Laboratory 2 - Descriptive Statistics

Descriptive statistics describes the main features of a sample and consists in finding simple measures of sample and graphical representations.

Univariate analysis is the study of only one variable or attribute of the sample. This **variable** is a property that can be measured or just named. We have two different classifications of variables:

a1) Discrete variable: the values are finite points in a certain interval. Examples: the number of highway accidents, difficulty levels (very easy, easy, normal, difficult, very difficult etc), type of race (mongoloid/caucasian/negroid).

a2) Continuous variable: the values can be any point in a certain (real) interval. Examples: weight, height, speed.

b1) Quantitative variables:

- discrete (number of errors, number of children per family etc);
- continuous (speed, volume, weight etc).

b2) Qualitative variables (or categorical)[1]:

- ordinal (worse, good, better; difficulty levels);
- nominal (european/non-european; employed/jobless; married/unmarried/divorced/widow).

I. Graphical representations of the sample distribution

Data are grouped in classes (usually intervals): we sort the value in each interval and associate to each class a frequency. (Frequencies can be percentages).

RStudio. Don't forget to set the working directory: **Session → Set Working Directory → Choose Directory**.

Types of graphical representation:

1. **Stem and leaf** plot: for quantitative variables (usually discrete) and small samples (at most 30 - 40 individuals). Example. For the following sample (quantitative variable)

$$\begin{array}{cccccccccc} 0.6 & 0.2 & 1.6 & 2.0 & 1.1 & 0.5 & 1.5 & 2.3 & 3.4 & 1.9 \\ 0.4 & 0.5 & 1.2 & 0.9 & 2.1 & 1.6 & 1.8 & 2.6 & 3.1 & 2.5, \end{array}$$

the stem is the digit to the left of the decimal point, and the leaf is the digit to the right:

$$\begin{array}{c|cccccc} 0 & 6 & 2 & 5 & 4 & 5 & 9 \\ 1 & 6 & 1 & 5 & 9 & 2 & 6 & 8 \\ 2 & 0 & 3 & 1 & 6 & 5 \\ 3 & 4 & 1 \end{array}$$

Solved exercise. Draw a stem-and-leaf plot for the following sample

$$11\ 14\ 21\ 32\ 17\ 24\ 21\ 35\ 52\ 44\ 21\ 28\ 36\ 49\ 41\ 19\ 20\ 34\ 37\ 29$$

[1] Qualitative variables are discrete because we can associate to each value a number: 1, 2, 3 etc.

```
> x = c(11, 14, 21, 32, 17, 24, 21, 35, 52, 44, 21, 28, 36, 49, 41, 19, 20, 34, 37, 29)
> stem(x)
 The decimal point is 1 digit(s) to the right of the |
 1 | 1479
 2 | 0111489
 3 | 24567
 4 | 149
 5 | 2
```

2. **Histograms**: we split the range in a number of intervals[2] and display the corresponding frequencies as bars. The function used is *hist*().

   Solved exercise. In the file sample1.txt we have a sample for which the histogram is drawn as follows

```
> sample = scan("sample.txt")
> min = min(sample)
> max = max(sample)
> min
[1] 41
> max
[1] 96
```

   We can choose the intervals $[40, 50)$, $[50, 60)$ etc the last being $[90, 100)$ - there are six intervals. The histogram is displayed with

```
> interval = seq(40, 100, 10)
> hist(sample, breaks = interval, right = F, freq = T)
```

   Or with

```
> int = 6
> hist(sample, breaks = int, right = F, col = "blue")
```

   - *breaks* is the vector of interval boundaries (from 40 to 100) or a number which gives the number of intervals.
   - *right* is TRUE if the intervals are right closed and FALSE otherwise,
   - a similar parameter, *include.lowest*, concerns the left limits,
   - *freq* is TRUE if we represent the frequencies and FALSE otherwise - although the relative height of the bars will be the same.

3. **Bar chart** (Pareto): it is a similar representation used mainly for discrete variables. First we have to compute the frequencies and then display them. The used function is *barplot*().

   Solved exercise. Suppose that the following values are the frequencies for a certain sample.

   $$9\ \ 8\ \ 12\ \ 3\ \ 17\ \ 41\ \ 29\ \ 35\ \ 40\ \ 19\ \ 8$$

   We represent them like this

[2] A rule for choosing their common length: $L = 1 + \frac{\ln n}{ln2}$

```
> frecv = c(9, 8, 12, 3, 17, 41, 29, 35, 32, 40, 19, 8)
> barplot(frecv, space = 0)
```

Proposed exercises.

I.1 Draw a stem-and-leaf plot for the sample from "sample1.txt".

I.2 The file "unemploy2012.csv" contains the unemployment rates from 2012 in a majority of european countries (with two columns named 'country' and 'rate'). Draw the corresponding histogram using as intervals $(0, 4]$, $(4, 6]$, $(6, 8]$, $(8, 10]$, $(10, 12]$, $(12, 14]$ and $(14, 30]$.

