# Univariate Descriptive Statistics

Bachelor in Computer Science and Engineering

2020/21

## 1. Introduction

The objective of this document is to present the most used techniques of Descriptive Statistics to resume the information of a dataset of one variable. The data we are going to analyze in this handbook are stored in the file `AlumnosIndustriales.xlsx`. These data correspond to 95 students of Industrial Engineering, to whom it was asked about some variables such as height, weight, number of siblings, and another seven variables. In this way we are going to use a simple dataset that will help us to learn the basic descriptive function of R.

First we read and view the data file. The figure shows the first five observations of this datafile. Note that the line `View (IndustrialStudents)` appears as a comment, to execute it, simply delete the symbol `#`.

```
library(readxl)
AlumnosIndustriales <- read_excel("AlumnosIndustriales.xlsx")
#View(AlumnosIndustriales)
```

AlumnosIndustriales

Filter

| | nacimiento | altura | peso | zapato | sexo | dinero | tiempo | locomocion | residencia | hermanos | Variables |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 180 | 72 | 44 | 1 | 1100 | 35 | 3 | 1 | 2 | Nacimiento: mes de nacimiento |
| 2 | 1 | 161 | 55 | 39 | 0 | 287 | 45 | 4 | 1 | 2 | Sexo: 1 chico 0 chica |
| 3 | 1 | 180 | 45 | 41 | 1 | 2000 | 100 | 4 | 3 | 2 | Dinero: ptas en el bolsilo |
| 4 | 1 | 180 | 99 | 44 | 1 | 25 | 40 | 3 | 2 | 2 | Tiempo: en llegar a la universidad |
| 5 | 1 | 178 | 68 | 41 | 1 | 3225 | 40 | 1 | 3 | 2 | Hermanos: que tiene |

## 2. Description of categorical variables

The variable `residencia` corresponds to the home location of the students. This variable is categorical. The set of values it can have is

- Madrid Sur (1)
- Madrid Centro (2)
- Madrid-otros (3)
- Fuera de Madrid (4)

To describe this variable we first obtain a frequency table.

```
ABStable <- table(AlumnosIndustriales$residencia)
lbls <- c("Madrid Sur", "Madrid Centro", "Madrid-otros", "Fuera de Madrid")
row.names(ABStable) <- lbls
ABStable
```

```
##
##      Madrid Sur   Madrid Centro    Madrid-otros Fuera de Madrid
##              46              36              12               1
```

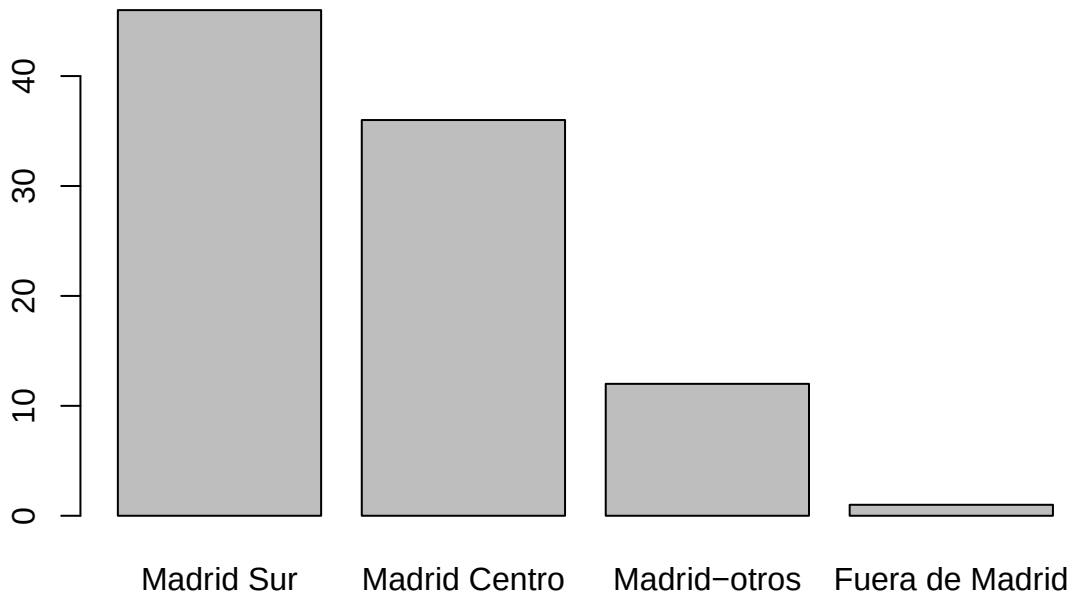
```
RELtable <- prop.table(ABStable)
RELtable
```

```
##
##      Madrid Sur   Madrid Centro    Madrid-otros Fuera de Madrid
##      0.48421053      0.37894737      0.12631579      0.01052632
```

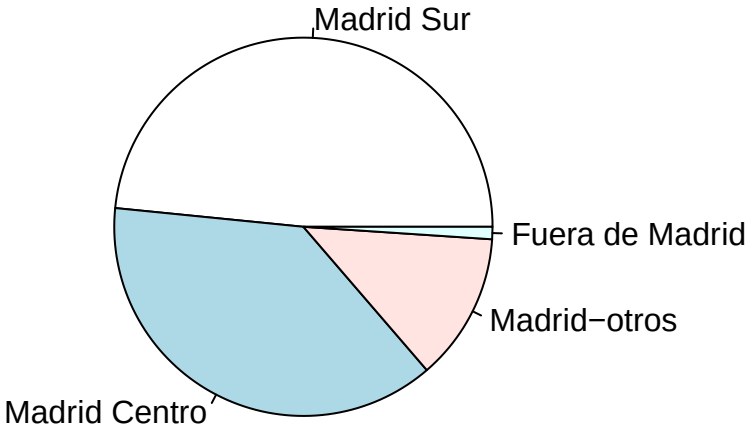
where we can see that the biggest group of students are the ones that come from Madrid Sur, made of 46 students which account for the 48.4% of the sample.

