BIOSTATISTICS

Descriptive statistics: exploratory data analysis and data visualization

Angela Jimeno Martin (ajimeno@usj.es)

Contents

Introduction to statistics in R

Data import

In biostatistics it is very common that the data to be worked with has to be imported from external files. The fundamental objective will be to incorporate the data in an object within R.

Plain text files

Normally these files have a .txt extension, although they can have any other extension. The most common is that the files to be loaded have the information separated by tabs. Sometimes they can also be found as tab separated value or .tsv, which corresponds to the character. The command to load these files is `read.table()`:

```
record <- read.table(file = "DataReg.txt",header = TRUE,  sep = "\t")
```

The type of object created is a `data.frame`.

.csv files

These files are groups of data whose values are separated by commas. If they are generated by excel type programs in their Spanish local language settings, instead of being separated by commas, they will be separated by semicolons. Therefore, different commands must be used. Usually, we will work with files separated by ;.

```
 lab.data <- read.csv2(file="EsterData.csv", header = TRUE,  sep = ";")
#The generated object is also a data.frame
```

Basic commands explore data

Once the data have been collected, either from experiments or from repositories, the first step is to carry out an initial exploration of the information. The objective is to gain knowledge about the data, to know the values that the variables acquire and to see how they change depending on the samples, i.e. to identify their distribution.

For a variable, the distribution shows the possible values it can take, the likelihood of observing those values, and the frequency with which we expect to find them for a random sample of a population. The initial screening allows for a sifting of the data to focus the analysis on those aspects that are most striking or relevant. The following is a summary of the basic R functions for obtaining information from datasets.

Function	Task description
sum(x)	Adds the elements in the vector x
prod(x)	Multiplies the elements in the vector x
max(x)	Maximum element in the vector x
min(x)	Minimum element in vector x
range(x)	Range (minimum and maximum) of the elements of x
length(x)	Number of elements in vector x
mean(x)	Mean of elements in x
median(x)	Meidan of elements in x
var(x)	Variance of elements in x
sd(x)	Standard deviation of the elements in x
cor(x,y)	Correlation between the elements of the vectors x and y
quantile(x,p)	The quantile pth of x
cov(x,y)	Covariance between x and y.

Function summary

Data collected in an assay of three different batches of the same enzyme are shown below. Each assay digests a protein of 200 amino acids by the enzyme. The result of the time taken to digest it is as follows:

```
enzyme
```

```
##   VarietyA timeA VarietyB timeB VarietyC timeC
## 1        a  0.12        b  0.12        c  0.13
## 2        a  0.13        b  0.14        c  0.12
## 3        a  0.13        b  0.13        c  0.11
## 4        a  0.12        b  0.15        c  0.13
## 5        a  0.13        b  0.13        c  0.12
## 6        a  0.12        b  0.14        c  0.13
```

To quickly obtain information on which of the enzyme batches is best, the summary function is very useful. It performs the function of several of the commands in the table above.

```
summary(enzyme)
```

```
##    VarietyA             timeA          VarietyB             timeB
##  Length:6           Min.   :0.120   Length:6           Min.   :0.120
##  Class :character   1st Qu.:0.120   Class :character   1st Qu.:0.130
##  Mode  :character   Median :0.125   Mode  :character   Median :0.135
##                     Mean   :0.125                      Mean   :0.135
##                      3rd Qu.:0.130                      3rd Qu.:0.140
##                      Max.   :0.130                      Max.   :0.150
##    VarietyC              timeC
##  Length:6            Min.   :0.1100
##  Class :character    1st Qu.:0.1200
##  Mode  :character    Median :0.1250
##                      Mean   :0.1233
##                      3rd Qu.:0.1300
##                      Max.   :0.1300
```

It can be observed that the mean of variety B is greater in time than those of varieties A and C. This data could be statistically significant, and therefore be used to assess the differences between batches.

