SMDE FIRST ASSIGNMENT (20% OF THE FINAL MARK, INDIVIDUAL)

FIRST QUESTION: VISUALISATION, CHI SQUARE AND T-TEST (15% OF THE FIRST ASSIGNMENT).

Download the data set decathlon from the web site of Kaggle given in the link below and read it in R.
https://www.kaggle.com/drisskaouthar/decathlon#decathlon.csv

a) Analyze the distribution of "X100m" according to the type of competition by using boxplot. Write your conclusion.

b) Create a new categorical variable with two categories from the variable "X100m" by using 11 seconds as the cut-off point. Make a cross table from the new categorical variable and the "Competition". Are these two variables independent? Write your conclusion by checking marginal probabilities and test the independency of two variables by using Chi-Square test.

c) Visualize the distribution of quantitative variables by using proper graph. Which of these variables follow a Normal distribution?

d) Generate three Normally distributed random variables of length 50. Two of them should have the same mean, different standard deviations while the third one has a different mean but the same standard deviation with the first distribution. Use t test to compare mean differences between three variables.

e) Test if there is a difference between two type of competitions according to the variables "X100m" and "X400m" by using t test.

SMDE FIRST ASSIGNEMENT (20% OF THE FINAL MARK, INDIVIDUAL)

SECOND QUESTION: ANOVA (25% OF THE FIRST ASSIGMENT).

Generate three populations that follow a normal distribution, **using your own algorithm**. As an example, the first is a population that follows a normal distribution with a parameter mean=0, the second with mean=10, and the third with mean=0. Select the SAME variance for the three distributions at your convenience (a value >0).

We want to analyze using an ANOVA if these three populations are different (or not) depending on the parameter selected.

Analyze and explain the results obtained. Justify your answers.

Remember to test the ANOVA assumptions. What do you expect on the assumptions?

Once you finish the analysis and you are familiar with ANOVA test: on the dataset contained on Kaggle, named "Red and White Wine Quality", we want to analyze if in both (type or quality) affects some properties of the wine. After combining the two datasets (one for red wines and one for white wines), you should create two variables. First, "type" that identifies if the wine is red or white, and second, wine quality categorized in three groups: <5 (low), 5-6 (medium) and >6 (high). Once you complete preprocessing steps, please answer to the following questions applying appropriate statistical techniques:

1) Which of the chemical properties influence the quality of the wines?
2) Which of the chemical properties are related with type of the wines?
3) How does type and quality of wines affect (separately and together) percentage of alcohol present in the wine?
4) Detail the results of Two-Way ANOVA considering as dependent variable "fixed acidity", and independent variable "type" and "quality".

SMDE FIRST ASSIGNEMENT (20% OF THE FINAL MARK, INDIVIDUAL)

THIRD QUESTION: DEFINE A LINEAR MODEL FOR AN ATHLETE IN THE 1500 M (25% OF THE FIRST ASSIGMENT).

To start: load the package RCmdrPlugin.FactoMinerR.

Load the data “decathlon” located in the package.

The data represents a data frame with observations for different athletes.

What is the linear expression that better predicts the behavior of an athlete for 1500m?

Explore different expressions describing the power and the features of each one of them.

Justify your answers.

Remember to test the assumptions of the linear model.

Now use the expression to **predict** the behavior for a specific athlete.

Analyze and explain the results obtained.

Is the model accurate? What do you expect?

Justify your answers.

ANNEX: THE DISTINCTION BETWEEN CONFIDENCE INTERVALS, PREDICTION INTERVALS AND TOLERANCE INTERVALS.

When you fit a parameter of a model, the accuracy or precision can be expressed as (i) confidence interval, (ii) prediction interval or (iii) tolerance interval. Assume that the data really are randomly sampled from a Gaussian distribution.

Confidence intervals tell you about how well you have determined the mean. If you do this many times, and calculate a confidence interval of the mean from each sample, you'd expect about 95 % of those intervals to include the true value of the population mean. The key point is that the confidence interval tells you about the likely location of the true population parameter.

Prediction intervals tell you where you can expect to see the next data point sampled. Collect a sample of data and calculate a prediction interval. Then sample one more value from the population. If you do this many times, you'd expect that next value to lie within that prediction interval in 95% of the samples. The key point is that the prediction interval tells you about the distribution of values, not the uncertainty in determining the population mean.

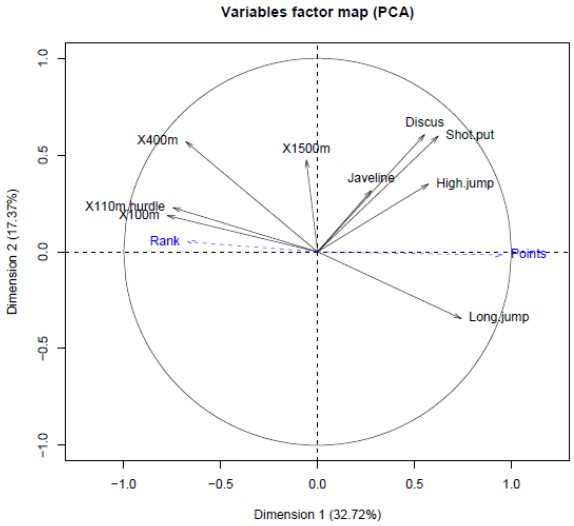
Prediction intervals must account for both the uncertainty in knowing the value of the population mean, plus data scatter. So a prediction interval is always wider than a confidence interval.

SMDE FIRST ASSIGNEMENT (20% OF THE FINAL MARK, INDIVIDUAL)

FOURTH QUESTION: WORKING WITH REAL DATA (25% OF THE FIRST ASSIGMENT)

FOR DECATHLON DATASET

Define a PCA for the decathlon dataset and discuss why the model you own works well (or not) looking the variables chart. We recommend using FactoMineR package to do this analysis.



Be focused on the Variables chart to understand, at a glimpse, the relations between the different variables of interest.

PRINCIPAL COMPONENT REGRESSION

Now we are going to continue with principal component regression.

We want to construct a linear regression model to predict the points of each athlete. To do so you must first decide the number of principal components to be included in the regression as independent variables. Justify your answer.

Check the assumptions of the regression model.

Is the prediction accurate enough?