In order to get a bar chart or a pie chart, we use the following instructions:

```
barplot(ABStable)
```




```
pie(ABStable)
```




# 3. Description of quantitative variables

## 3.1 Graphical analysis of discrete variable with only few values

In the case of quantitative discrete variables assuming only few values, the graphic analysis is the same as the one we saw for categorical variables. We can now plot a bar chart. The frequency table would be similar to the one generated by using the analysis of categorical data.

For example, the variable `hermanos` gives the number of siblings that each student has.

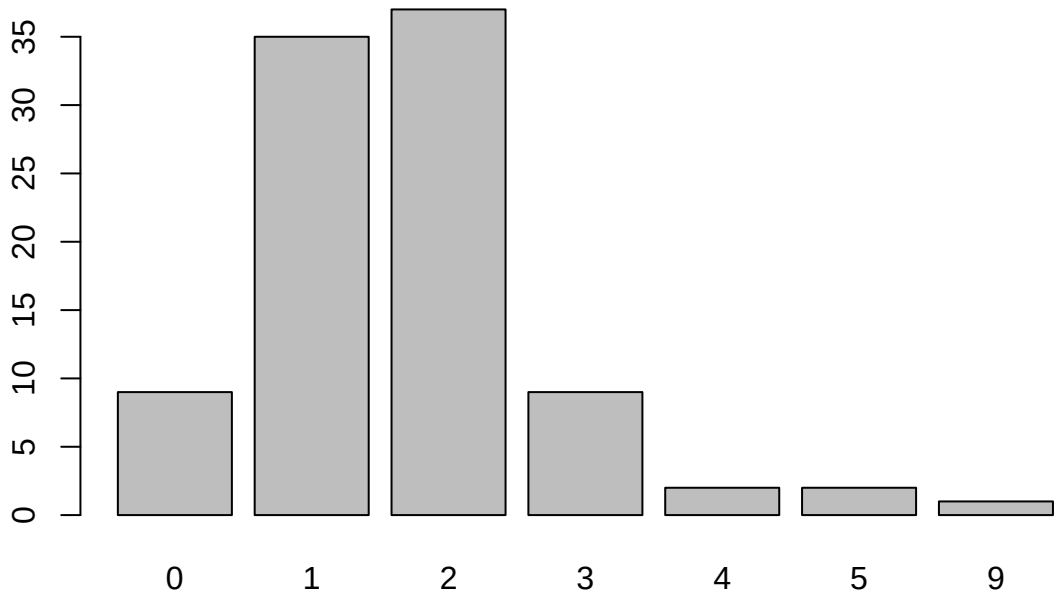
```
ABStable <- table(AlumnosIndustriales$hermanos)
ABStable
```

```
##
##  0  1  2  3  4  5  9
##  9 35 37  9  2  2  1
```

```
RELtable <- prop.table(ABStable)
RELtable
```

```
##
##          0          1          2          3          4          5          9
## 0.09473684 0.36842105 0.38947368 0.09473684 0.02105263 0.02105263 0.01052632
```

```
barplot(ABStable)
```



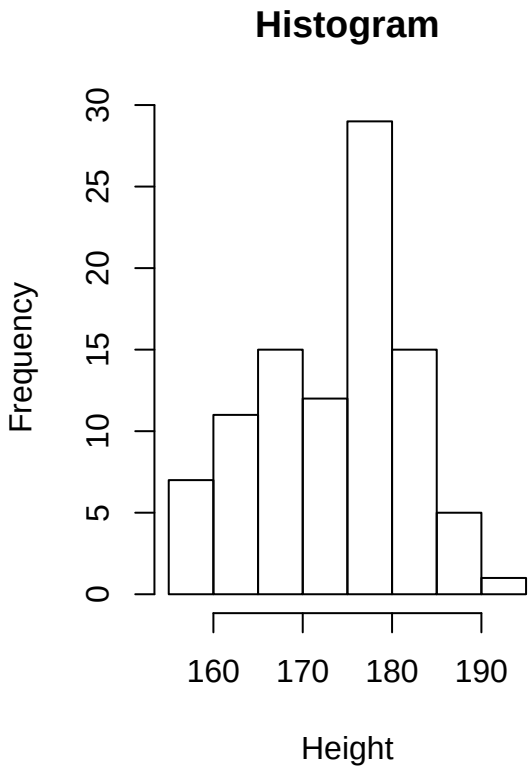

From the picture we can see that the families more frequent (among those who have children studying Industrial Engineering at UC3M) are the ones with 2 or 3 children (1 or 2 siblings).

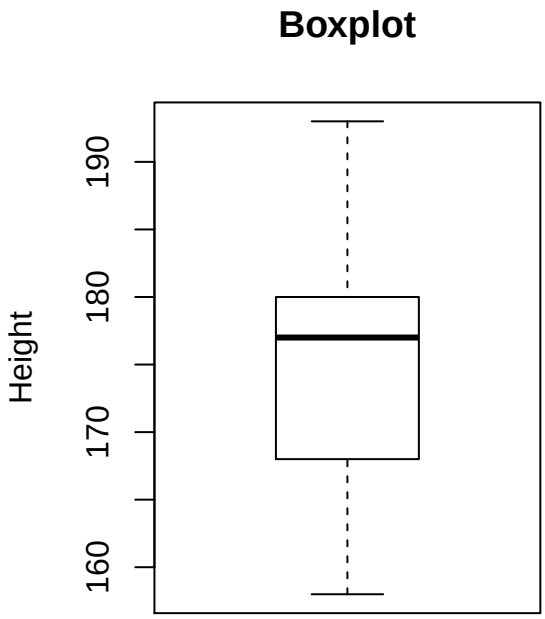
### 3.2 Graphical analysis of quantitative variables

The graphical analysis of generaL quantitative variables can be done using the functions `hist` and `boxplot`.