Types of variables

The variables that we can represent can be of two types: • Qualitative. • Quantitative. The following example contains both types of variables. It is an example within the MASS package. called birthwt which includes information on 189 newborns born in Springfield in 1986:

```
require(MASS)
```

```
## Loading required package: MASS
```

```
head(birthwt)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

The variables for which data are available are:

```
names(birthwt)
```

```
##  [1] "low"   "age"   "lwt"   "race"  "smoke" "ptl"   "ht"    "ui"    "ftv"
## [10] "bwt"
```

- `low`: index of births with babies under 2.5 kg (0 = normal birth weight, 1 = low birth weight).
- `age`: mother's age in years.
- `lwt`: mother's weight in pounds at her last menstrual period.
- `race`: mother's race (1 = white, 2 = African American, 3 = other).
- `smoke`: mother smoked during pregnancy (0 = non-smoker, 1 = smoker).
- `ptl`: number of previous preterm deliveries.
- `ht`: history of hypertension (0 = no, 1 = yes).
- `ui`: presence of uterine irritability (0 = no, 1 = yes).
- `ftv`: number of physical visits during the first quarter.
- `bwt`: weight in grams at birth.

The variables age, lwt, ptl, ftv and bwt are quantitative variables. The rest, low, race, smoke, ht and ui are qualitative or categorical, although they are coded with numerical values. In this case, R does not recognise them as categorical variables, but as numerical by default, so they must be recoded to factors with as.factor().

Exploring qualitative variables

Absolute frequency is the number of times a specific category, c, is observed nc . For the above dataset, the number of white women was 96, African-American women was 26, while other race was 67.

The sum of the frequencies of all categories has to be equal to the total number of observations in the sample.

$$\sum_c n_c = n_1 + n_2 + n_3 = 96 + 26 + 67 = 189$$

Relative frequency We have obtained that n1 = 96, n2 = 26 and n3 = 67, for white, African-American and other women, respectively. If we would like to express that the sample size is representative for American society so that the data should also be representative, then we can calculate the relative frequency. It would therefore be defined as the proportion of each possible category in relation to the total population. And it is obtained:

$$p_c = \frac{n_c}{n},$$

where p_c is the relative frequency, n_c is the absolute frequency found above and n is the number of observations. Sometimes the value is usually presented in percentages, which can be done by simply multiplying the result by 100. Therefore, the above values of relative frequencies would be:

$$p_1 = 96/189 = 0.508$$

$$p_2 = 26/189 = 0.138$$

$$p_3 = 67/189 = 0.354$$

The sum of the relative frequencies of all categories must always equal 1 (100 if the percentages are added together).

$$\sum_c n_c = 1,$$

$$\sum_c n_c = n_1 + n_2 + n_3 = 0.508 + 0.138 + 0.354 = 1$$

NOTE: The mode, which is a measure of central tendency and represents the most repeated value, will coincide with the value with the highest frequency.

Exploring numerical variables

In the visualisation of quantitative variables, the exploratory study of their distribution, i.e. how concentrated and how dispersed their values are, is of special interest. Concentration refers to the central tendency of the values, i.e. around which point most of the values are clustered. Measures of central tendency are representative of a data set. The dispersion of a distribution refers to how dispersed the possible values are around a location.

Graphics

Graphical representations are often of great help in biostatistics during data exploration, as a lot of information can be obtained at a glance. For each type of variable, there are a number of graphs that help to represent the corresponding data.

Categorical variables

Bar charts

It is the simplest way to visualise data. You can see: • The possible values per category. • They are usually used for qualitative variables. • They present the number of times each category is observed in the sample. In short, they are the representation of the absolute frequencies obtained for a group of data.

For this theoretical part, we will also work with the dataset diabetes, which shows data collected in a hospital in India on diabetic women, containing the following information: • `Pregnancies`, number of pregnancies. • `Glucose`, plasma glucose concentration in the oral glucose tolerance test, • `BloodPressure`, diastolic blood pressure, • `SkinThickness`, thickness of the skin fold of the triceps, • `BMI`, body mass index, • `DiabetesPedigreeFunction`, diabetic pedigree function, • `Age`, • `Outcome`, disease status, 1 diabetic (Yes), 0 non-diabetic (No).

Representing the Outcome variable with a bar chart shows the frequency of the values obtained in the study for female patients.

```r
# First we calculate absolute frequency
tabla.diabetes <- table(diabetes$Outcome)

