Tutorial 2 : hypothesis testing tests

I hope this tutorial will help you master hypotheses testing. Don't hesitate to ask me questions if you need further clarification. That's what I'm here for.

Learning objectives:

Familiarize yourself with R's functions for run hypothesis testing, as well as well as the steps to do to check applications conditins.

I. Link between one qualitative variable and one quantitative variable

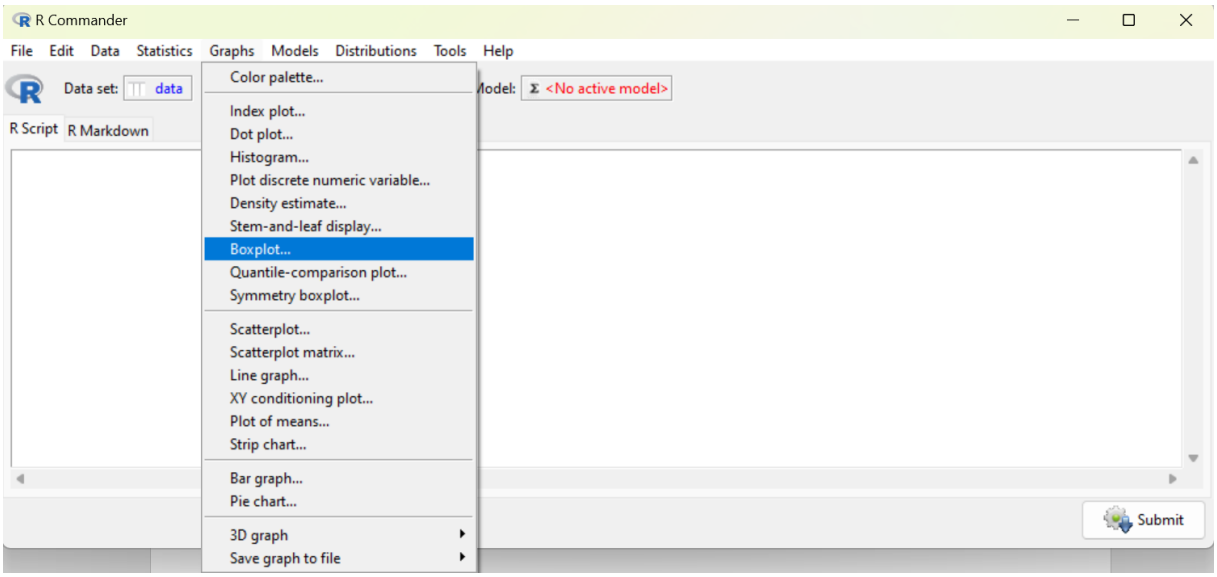
1. Data loading

Download the dataset « Wand_HP.csv ».

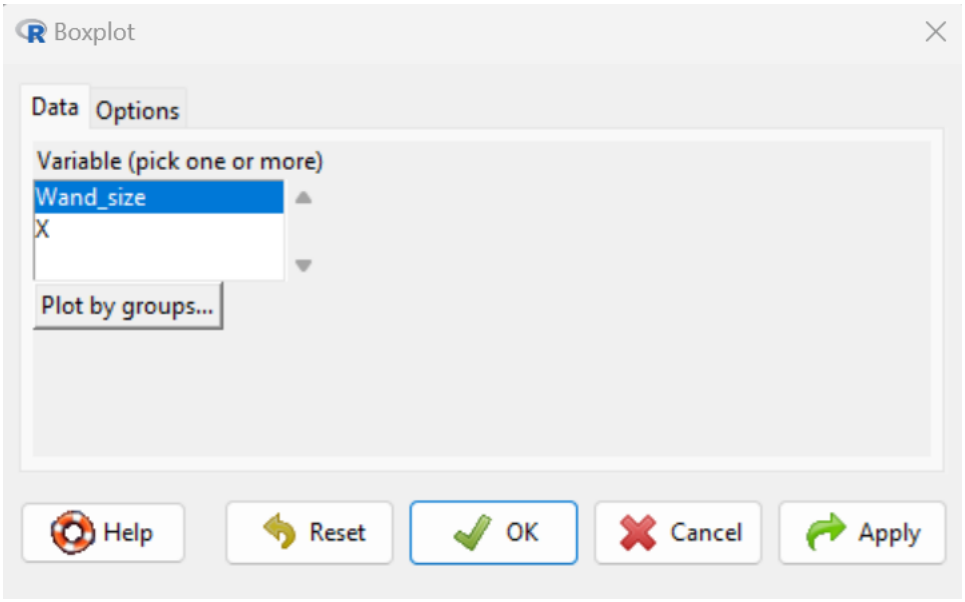
2. Data visualization

The first step is to visualize the data with a graph. In this case, we want to visualize Wand size as a function of gender. In other words, we're trying to visualize a quantitative variable as a function of a qualitative variable. So we're going to create a boxplot.