The variable `altura` contains the heights of the students. Here, we plot its histogram and its boxplot.

```
par(mfrow=c(1,2))
hist(AlumnosIndustriales$altura, xlab = "Height", main = "Histogram")
boxplot(AlumnosIndustriales$altura, ylab = "Height", main = "Boxplot")
```

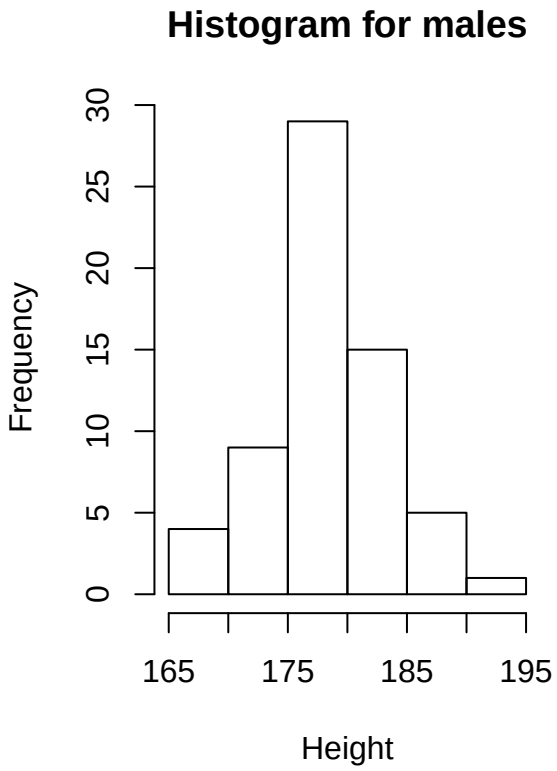


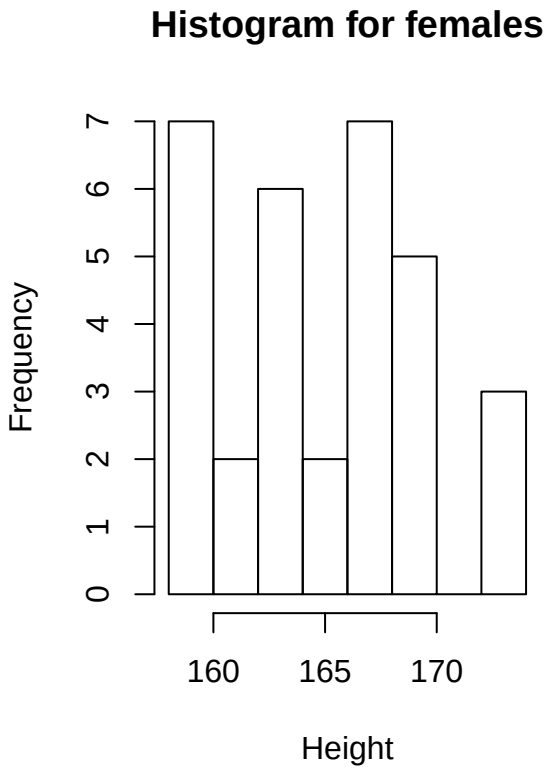




The boxplot show that the height distribution is asymmetric. The central box shows a skewness to the left even if the tail are not very long. This effect is visible in the histogram as well. Also, we can see two modes suggest that the sample is not homogeneous. Probably this is due to the joint presence of males' and females' heights.

The variable `sexo` contains the gender of the students (1 = male, 0 = female). We can use this variable to select the heights of mans or of woman and in this way to check if these two groups have the corresponding height data concentrated around the two modes.

```
par(mfrow=c(1,2))
hist(AlumnosIndustriales$altura[AlumnosIndustriales$sexo == 1], xlab = "Height",
     main = "Histogram for males")
hist(AlumnosIndustriales$altura[AlumnosIndustriales$sexo == 0], xlab = "Height",
     main = "Histogram for females")
```




We see that displaying only the males' heights, the distribution looks more symmetric, unimodal, and with mode in the interval [175, 180], with a high concentration around it. Repeating the same for females we get that its distribution does not have a unimodal bell shape like the one of the males. Maybe this can be due to the fact that we have less data (only 32 individuals) or maybe because they are heterogeneous by themselves.

### Frequency table

The frequency table gives us the same information contained in the corresponding histogram, but it let us to see the numerical values of the frequencies in each interval. To obtain the frequency table we can use the following code:

```
Altura <- AlumnosIndustriales$altura
range(Altura)
```

```
## [1] 158 193
```

```
breaks = seq(155, 195, by=5)
breaks
```

```
## [1] 155 160 165 170 175 180 185 190 195
```

```
Altura.cut = cut(Altura, breaks, right=TRUE)
Altura.table = table(Altura.cut)
Altura.table
```

```
## Altura.cut
## (155,160] (160,165] (165,170] (170,175] (175,180] (180,185] (185,190] (190,195]
##         7        11        15        12        29        15         5         1
```

```
prop.table(Altura.table)
```

```
## Altura.cut
##  (155,160]  (160,165]  (165,170]  (170,175]  (175,180]  (180,185]  (185,190]
## 0.07368421 0.11578947 0.15789474 0.12631579 0.30526316 0.15789474 0.05263158
##  (190,195]
## 0.01052632
```

These tables show that the modal interval are around the values (midpoint), 167.5 and 177.5, and that the interval with highest frequency, with midpoint 177.5, contains more than 30% of the students.