# Then we represente that frequency
barplot(tabla.diabetes, main = "Bar chart of diabetes frequency"
        ,xlab = "Outcome", ylab="Frequency", col = c("#a1e9f0", "#d9b1f0"))
```

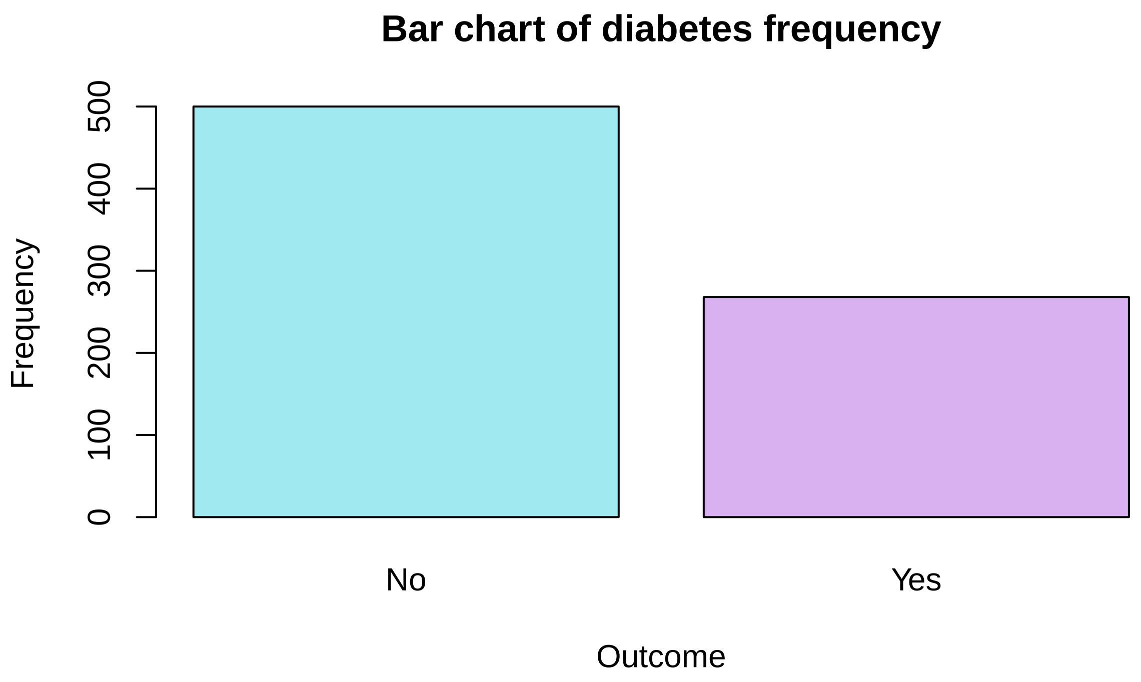

Figure 1: Barplot

Pie chart

Pie charts are used for the visualisation of the relative frequencies of a categorical variable. The area of a circle will be divided into as many sectors as possible categories. The area of each sector shall be proportional to the calculated relative frequency.

```
# Using absolute frequencies, it is then straightforward to represent
# its relative counterparts
pie(tabla.diabetes, main="Diabetes Frequency")
```

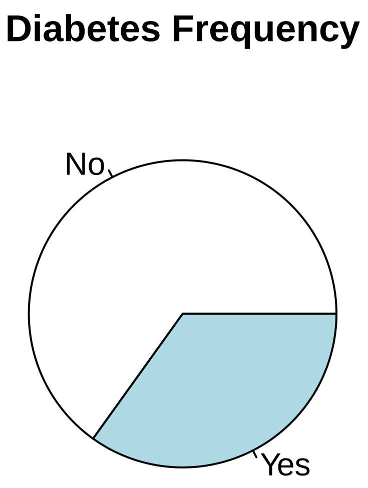

Figure 2: Pie chart

Quantitative variables

Histograms They are very common and similar to a bar chart but here the variable to be studied is quantitative and continuous and is grouped into a finite number of intervals. For each interval, the height of the bar corresponds to the frequency of that observation in that interval. The function to use is `hist()`. For the diabetes dataset variable body mass index (`BMI`), a histogram would look like this.

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

```
## NULL
```

```
# Absolute frequency
hist(diabetes$BMI, main="BMI absolute frequency histogram", xlab="BMI"
     , col="#b9e38d", freq = TRUE)
```