BUTTON CLICK CODING

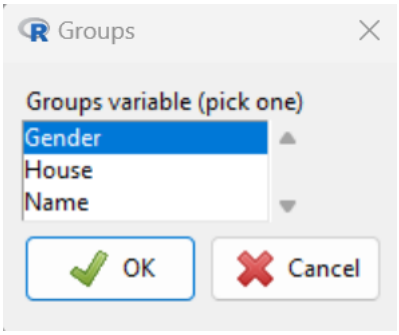


The quantitative variable to be selected is Wand size.



Then click on « plot by groups ».

Then select Gender.



Then click on OK.

CODING TEAM

```
boxplot(Wand_size ~ Gender, data = HP,col = c("purple", "gray"))
```

I'll leave it to you to deduce the presence of extreme data or not.

3. Check for normality

You need to check the normality of the data by gender. The first step is to make two sub-tables, one for girls and one for boys.

For this step, everyone is in the CODING Team for everyone.

To create subsets in R, here is the code (given dataset name is HP) :

```
HP_F=HP[which(HP$Gender=="Female"),]
HP_M=HP[which(HP$Gender=="Male"),]
```

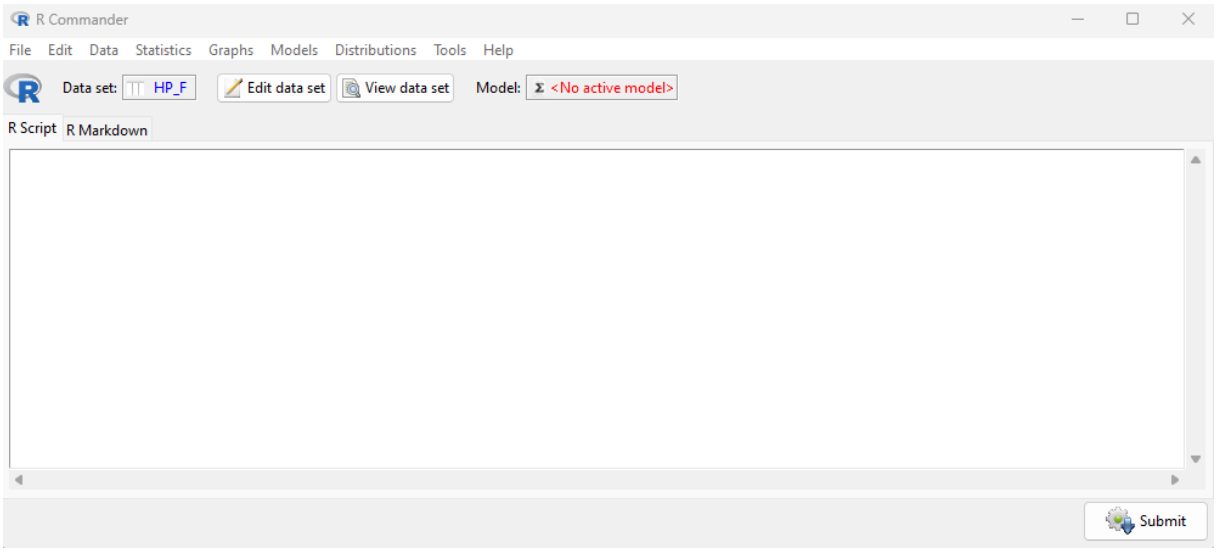
To go further : In R, you can access specific subsets of a data frame using square brackets [,]. The expression before the comma specifies which rows to select, while the expression after the comma specifies which columns to select. If you leave the part after the comma blank, R will include all columns by default. The which() function in R is particularly useful for row selection. It returns the indices of all rows that satisfy a given condition.

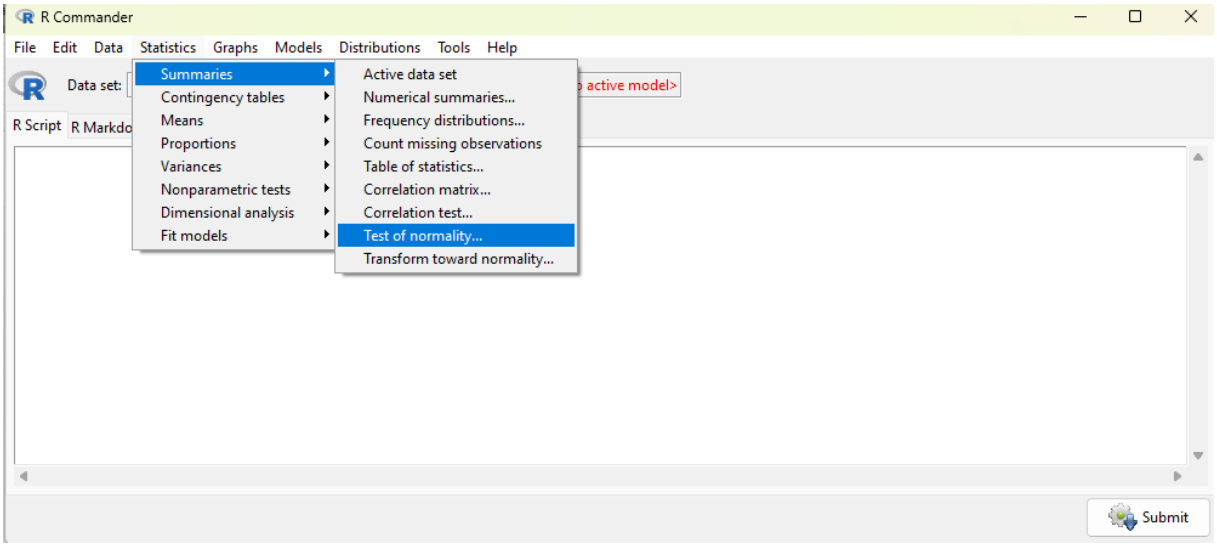
Once the two sub-tables have been generated, there are several ways to check the normality of the data.