## 3.4 Numerical summary measures of quantitative variables

To compute the summary measures of the variable `altura` we can use the function `summary`:

```
summary(Altura)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   158.0   168.0   177.0   174.6   180.0   193.0
```

It gives position summary measures such mean, median, first and third quartiles, the minimum and the maximum.

Another summary statistics can be obtained using function `descr` from package `summarytools`:

```
suppressWarnings(library(summarytools))
descr(Altura)
```

```
## Descriptive Statistics
## Altura
## N: 95
##
##                      Altura
## ----------------- --------
##              Mean   174.62
##           Std.Dev     8.23
##               Min   158.00
##                Q1   168.00
##            Median   177.00
##                Q3   180.00
##               Max   193.00
##               MAD     7.41
##               IQR    12.00
##                CV     0.05
##          Skewness    -0.29
##       SE.Skewness     0.25
##          Kurtosis    -0.92
##           N.Valid    95.00
##         Pct.Valid   100.00
```

It includes the previous position statistics and some dispersion and shape measures such as the standard deviation, the coefficient of variation, the MAD, the skewness and the kurtosis coefficients.

It is now necessary to clarify some things about these summary measures:

- The variance (as a consecuence, the standard deviation) that is used is indeed the sample variance that is computed using the following formula $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ instead of the variance formula

$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$. The convenience of dividing by $n-1$ instead of $n$ is not immediate, and their theoretical justification will be seen on more advanced topics.

- The kurtosis coefficient computed is what is known as 'excess kurtosis', defined as $\kappa = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{ns^4} - 3$. Therefore for a bell shaped variable the kurtosis is equal to 0.

In the following it is shown a comparison between males and females by mean of some characteristic measures:

```
suppressWarnings(library(summarytools))
descr(Altura[AlumnosIndustriales$sexo==1])
```

```
## Descriptive Statistics
## Altura
## N: 63
##
##                      Altura
## ----------------- --------
##              Mean   179.35
##           Std.Dev     5.04
##               Min   165.00
##                Q1   177.00
##            Median   180.00
##                Q3   182.00
##               Max   193.00
##               MAD     2.97
##               IQR     4.50
##                CV     0.03
##          Skewness    -0.28
##       SE.Skewness     0.30
##          Kurtosis     0.66
##           N.Valid    63.00
##         Pct.Valid   100.00
```

```
descr(Altura[AlumnosIndustriales$sexo==0])
```

```
## Descriptive Statistics
## Altura
## N: 32
##
##                      Altura
## ----------------- --------
##              Mean   165.31
##           Std.Dev     4.43
##               Min   158.00
##                Q1   161.50
##            Median   165.00
##                Q3   168.50
##               Max   174.00
##               MAD     5.19
##               IQR     6.50
##                CV     0.03
##          Skewness     0.21
##       SE.Skewness     0.41
##          Kurtosis    -1.14
##           N.Valid    32.00
##         Pct.Valid   100.00
```

We can see now that the men of this sample are in average taller than the women. The average height of men is 179 cm while for the women it is 165 cm. In both genders the mean is almost equal to the median; this is clearly visible looking at the symmetry of the boxplot (see section 4) and the low absolute value of the skewness. This concentration around the median value is visible also when looking at the interquartile range. 50% of the men (women) located in the central positions have a height that differs from the median value by less than 3 cm.

### 3.5 Percentiles

To obtain the percentiles, we use the function `quantile`

```
quantile(Altura, probs = seq(0, 1, 0.1))
```

```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
## 158.0 163.0 167.0 169.2 173.0 177.0 179.0 180.0 181.0 184.6 193.0
```

So, we can conclude that 20% of the students is less than 167 cm tall and 80% of them is less than 181 cm. The argument `probs` can have any value in the interval $[0, 1]$.

## 4 Simultaneous description of more than one variable

In many cases we are interested in comparing several variables, or compare the values of a variable divided into two or more groups of individuals as in the case of height by gender. In such cases it is more interesting to produce graphs and statistical summaries together in the same window to make this comparison, instead of performing the univariate analysis of each variable separately. For example, we would want to generate the boxplots of each variable on the same graph.