```
# Relative frequency
hist(diabetes$BMI, main="BMI relative frequency histogram", xlab="BMI"
     , col="#eb8060", freq = FALSE)
```

Both histograms look like very similar, only by changing the values of the y-axis. However, it is more common to represent the second one, with the value of `Density`. This would be similar to the concept of relative frequency, but as it is a quantitative variable, it is more correct to refer to the data obtained as density. The density is the relative frequency of a unit interval. It is obtained by dividing the relative frequency by the width of the interval, c:

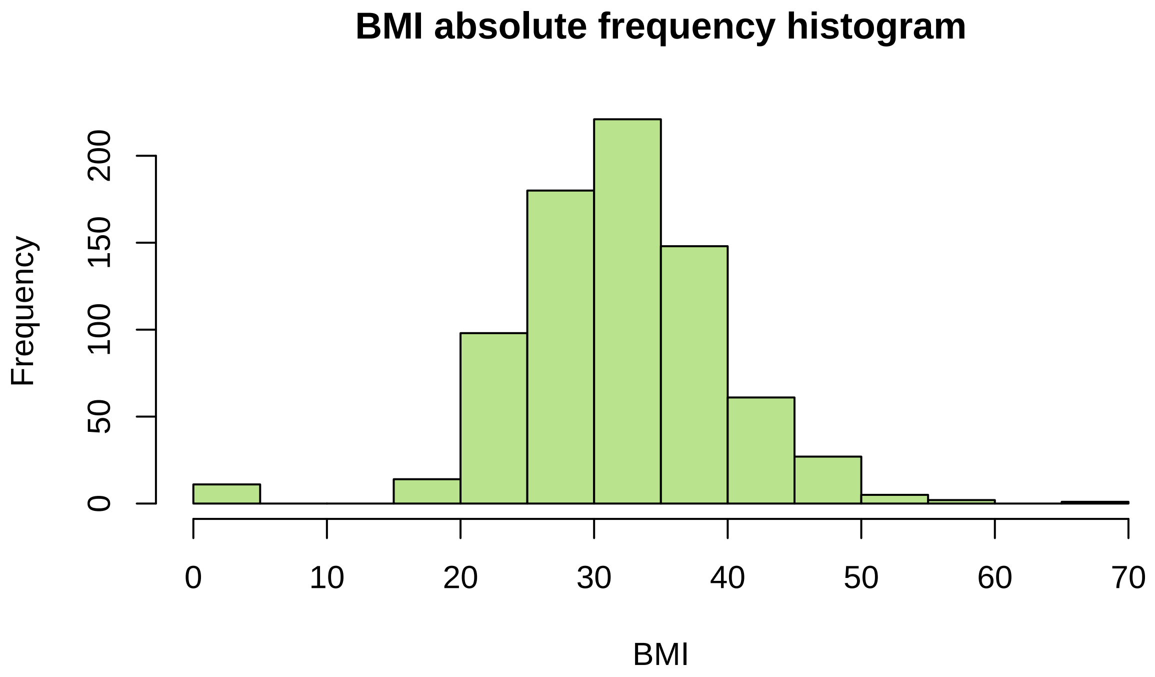

Figure 3: Histogramas

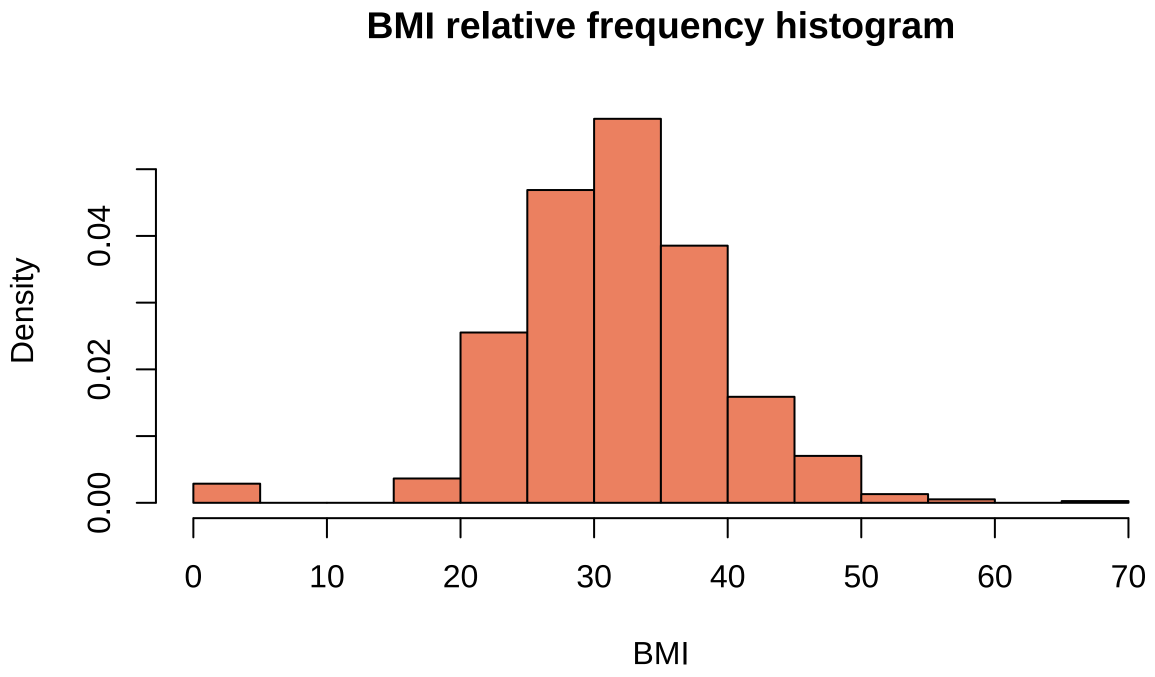

Figure 4: Histogramas

$$f_c = \frac{p_c}{w_c}, \text{ where}$$

$p_c = n_c/n$ is the relative frequency with n_c as the frequency of the interval c and n the total sample size, w_c is the size of the interval of c. For the interval of (30,35], which is the seventh interval, $c = 7$, there are 221 observations (n_7). Therefore the relative frequency $p_7 = 0.2877604$. Considering that $w_c = 5$, the density will be:

$$f_7 = 0.289/5 = 0.0578$$

* Shape of histograms The way the data is reflected in the histograms is also very informative. It gives an idea of how the values are spread around the location. The different shapes we can find are: ** Symmetric: the observed densities are the same on each side of the centre of the graph. It is very rare to find a dataset that generates an exact symmetric histogram, so it is usually called symmetric if the densities on each side are approximate. This would be the case for BMI data.

** Asymmetric: – Left-skewed: the histogram shows most of the data on the left side of the graph, there is little data density on the right side of the graph. Also called negative skewness. – Right-skewed: the histogram shows most of the data on the right side of the graph, there is little data density on the left side of the graph. Also called positive skewness. Some of the functions, such as the Normal, are symmetric.

Boxplot

In order to study box plots, it is necessary to review some basic statistical concepts:

Arithmetic mean

This is the mean value of the observed data. For values x_1, …, xn, the arithmetic mean of the sample is $\overline{x}$ and it is calculated:

$$\overline{x} = \frac{\sum_i x_i}{n}$$

where x_i is the each observed ith value of the vector of size n. The function to calculate it is `mean`. If the vector has very large or very small values, which can sometimes be considered unusual outliers, the mean will be greatly affected by the extreme values.