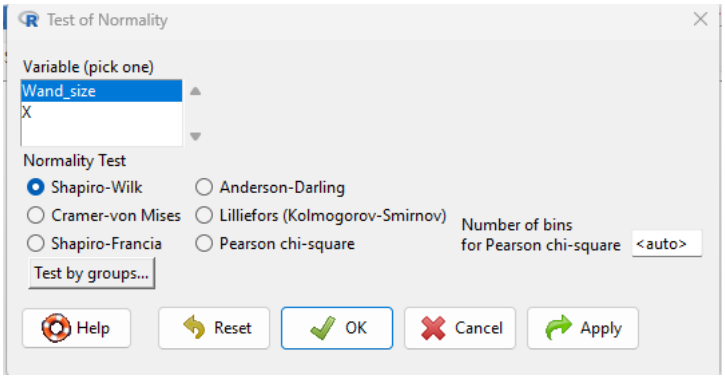
3.1 With shapiro test.

CLICK BUTTON TEAM

Don't forget to select the right dataset on Rcmdr (HP_F vs HP_M).







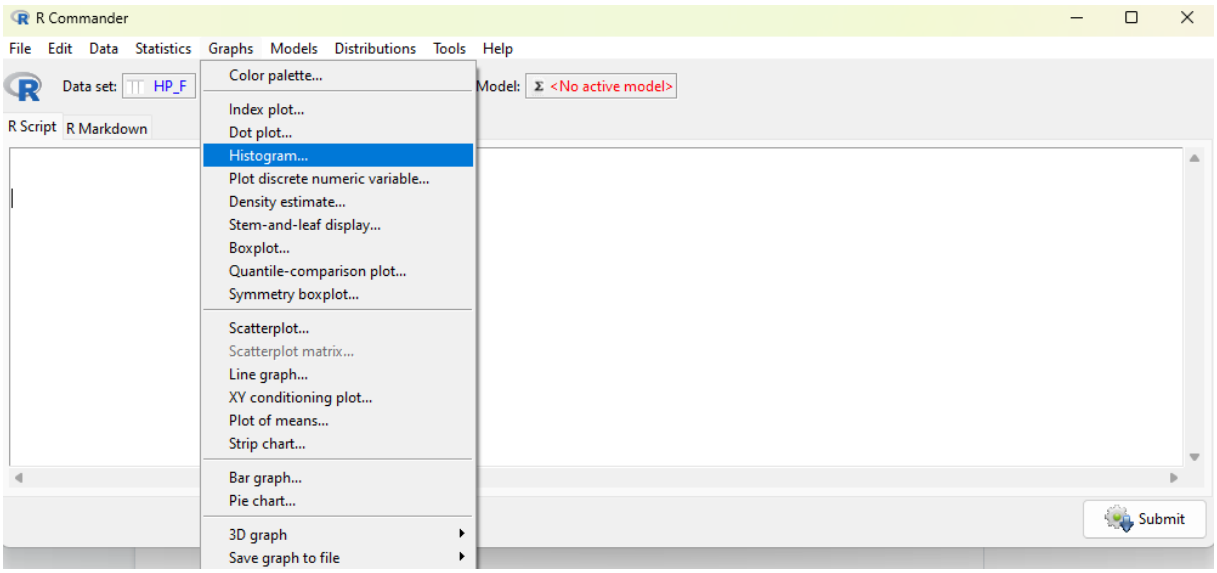
CODING TEAM

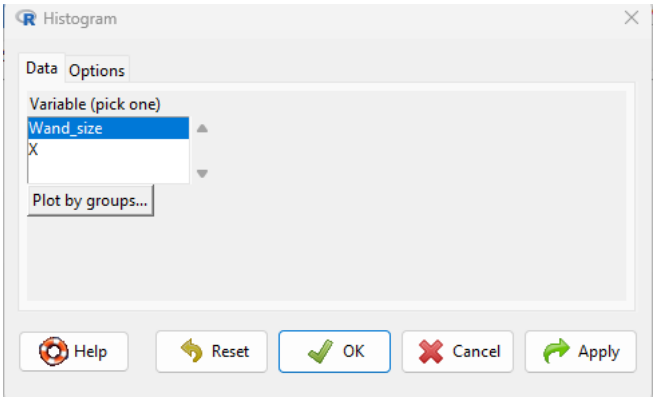
```
shapiro.test(HP_F$ Wand_size)
```

Do the same for males.