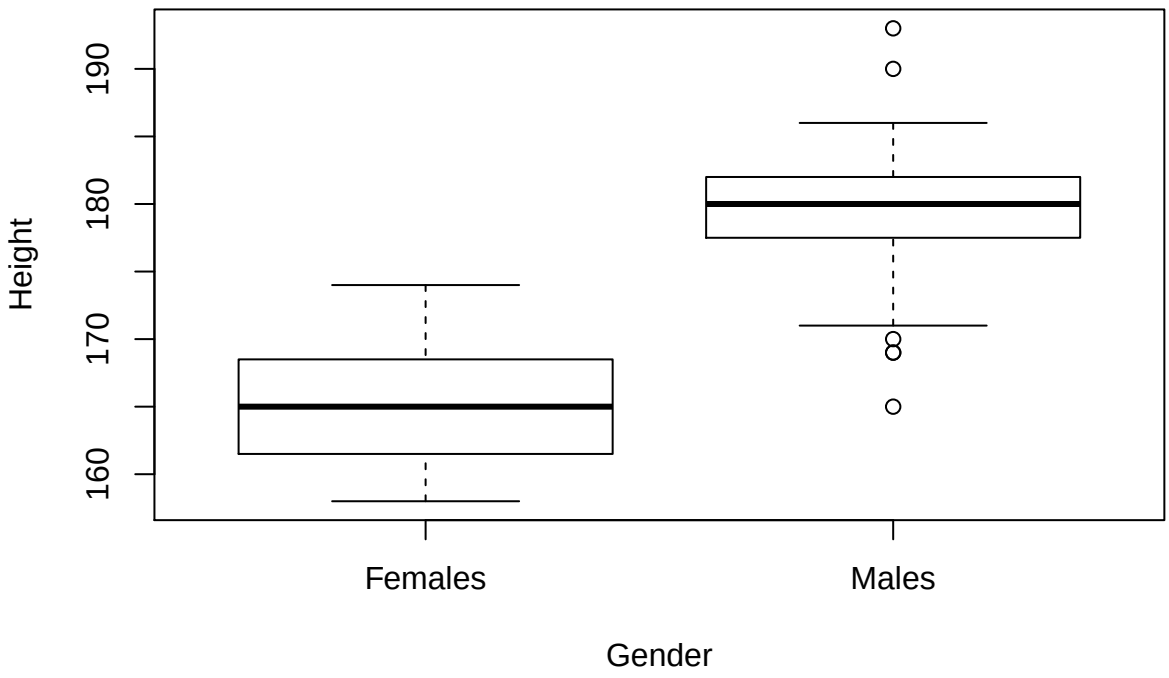
### 4.1 Multiple boxplots

Our goal is to create a graph that has the boxplot diagrams of several variables in it or the same variable divided in more than one subgroup. This graph will allow a better comparison of these variables.

#### 4.1.1 One quantitative variable by subgroups

We are interested in analyzing how the distribution of the values of one variable is when the dataset is subdivided into subgroups according to some criterion. For examplem we want to study the heights of a group of students according to their gender. The variable `sexo` just takes values 1 and 0. These values are needed only to distinguish the members of each group, so the number they assume is irrelevant. It could be -1 and 1, or even characters.

```
boxplot(AlumnosIndustriales$altura ~ AlumnosIndustriales$sexo, xlab = "Gender",
        ylab = "Height", names = c("Females", "Males"))
```




It may be seen that men are generally taller than women. Simultaneously displaying the two box-whisker plots we can interpret that approximately only 25% of the women have higher heights comparable to 25% of the shorter mens. Also, some outliers at the men's boxplot are visible.

### 4.1.2 Several quantitative variables

If we have several variables of the same magnitude or different magnitudes, we can also do multiple boxplot of these variables in one graph. As an example we will use the data file `Rotura.xlsx`. This file contains the breakdown tension or ultimate tensile strength (that is the strengh at the moment that the rupture occurs) of a set of identical parts in order to test the resistance of the material. The file contains ten variables, LAB1 to LAB10, where each sample of breakage tensions is obtained in ten different laboratories. In each laboratory 100 different pieces are broken and the corresponding pressure values are recorded. We are going to show the boxplots of these ten variables:

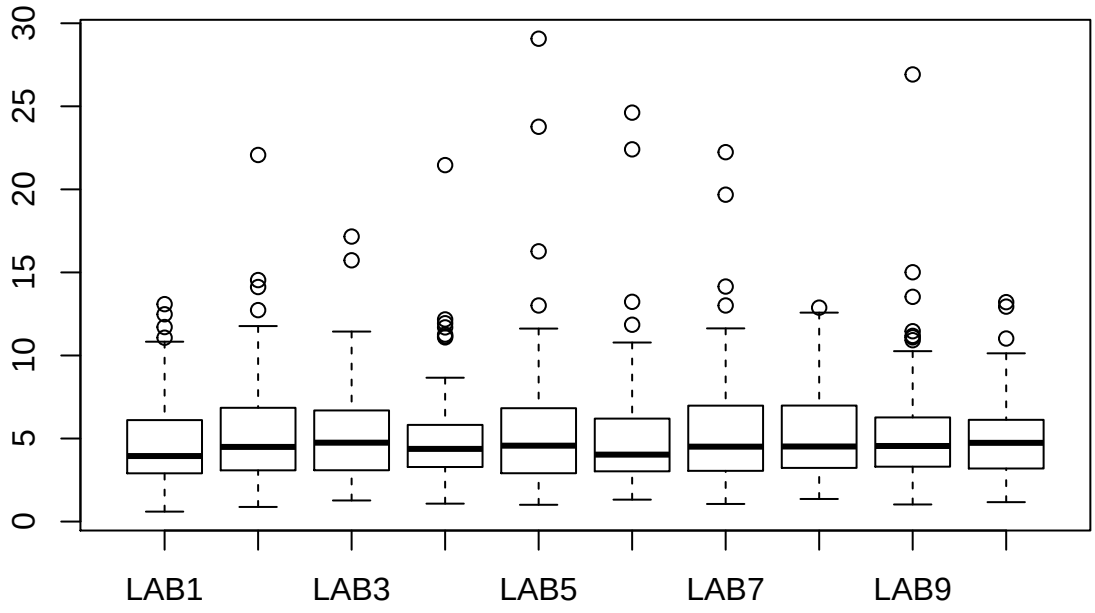
```
library(readxl)
Roturas <- read_excel("Roturas.xlsx")
#View(Roturas)
```

