

### Validating Assumption 4 with normality test shapiro.test()

To say that a distribution is normal, the p-value needs to be greater than 0.05. H0 = Data follows a normal distribution.

```
shapiro.test(sleep$extra[group_two]) # p-value = 0.3511 > 0.05
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sleep$extra[group_two]
## W = 0.9193, p-value = 0.3511
```

```
shapiro.test(sleep$extra[! group_two]) # p-value = 0.4079 > 0.05
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sleep$extra[!group_two]
## W = 0.92581, p-value = 0.4079
```

The test p-value of each group is greater than 0.05 and so we fail to reject H0. We can assume that the data follows a normal distribution.

## Stage 2 - Validating Assumption 5 with Test F

```
# First we check for missing values
colSums(is.na(sleep))
```

```
## extra group    ID
##     0     0     0
```

```
# Let's see a statistical summary of the dataset
sleep %>%
  group_by(group) %>%
  summarize(
    count = n(),
    mean = mean(extra, na.rm = TRUE),
    sd = sd(extra, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   group count  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 1        10  0.75  1.79
## 2 2        10  2.33  2.00
```

To reject the null hypothesis that the group means are equal, we need a high F-value. H0 = Means of data extracted from a normally distributed population have the same variance.

```
result_test_f <- var.test(extra ~ group, data = sleep)
result_test_f
```

```
##
##  F test to compare two variances
##
## data:  extra by group
## F = 0.79834, num df = 9, denom df = 9, p-value = 0.7427
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.198297 3.214123
## sample estimates:
## ratio of variances
##          0.7983426
```

The p-value is 0.7427, so greater than 0.05. We failed to reject H0. There is no significant difference between the variances of the 2 groups.

Validated assumptions. Now we can apply the t Test.

## Stage 3 - Apply the t Test

We apply the t Test to answer the question: H0 (Null Hypothesis) – There is no significant difference between the means of the 2 groups

```
# First we check for missing values
t_test_result <- t.test(extra ~ group, data = sleep, var.equal = TRUE)
t_test_result
```

```
##
##  Two Sample t-test
##
## data:  extra by group
## t = -1.8608, df = 18, p-value = 0.07919
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -3.363874  0.203874
## sample estimates:
## mean in group 1 mean in group 2
##            0.75            2.33
```

## Stage 3 - Conclusion

The p-value of the test is 0.07919, so greater than 0.05. We failed to reject H0. We can conclude that the 2 groups have no significant difference. There is no significant difference between the drugs applied to treat sleep disorders.

## Disclaimer:

Disclaimer: a good part of this project was largely done in the Data Science Academy, Big Data Analytics with R and Microsoft Azure Machine Learning course (part of the Data Scientist training)

## End