3.2 Graphically with an histogram

CLICK BUTTON TEAM





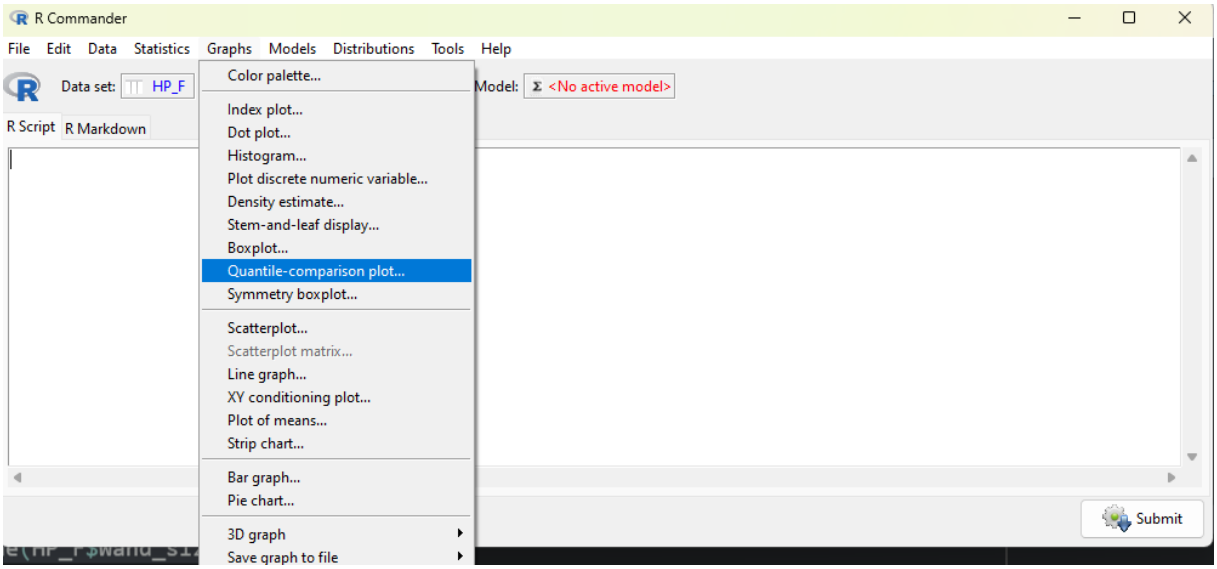
CODING TEAM

```
hist(HP_F$Wand_size, breaks = 10, prob = TRUE, main = "Histogramme de données")
```

Do the same for males.

3.3 Graphically with a qqplot

CLICK BUTTON TEAM



CODING TEAM

```
qqnorm(HP_F$Wand_size)
qqline(HP_F$Wand_size)
```

Do the same for male.

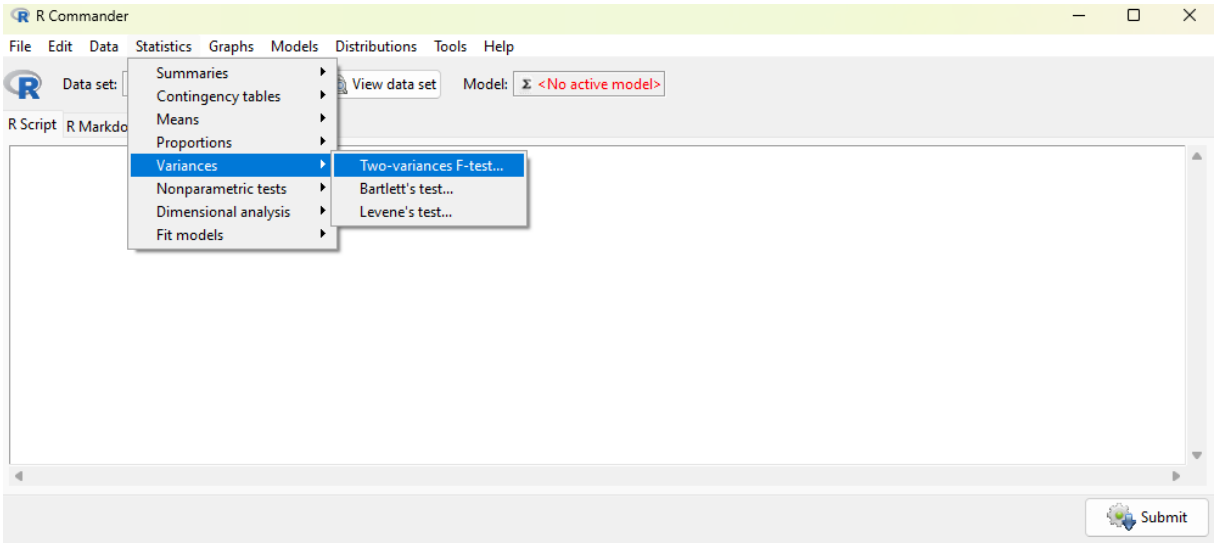
What do you conclude from these three approaches? Are the data distributions normal in each of the two groups?