Roturas

Filter

| | LAB1 | LAB2 | LAB3 | LAB4 | LAB5 | LAB6 | LAB7 | LAB8 | LAB9 | LAB10 | MEDIAS | VARIANZAS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5.87 | 3.79 | 6.81 | 4.25 | 6.89 | 2.97 | 9.94 | 9.75 | 2.13 | 3.07 | 4.766 | 7.294 |
| 2 | 3.13 | 3.21 | 3.82 | 6.03 | 11.44 | 3.25 | 2.79 | 2.85 | 5.66 | 1.23 | 5.452 | 11.014 |
| 3 | 4.02 | 3.11 | 7.29 | 4.79 | 8.20 | 3.40 | 2.85 | 4.14 | 2.92 | 3.30 | 5.243 | 8.645 |
| 4 | 6.41 | 5.23 | 7.53 | 5.67 | 6.91 | 9.02 | 2.24 | 4.01 | 4.03 | 3.33 | 5.033 | 8.285 |
| 5 | 3.24 | 9.38 | 2.81 | 4.89 | 2.87 | 5.53 | 2.38 | 12.50 | 6.05 | 3.06 | 5.589 | 17.900 |

```r
boxplot(Roturas[,1:10])
```



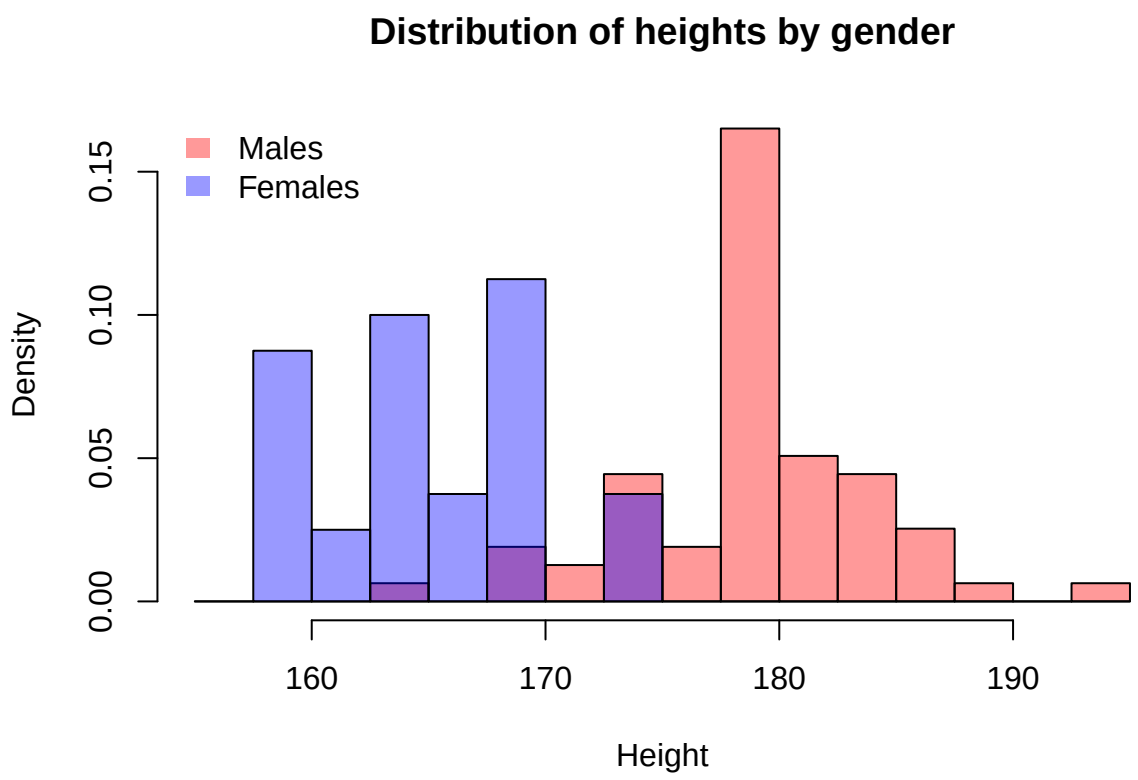

It is seen that the results of the 10 laboratories are similar. From the whiskers of the third quartile and from the outliers it can be seen that in general, the distributions are positive skewed.

## 4.2 Two overlapped histograms

If we want to compare two histograms, it is useful to put them in the same graph. There are many ways to produce overlappeed histograms but here we show how to do it using the base R graphics. We will use the example of heights by gender:

```r
# notice that plot = FALSE
histMales <- hist(Altura[AlumnosIndustriales$sexo == 1], breaks = seq(155,195,2.5),
                  plot = FALSE)
histFemales <- hist(Altura[AlumnosIndustriales$sexo == 0], breaks = seq(155,195,2.5),
                    plot = FALSE)
# calculate the range of the graph
xlim <- range(histMales$breaks,histFemales$breaks)
ylim <- range(0,histMales$density,histFemales$density)
# plot the first histogram
plot(histMales,xlim = xlim, ylim = ylim, col = rgb(1,0,0,0.4), xlab = 'Height',
     freq = FALSE, ## relative, not absolute frequency
     main = 'Distribution of heights by gender')
## plot the second histogram on top of this
opar <- par(new = FALSE)
plot(histFemales,xlim = xlim, ylim = ylim,
     xaxt = 'n', yaxt = 'n', ## don't add axes
```

