Tutorial course N° 3

Descriptive statistics

Exercise 0. Change the working directory and create a script named `TP3.R` that you will use to save all the results of this session.

1. Quantitative variable

In this section, you will perform a descriptive analysis of the variables contained in the file `cats.txt`. The file contains data concerning the sex, the body-weight (in Kilograms) and the heart-weight (in grams) of cats.

Download the file `cats.txt` and import the dataset. To have an idea about the data structure, you can display the names of variables (columns) with the command

```
> names(cats)
```

and the number of observations (number of rows) with

```
> nrow(cats)
```

When the table is very large it is preferable to display *only* some rows of the dataframe. For instance, use the command `head(nom.data.frame,n)` to display only the first `n` rows.

Exercise 1. *Get familiar with the dataset.*

(1) What are the variable names ?

(2) How many variables and observations does the dataset contain ?

(3) Display the first 10 observations. What is the sex and weight of the heart of the cat number 6 ?

The `attach()` function allows to use the names of the variables in the table without recalling the name of the dataframe, i.e. instead of `cats$Bwt` just type `Bwt`. First try :

```
> Bwt
```

Next, execute the command :

```
> attach(cats)
```

Finally, retry :

```
> Bwt
```

Exercise 2. *Descriptive statistics*

(1) The following table shows the functions in `R` to calculate descriptive statistics of central tendency as well as statistics of dispersion. [1]

1. **Warning :** the sample variance calculated with the `var()` function in `R` returns the unbiased-variance given by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (see the lecture). Do not confuse with biased variance $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Statistic	function
Sample mean	`mean(variable_name)`
Sample median	`quantile(variable_name,0.5)`
Sample variance	`var(variable_name)`
Sample standard deviation	`sd(variable_name)`
Quantile of order p	`quantile(variable_name,p)`
Range	`diff(range(variable_name))`
Interquartile range	`IQR(variable_name)`

Calculate the sample mean for the variable `Bwt` ; the median, quartiles of order 1, 2 and 3 ; the sample variance and the standard deviation, the range and the interquartile range. Comment on the results.

(2) The command `summary()` returns descriptive statistics for a dataframe. Execute the command

```
> summary(cats)
```

and compare the obtained results with those of the previous question.

Exercise 3. *Boxplot*

(1) A **boxplot** is a graphical representation describing the main characteristics of a dataset (median, quartiles, minimum, maximum and outliers). The `R` function which plots the boxplot of a variable is `boxplot(nom_variable)`.
Plot the boxplots of the variables representing the body-weight and the heart-weight of the cats. Comment on the results obtained and compare them to those obtained with `summary()`.

Exercise 4. *Histogram of frequencies*

The `hist()` function displays the histogram of a variable in the dataset. By default, `hist()` represents the frequencies in the histogram, that is, the number of observations per class. With the option `freq=FALSE` the relative frequencies are plotted (so that the histogram has a total area of one). Additional options of the function `hist()` allow to :

— define the number of classes, with the option `breaks=n` we obtain a histogram with `n`+1 classes.
— define the intervals on which the histogram is built. With `breaks=vec` on can obtain a histogram for which the limits of the intervals (the classes) are given by the values of the vector `vec`,
— change the color : e.g. `col='blue'`

(1) Plot the histogram associated with the variable of the weight of cats. Try different values for the number of classes : 2, 20, 200 and 2000. What do you observe ? What number of classes is preferable ?

(2) The command `hist()` returns a list-type object that allows you to find the elements of the frequency table associated to the histogram. For example, if the object returned by `hist()` is `histo`, the command `histo$breaks` allows to find the intervals of classes, `histo$counts` returns the frequencies for each interval. Draw the table of frequencies associated to a histogram with 4 classes.

2. Qualitative variable

A qualitative variable, also called categorical, or factor is a variable whose taken on values are *categories*, which can be ordinal or nominal. The categories are also called *levels*. For instance, the variable `Sex` :

```
> class(Sex)
[1] "factor"
```

The function `levels()` is used to find out the set of values taken on by a qualitative variable.

```
> levels(Sex)
[1] "F" "M"
```

To graphically analyze qualitative variables, we can plot bar charts using the functions `barplot()`. The argument of the function is a vector with the heights of the bars. This latter can be obtained from the function `table()` which returns the frequency table of a qualitative variable. For instance, type

```
> table(Sex)
```

Exercise 5.

(1) Use the function `table()` to calculate the relative frequency of each category of the variable `Sex`.

(2) Use the `barplot()` function to display a barplot for the variable `Sex`.

(3) It is also possible to plot a pie chart using the function `pie()`. Draw the pie chart. Which representation do you prefer ?

Qualitative variables are useful for plotting boxplots by group. More precisely, the following command can be used to build the boxplots of the variable `var.num` by group according to the values taken on by the variable `var.factor` :

```
> boxplot(var.num ~ var.factor)
```

Note that the command `plot(var.num ~ var.facteur)` is equivalent to the previous one.

Exercise 6. * Plot the boxplots of the weight of male and female cats and interpret the result.

3. Study of the relationship between two quantitative variables

To study the relationship between two quantitative variables, we can display the scatter plot.

Exercise 7.

(1) Use the `plot()` function to display the scatter plot of the variables `(Bwt,Hwt)`. What can you say about the relationship between these two variables ?

(2) Calculate the sample covariance and correlation coefficient of these variables using the functions `cov()` and `cor()`. Comment on the results.

(3) To visualize the impact of the sex gender on the variables `Bwt` and `Hwt`, plot the points using different colors for males and females. Comment on the plot.