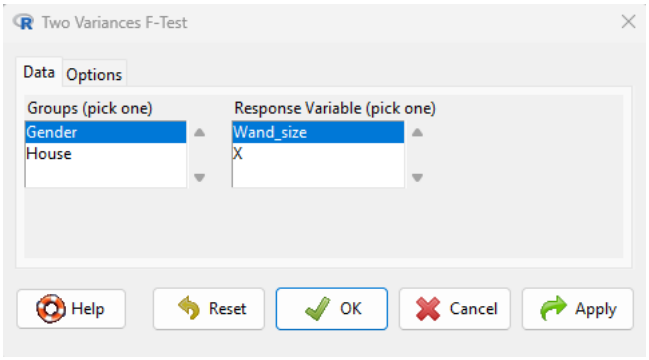
Conclude

4. I verify the homoscedasticity hypothesis (homogeneity of variances)

Note that we will use the whole dataset, not the subset.

CLICK BUTTON TEAM





CODING TEAM

```
var.test(Wand_size ~ Gender, data = HP)
```

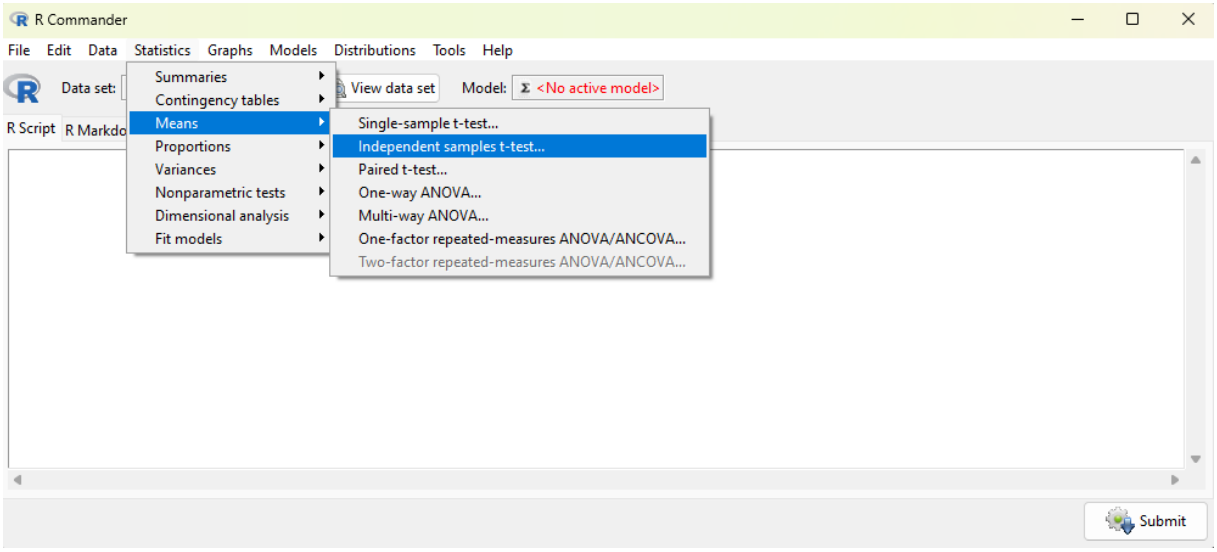
Conclude

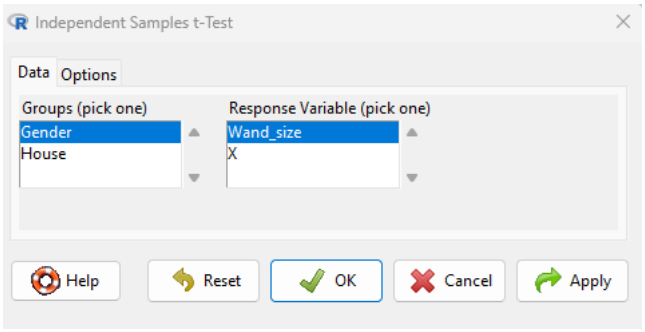
5. T-test run

There are two possible options: t-test or Wilcoxon test. I'll leave it to you to choose the right one.