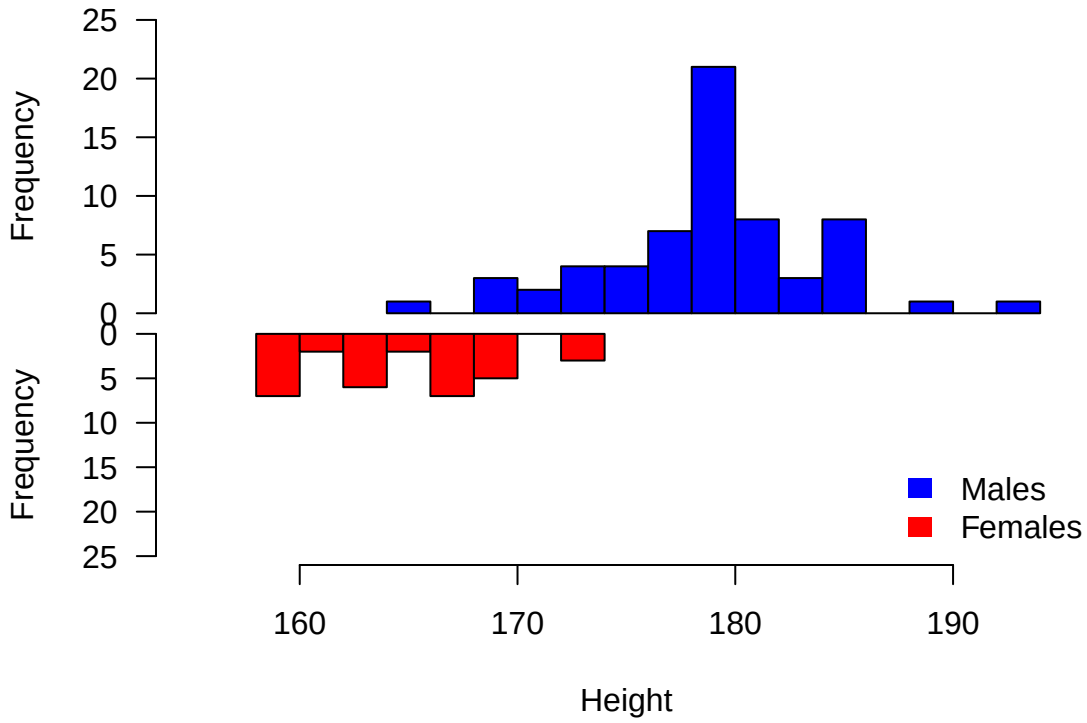
```
     col = rgb(0,0,1,0.4), add = TRUE,
     freq = FALSE) ## relative, not absolute frequency
## add a legend in the corner
legend('topleft',c('Males','Females'), fill = rgb(1:0,0,0:1,0.4), bty = 'n', border = NA)
```




```
par(opar)
```

Another posibility is to obtain mirrored histogram:

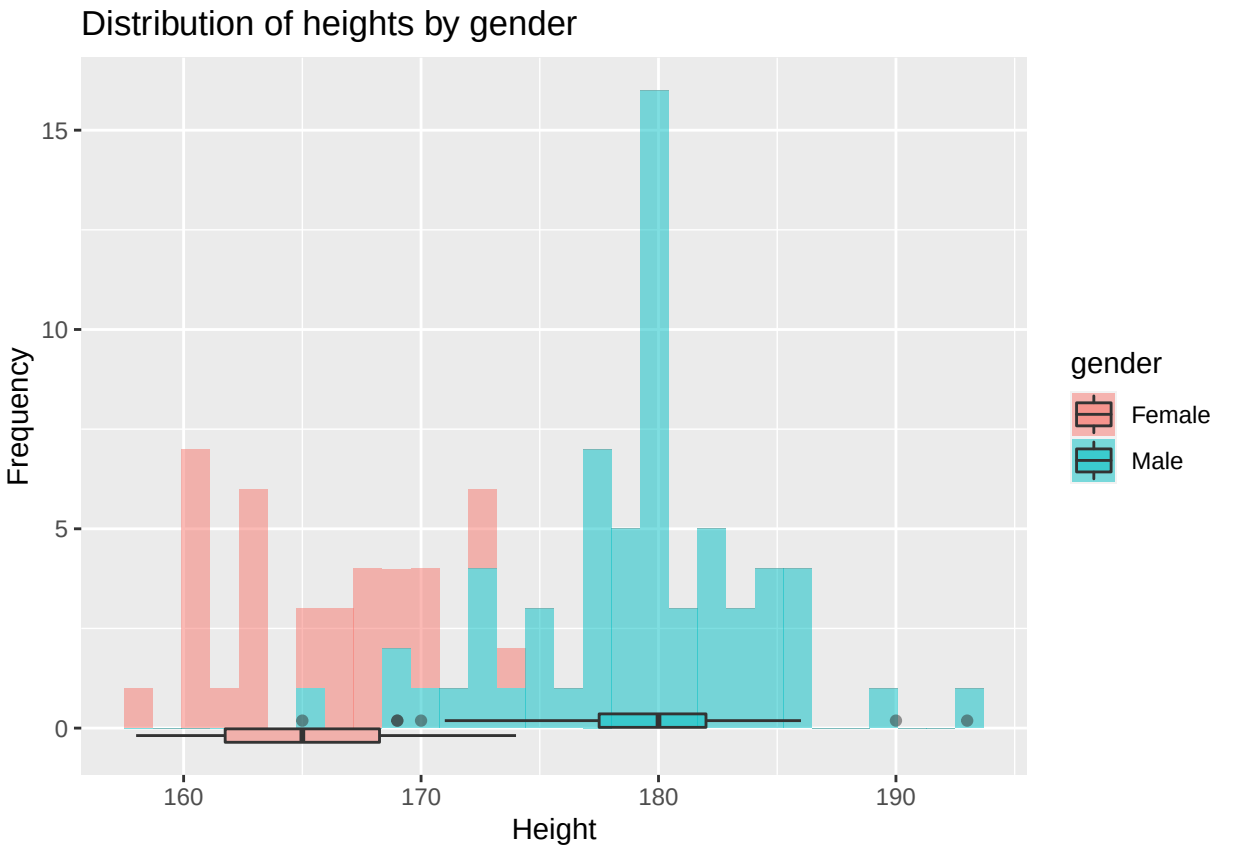
```
par(mfrow=c(2,1))

#Make the plot
par(mar=c(0,5,3,3))
hist(Altura[AlumnosIndustriales$sexo == 1], main="" , xlim=c(155,195),
     ylab="Frequency", xlab="", ylim=c(0,25) , xaxt="n", las=1 ,
     col="blue", breaks=10)
par(mar=c(5,5,0,3))
hist(Altura[AlumnosIndustriales$sexo == 0], main="" , xlim=c(155,195),
     ylab="Frequency", xlab="Height", ylim=c(25,0), las=1,
     col="red"  , breaks=10)
legend('bottomright',c('Males','Females'), fill = c("blue", "red"), bty = 'n', border = NA)
```