Median

It is a measure of central tendency, which is not affected by extreme values, hence it is said to be robust. For values x_1, …, x_n , the median of the sample is $\tilde{x}$, and would be calculated by ordering the observed values from smallest to largest and selecting the central value. If the sample size, n, is odd, it will be the central value, if it is even, it will be the mean of the central values. The function to calculate it is median.

Variance

To finish describing the distribution of the data it is necessary to study the dispersion. It is based on the deviation of the observed values from the mean, which would be: : $xi - \overline{x}$. The sum of all deviations is 0, so the deviation itself cannot be used as a measure of dispersion. However, if we remove the sign from the calculation of the deviation, for example by taking the absolute value of the deviation or by squaring it (as is the case for the variance), we can observe the dispersion. The variance, s^2, is based on the calculation of the squared deviations divided by the sample size, n:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

The function to calculate is `var`.

Standard deviation

It is also a measure of dispersion of sample data and is calculated from the square root of the variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

The function to calculate is `sd`. By taking the square root, this measure of dispersion is expressed, where appropriate, in the same units as the mean.

Quartiles

The quartile or percentile is the result of placing in four different blocks the values ordered from smallest to largest in a sample. Therefore there are 4 quantiles, they are values, Q_1, Q_2, Q_3 y Q_4. • Q_2 corresponds to the median. • Q_{1}1 will be the first quartile. • Q_3 is the third quartile. • Q_4 is the fourth quartile.

50% of the values lie between Q_1 and Q_3 . And the **interquartile range (IQR)** is the difference between the value Q_3 and Q_1 (IQR = Q_3 - Q_1). Not to be confused with the range, which will be the difference between the maximum and minimum value of the sample. The interquartile range will be a more robust measure as it is less sensitive to unusual values, as they are usually below Q_1 and above Q_3 . To visualise how the data are distributed in the quartiles, the **boxplot**. This graph also allows you to study symmetry, the distribution of data and to compare populations.