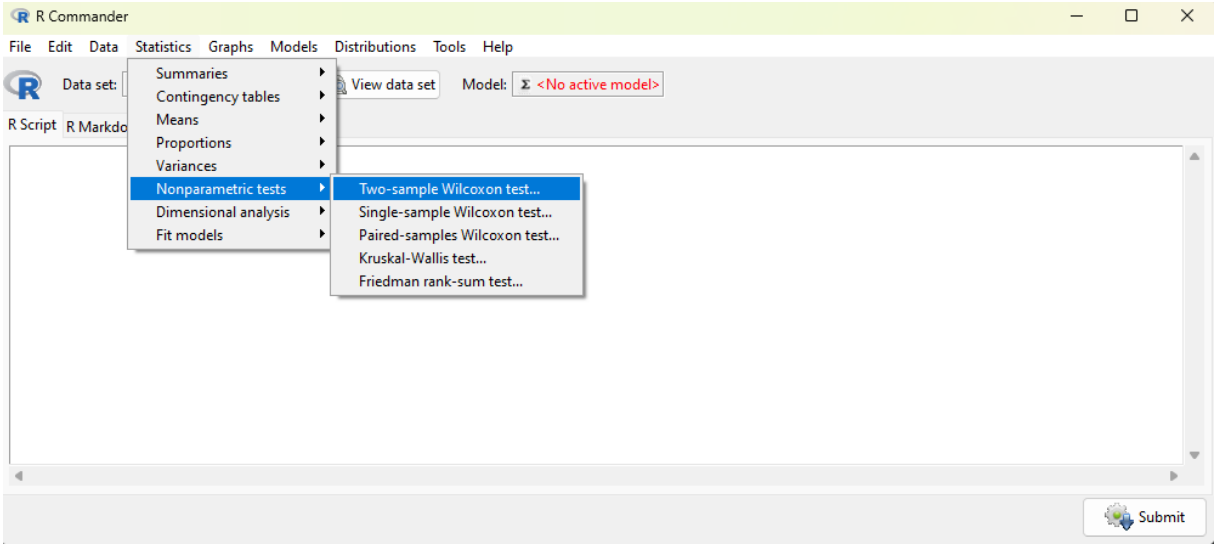
TEAM CLIC BOUTON

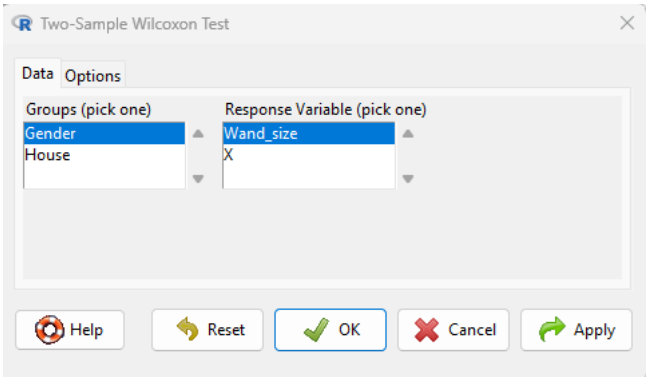
To perform a t-test :





To perform a Wilcoxon test





CODING TEAM

```
t.test(Wand_size ~ Gender, HP)

wilcox.test(Wand_size ~ Gender,HP)
```

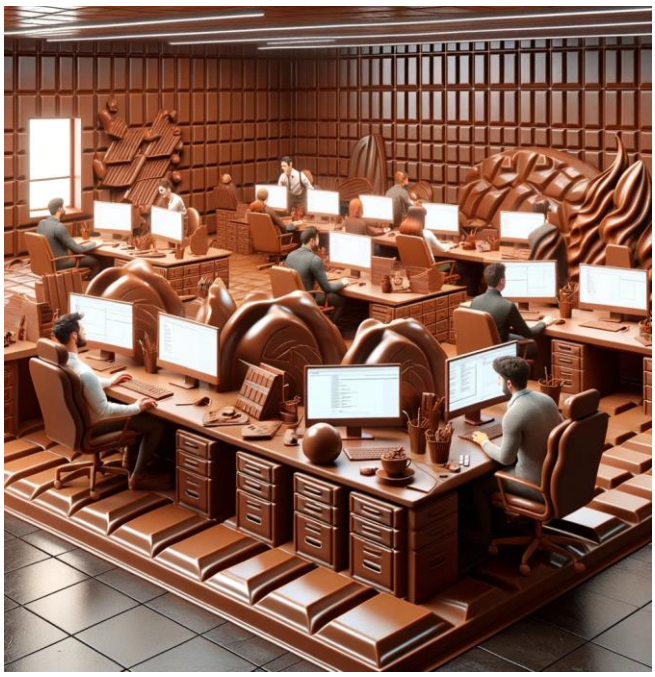
Conclude

Is there a significant difference between the two groups?

Which group has a higher baguette detail?

Is this size difference large or not?