The ggplot2 package, created by Hadley Wickham, offers a powerful graphics language for creating elegant and complex plots and it is very popular at the R community. Mastering the ggplot2 language can be challenging. There is a helper function called qplot that can hide much of this complexity when creating standard graphs such as histograms. Moreover, we can combine histogram and boxplot in the same graph:

```
suppressWarnings(library(ggplot2))
AlumnosIndustriales$gender <- "Male"
AlumnosIndustriales$gender[AlumnosIndustriales$sexo == 0] <- "Female"
qplot(altura, data=AlumnosIndustriales, geom=c("histogram", "boxplot"), fill=gender,
      alpha=I(.5), main="Distribution of heights by gender",
      xlab="Height", ylab="Frequency")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




### 4.3 Numerical summary measures of several variables-

In general, we start our data analysis by looking at the graphs. In a second moment we look for measures that can summarize in a quantitative way the features of most interest.

We will use the data.frame `AlumnosIndustriales`. It should be notice that variables `sexo`, `locomocion` and `residencia` are coded as numeric and variable `gender` is character, so we can change their class in order to obtain valid summary statistics.

```
AlumnosIndustriales$sexo <- as.factor(AlumnosIndustriales$sexo)
AlumnosIndustriales$residencia <- as.factor(AlumnosIndustriales$residencia)
AlumnosIndustriales$locomocion <- as.factor(AlumnosIndustriales$locomocion)
AlumnosIndustriales$gender <- as.factor(AlumnosIndustriales$gender)
summary(AlumnosIndustriales)
```

```
##    nacimiento         altura           peso           zapato      sexo
##  Min.   : 1.000   Min.   :158.0   Min.   :45.00   Min.   :36.00   0:32
##  1st Qu.: 3.000   1st Qu.:168.0   1st Qu.:56.00   1st Qu.:38.00   1:63
##  Median : 5.000   Median :177.0   Median :69.00   Median :42.00
##  Mean   : 5.463   Mean   :174.6   Mean   :67.77   Mean   :40.98
##  3rd Qu.: 7.500   3rd Qu.:180.0   3rd Qu.:75.00   3rd Qu.:43.00
##  Max.   :12.000   Max.   :193.0   Max.   :99.00   Max.   :46.00
##      dinero           tiempo       locomocion residencia    hermanos
##  Min.   :   0.0   Min.   :  1.00   1:19       1:46       Min.   :0.000
##  1st Qu.: 217.5   1st Qu.: 20.00   2: 2       2:36       1st Qu.:1.000
##  Median : 655.0   Median : 40.00   3:29       3:12       Median :2.000
##  Mean   :1039.2   Mean   : 41.42   4:37       4: 1       Mean   :1.716
```

```
##  3rd Qu.:1300.0   3rd Qu.: 60.00   5: 8                  3rd Qu.:2.000
##  Max.   :5000.0   Max.   :120.00                         Max.   :9.000
##   Variables            gender
##  Length:95          Female:32
##  Class :character   Male  :63
##  Mode  :character
##
##
##
```

If you are only interested on quantitative variables, you can use function `descr` from package `summaryTools`:

```
descr(AlumnosIndustriales)
```

```
## Non-numerical variable(s) ignored: sexo, locomocion, residencia, Variables, gender
```

```
## Descriptive Statistics
## AlumnosIndustriales
## N: 95
##
##                     altura    dinero   hermanos   nacimiento     peso   tiempo   zapato
## ----------------- -------- --------- ---------- ------------ -------- -------- --------
##              Mean   174.62   1039.23       1.72         5.46    67.77    41.42    40.98
##           Std.Dev     8.23   1200.14       1.25         3.26    11.80    24.74     2.73
##               Min   158.00      0.00       0.00         1.00    45.00     1.00    36.00
##                Q1   168.00    200.00       1.00         3.00    56.00    20.00    38.00
##            Median   177.00    655.00       2.00         5.00    69.00    40.00    42.00
##                Q3   180.00   1300.00       2.00         8.00    75.00    60.00    43.00
##               Max   193.00   5000.00       9.00        12.00    99.00   120.00    46.00
##               MAD     7.41    770.95       1.48         2.97    13.34    29.65     2.97
##               IQR    12.00   1082.50       1.00         4.50    19.00    40.00     5.00
##                CV     0.05      1.15       0.73         0.60     0.17     0.60     0.07
##          Skewness    -0.29      1.94       2.38         0.38     0.25     0.63    -0.33
##       SE.Skewness     0.25      0.25       0.25         0.25     0.25     0.25     0.25
##          Kurtosis    -0.92      3.42      10.77        -0.72    -0.59    -0.04    -1.11
##           N.Valid    95.00     95.00      95.00        95.00    95.00    95.00    95.00
##         Pct.Valid   100.00    100.00     100.00       100.00   100.00   100.00   100.00
```