```
boxplot(diabetes$BMI, main="Boxplot BMI ", xlab="BMI", col="#a1e9f0")
```

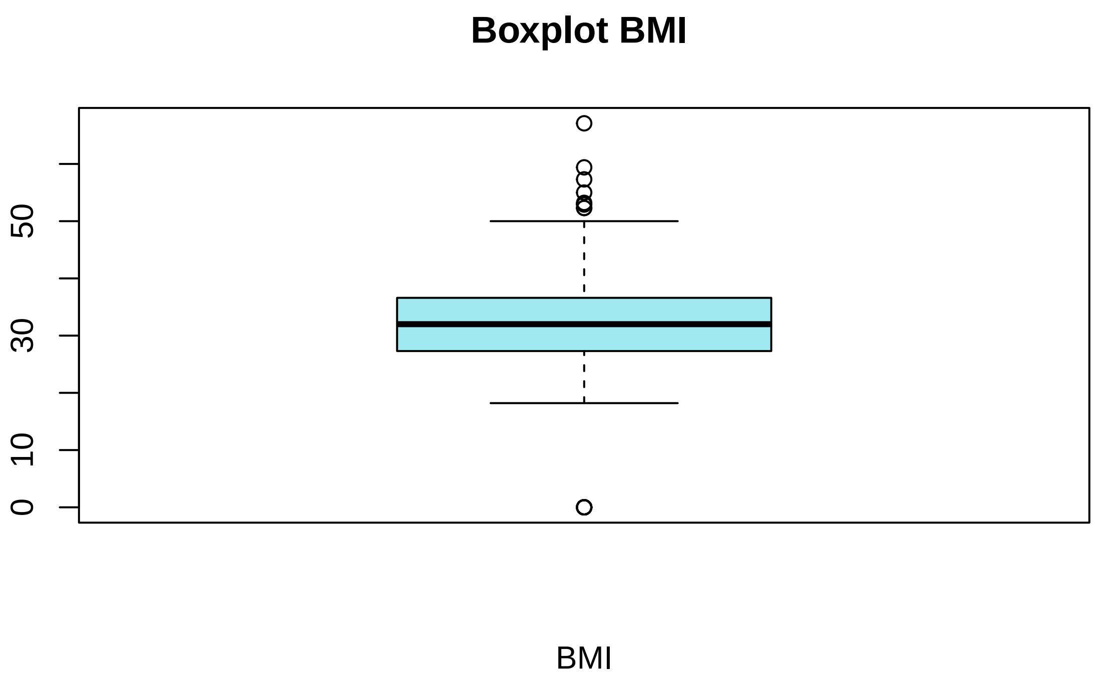

Figure 5: Boxplot

The bottom line of the box represents the Q_1, while the top line corresponds to the Q_3. The central horizontal line of the box corresponds to the Q_2. 50% of the data are contained in this box, i.e. it represents the interquartile range. The lines extending from the boxes, called whiskers, extend up and down to the value Q_3 + 1.5 x IQR and Q_1 - 1.5 x IQR. Data outside of these lines are potential *outliers*, and therefore, susceptible to be eliminated. However, this has to be done with caution.

Exercises

1. Produce an absolute frequency plot representing the data of the mother's race from the `birthwt` dataset.
2. Produce a relative frequency plot representing the data of the mother's race from the `birthwt` dataset.
3. Represent in a graph the density of the variable `lwt` of the dataset `birthwt` for 12 intervals (use the `break` parameter for the graph in order to specify the number of intervals). What is the shape of the graph?
4. Reason the following questions.
   - What will the shape of a histogram look like if the data have no *outlier*?
   - What will the shape be like if the data have *outliers* with very small values, and if they have very large *outliers*?
   - Do you think there are *outliers* in the `lwt` observations of the `birthwt` dataset?
5. Draw two diagrams showing the distribution of the data and the possible presence of outliers, as well as the interquartile ranges for:
   a) The variable `bwt` of `birthwt`.
   b) The variable `lwt` of `birthwt`.
6. Using the `summary` function, calculate manually the values Q_1, Q_3 and the interquartile range for:
   a) The variable `bwt` of `birthwt`.
   b) The variable `lwt` of `birthwt`.
7. Using the `summary` function, calculate manually how far the whiskers in question 6 extend for:
   a) The variable `bwt` of `birthwt`.
   b) The variable `lwt` of `birthwt`.

Bibliography

Crawley, Michael J. 2012. *The r Book*. John Wiley & Sons.

Mathur, Sunil K. 2009. *Statistical Bioinformatics with r*. Academic Press.

Seefeld, Kim, and Ernst Linder. 2007. "Statistics Using r with Biological Examples." *Durham: University of New Hampshire*.