II. Link between 2 quantitative variables



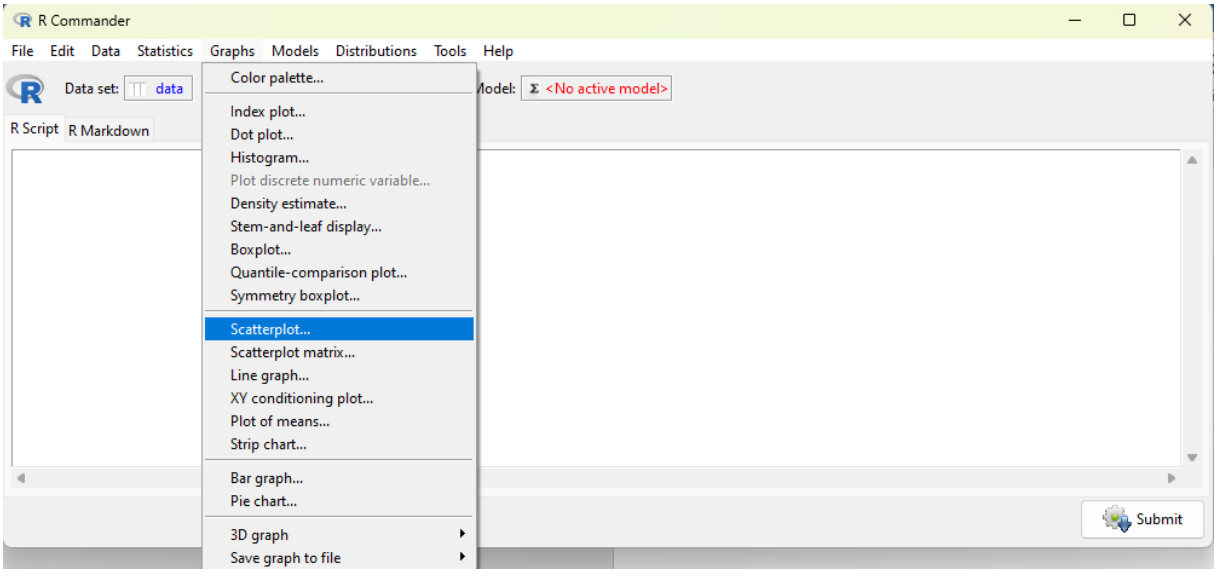
1. Data loading

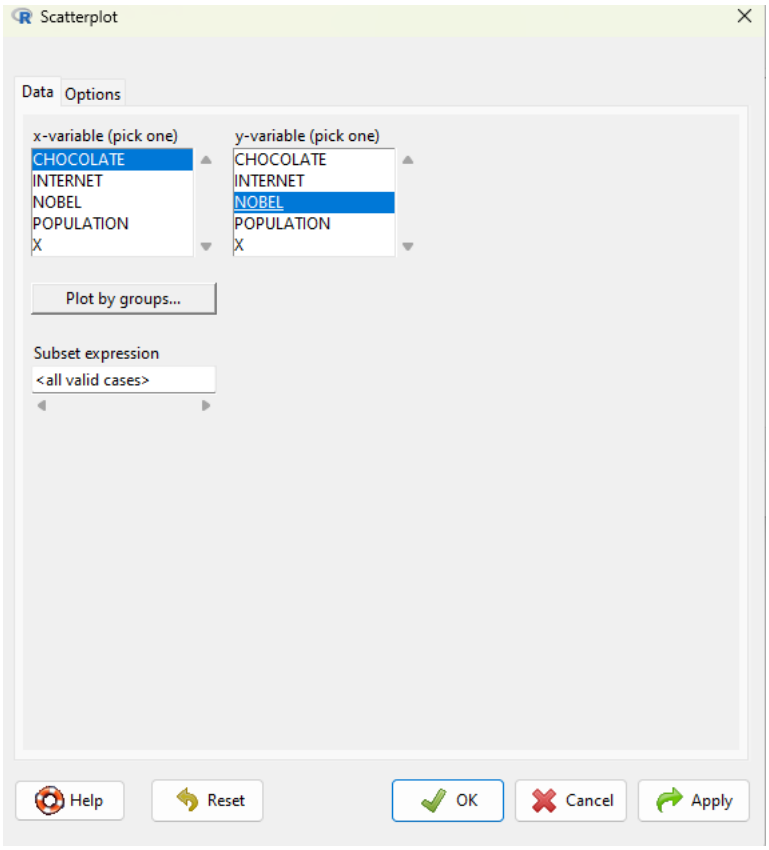
Download the dataset « chocolate_nobel.csv ».

2. Data visualization

Construct a scatter plot of the number of Nobel Prize winners as a function of chocolate consumption.

CLICK BUTTON TEAM





CODING TEAM

```
plot(data$NOBEL~data$CHOCOLATE)
```

Check for outliers by making a boxplot for each variable (procedure seen in part 1). Yes, you can make a boxplot on a single quantile variable.

3. Normality check

I check the normality of the distribution of the two variables. Here are the 3 options that you have seen in part 1 :

