Iris data  
dr. maryam al-berry  
2017-2018

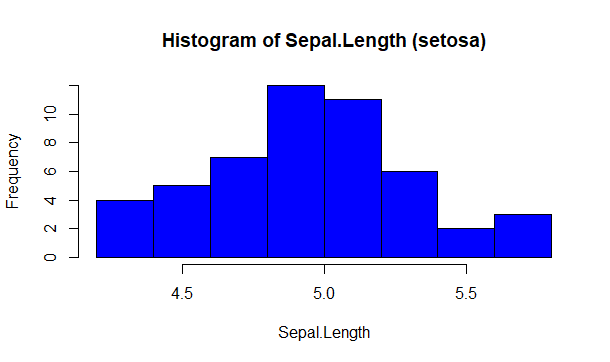
BY:  
Mirna Muhammad  
Esraa Yasser

# Introduction

In this project we decided to analyze some of the *Iris setosa* species data (Fisher's Iris data set) including the Sepal Length of an *Iris setosa* flower and its Petal Length, studying the correlation between them, and determining the Sepal Length population mean of the flower by hypothesis tests. All of this was implemented using R-programming language and this semester’s statistical analysis course.

Step 1. Descriptive Statistics

Using R-programming Language we were able to draw 2 kinds of plotting schemes:

* “Boxplot” for the Sepal Lengths, and the Petal Lengths of a Setosa flower.
* “Histogram” for the same data previously mentioned.

Sepal.Length Petal.Length

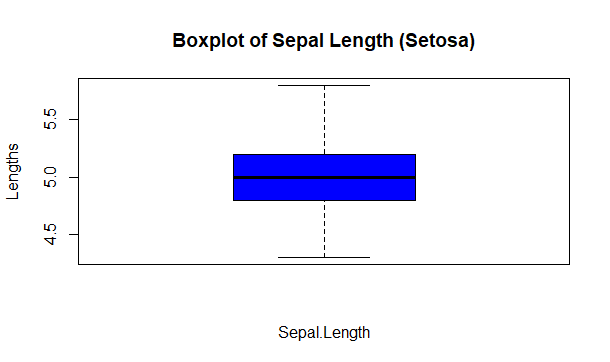
Min. :4.300 Min. :1.000

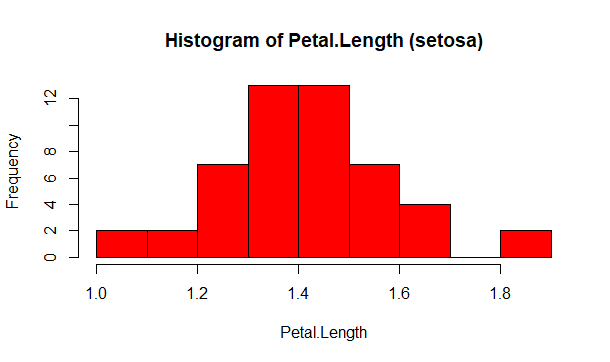
1st Qu.:4.800 1st Qu.:1.400

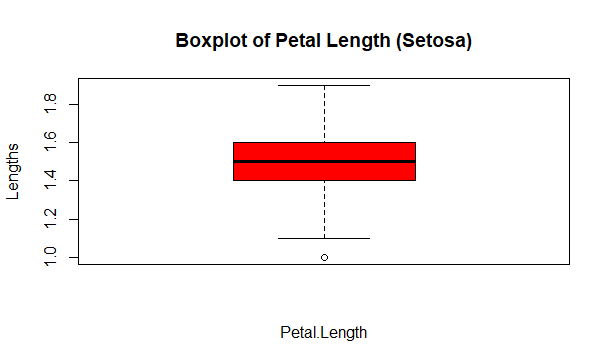
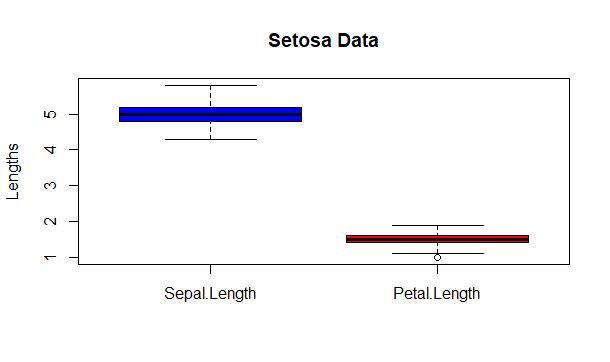
Median :5.000 Median :1.500

Mean :5.006 Mean :1.462

3rd Qu.:5.200 3rd Qu.:1.575

Max. :5.800 Max. :1.900



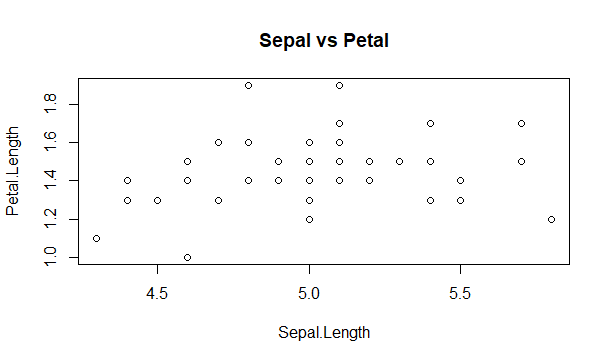


Step 2. Correlations

We wanted to see if there’s a correlation between a setosa flower sepal length and its petal length using Pearson’s **r**, and by plotting the scatterplot.

Pearson’s **r** = 0.2671758 (weak positive linear relation)

Which leads us to indicate the value of the Coefficient of Determination **r2** = 0.07138289

The correlation between the two variables is weak, **r2** = 0.07138 this means that only 7.13% of the variation in Petal Lengths depends on the variation in Sepal Length. So, we can’t estimate the value of one of them based on the other.   
  
  
  
  
  
  
  
  
  
  
  
  
 Scatterplot between the Sepal and Petal Lengths of Setosa flowers

Step 3. Inferential Statistics

We tested the hypothesis of our Sepal Length Data using inferential tests to determine the population’s average sepal length using the sample mean of the *setosa* species. A one-sided test is suitable because we're interested in knowing whether the length is less than 5 cm.  
  
The test has the null hypothesis that the average sepal length is equal to 5 cm, and the alternative hypothesis that the average sepal length is less than 5 cm. A significance level of 0.05 is to be used.  
Ho: µ = 5.0  
H1: µ < 5.0

data: sepal.Length  
t = 0.12036, df = 49, p-value = 0.5477  
alternative hypothesis: true mean is less than 5  
95 percent confidence interval:  
 -Inf 5.089575  
sample estimates:  
mean of x   
 5.006

From the output, we can see that the average sepal length for the sample is 5.006 cm.  
The one-sided 95% confidence interval tells us that the average sepal length is likely to be less than 5.0896 cm.

Since the p-value is not less than the significance level of 0.05, we cannot reject the null hypothesis that the population mean sepal length is equal to 5 cm. But our question of interest that the population’s average sepal length being less than 5 cm is rejected (µ < 5.0).

We used t-test as it can still be applied to larger samples and as the sample size **n** grows larger and larger, the results of a t-test and z-test become closer and closer. In the limit, with infinite degrees of freedom, the results of t and z tests become identical. ([*reference1*](https://en.wikibooks.org/wiki/Statistics/Testing_Data/t-tests))