- Shaporo test
- Histogram
- qqplot

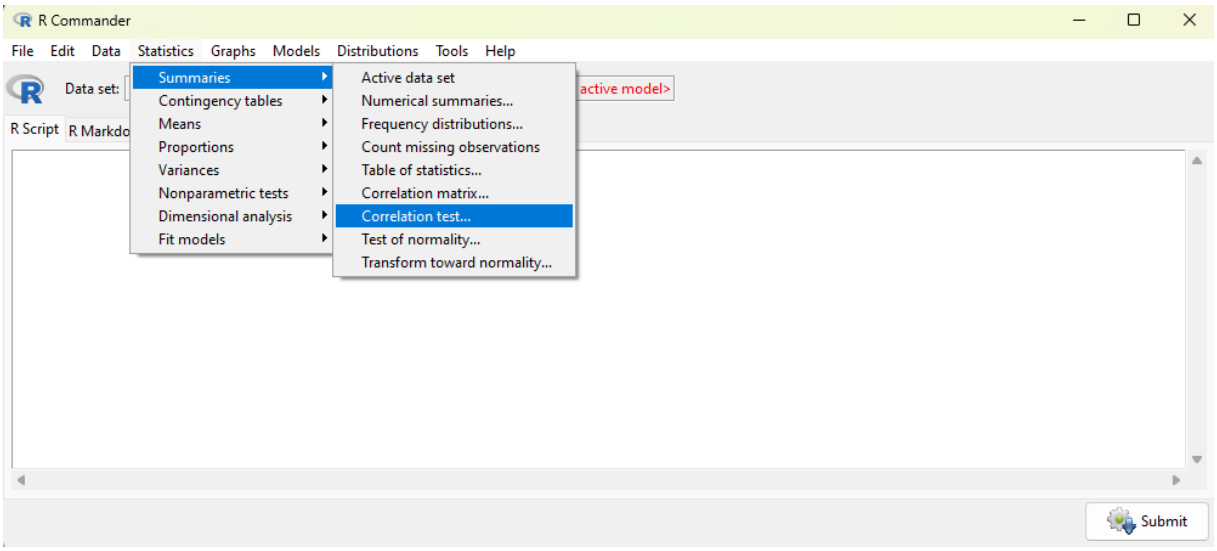
What do you conclude from these three approaches? Are the data distributions normal in each of the two variables?

4. Correlation test

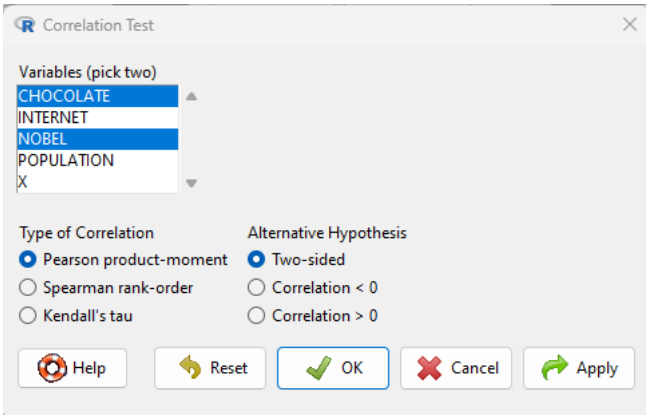
Should you use a Spearman or Pearson test?

Here are both options :

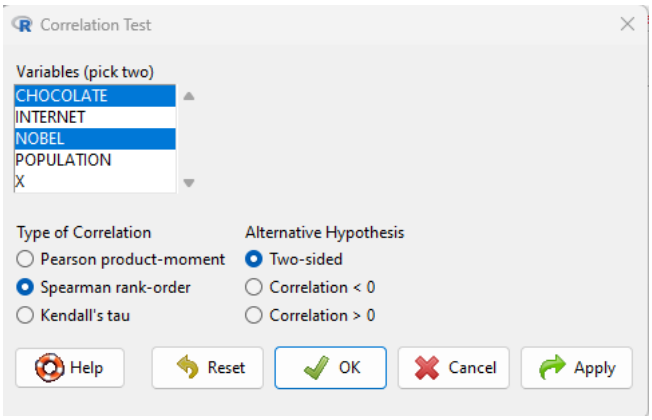
CLICK BUTTON TEAM



Then for pearson correlation test :



For Spearman correlation test :



TEAM CODAGE

Both options for both pearson correlation and spearman correlation.

```
cor.test(data$CHOCOLATE, data$NOBEL, method = "pearson")
cor.test(data$CHOCOLATE, data$NOBEL, method = "spearman")
```

Adapt.

Conclude

Is there a significant correlation between the number of and chocolate consumption by country?

Is this correlation positive or negative?

Is the correlation strong or weak?

III. Link between 2 qualitative variables

Only CODING TEAM for this part.

5. Data loading

Download the dataset « titanic.csv ».

6. Contigency table extraction

```
contingency_table <- table(titanic$Pclass, titanic$Survived)
```

7. Graphs

```
# Create frequency data frame from contingency table
data_freq <- as.data.frame(contingency_table)
names(data_freq) <- c("Pclass", "Survived", "Freq")

# Plot for Pclass vs. Frequency by Survival status
barplot(Freq ~ Pclass + Survived, data = data_freq, beside = TRUE,
        col = c("cornsilk2", "black"), legend.text = c("Survived", "Not Survived"),
        args.legend = list(title = "Survival", x = "topright"),
        main = "Count of Passengers by Pclass and Survival",
        ylab = "Count", xlab = "Pclass", ylim = c(0, max(data_freq$Freq) + 10))
```

8. Chi² test and theoretical counts extraction

```
chi_square_test <- chisq.test(contingency_table)
chi_square_test
chi_square_test$expected
```

9. In case, Fisher test's code

```
fisher.test(contingency_table)
```