Hint: read the sample like follows

```
> tablou = read.csv("unemploy2012.csv", header = T, sep = ';')
> rate = tablou['rate']
```

I.3 The file "life_expect.csv" contains the life expectancy (at birth, in 2012) in some european countries (with three columns named 'country', 'female' and 'male'). Draw the histograms for males and females using seven intervals in each case.

II. *Central tendency* analysis

Central tendency analysis is an approximation of the "center" of the distribution. (we will suppose that the values are sort in increasing order $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$, although not all of the statistics that follows need this ordering). The most important measures of the central tendency

- ***Mean*** - the arithmetic mean of all values; for the following sample

$$3, 6, 4, 3, 6, 7, 8, 5$$

the mean is $M = (3 + 6 + 4 + 3 + 6 + 7 + 8 + 5)/8 = 42/8 = 5.25$

$$M = \frac{1}{n}\left(\sum_{k=1}^{n} x_k\right)$$ in R: *mean(sample)*

- ***Median***: if n (the size of the sample) is odd the median is the value positioned in the middle, otherwise the median is the mean of the two middle values.

For the sample $3, 6, 4, 3, 6, 7, 8, 5$, after ordering: $3, 3, 4, 5, 6, 6, 7, 8$, we find $Me = \frac{5+6}{2} = \frac{11}{2} = 5.5$.

For the sample $3, 6, 4, 5, 2, 6, 9, 7, 8, 5, 4$, after ordering: $2, 3, 4, 4, 5, 5, 6, 6, 7, 8, 9$, $Me = 5$.

$$Me = \begin{cases} x_{k+1}, & \text{if } n = 2k+1 \\ \frac{x_k + x_{k+1}}{2}, & \text{if } n = 2k \end{cases}$$ in R: *median(sample)*

- ***Mode*** is the most frequent value in the sample. If there are many such values we say that we have a multi-modal distribution.

For the sample $3, 6, 4, 3, 6, 7, 8, 5, 3, 6,$ values 3 and 6 are the most frequent - we have a bimodal distribution.

For the sample $2, 6, 4, 3, 6, 7, 8, 5, 6, 4,$ the mode is 6 - the distribution is unimodal.

In R there is no specific function to compute the mode (only some packages has such functions).

Proposed exercises.

II.1 Compute the mean and the median for the sample from "sample1.txt".

II.2 Compute the mean and the median for the samples from "life_expect.csv".

II.3* Write a function to compute the mode for a given sample.

III. Variability and outliers.

The variability (or the *dispersion* of data) is a group of numerical characteristics that measure the spread of the data around the center.

- the ***range*** is the difference between the maximum and the minimum of the sample.

  For the sample $2, 6, 4, 3, 6, 7, 8, 5, 6, 4,$ the range $8 - 2 = 6$.

$$\boxed{Range = \max_{1 \leqslant k \leqslant n} x_k - \min_{1 \leqslant k \leqslant n} x_k}$$

- ***standard deviation of the sample*** ($\boldsymbol{s}$)

$$\boxed{s = \sqrt{\frac{\sum_{k=1}^{n}(x_k - M)^2}{n-1}}} \text{ in R: } \boxed{sd(sample)}$$

- ***variance of the sample*** ($\boldsymbol{s^2}$):

$$\boxed{s^2 = \frac{\sum_{k=1}^{n}(x_k - M)^2}{n-1}} \text{ in R: } \boxed{var(sample)}$$

- ***quartile*** and ***interquartile range*** (***IQR***): the first quartile Q_1 is the median of the sub-sample starting with the minimum (x_1) and ending with the median, the third quartile Q_3 is the median of the sub-sample starting with the median and ending with the maximum (x_n).

  The function *quantile*(*sample*) returns an object(*date frame*) that contains the following values: the minimum, the first quartile, the median, the third quartile, and the maximum. A certain quartile can be obtained like this

Q_i in R: $\boxed{as.vector(quantile(sample))[i+1]}$

$$\boxed{IQR = Q_3 - Q_1}$$

Solved exercise. The function *summary*(*sample*) determines the five (six) number summary: min, Q_1, Me, M, Q_3, and max.

```
> sample = c(9, 8, 12, 3, 17, 41, 29, 35, 32, 40, 19, 8)
> summary(sample)
   Min.   1st Qu.  Median  Mean   3rd Qu.  Max.
   3.00    8.75    18.00   21.08   32.75   41.00
```

- ***Outliers*** are those values which are too large or too small compared with the "center". Outliers can be the result of erroneous measurement, but can have natural causes. We can remove the outliers by two methods:

  - using the sample standard deviation: outliers are all those values that do not belong to the interval[3] $(M - 2s, M + 2s)$.
  - $(1.5 \cdot IQR$ rule) using the quartiles: outliers are all those values that do not belong to the interval[4] $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$.

  Solved exercise. Determine the outliers of the following sample using one of the above methods.

  1 91 38 72 13 27 11 19 5 22 20 19 8 17 11 15 13 23 14 17

```
> sample = c(1, 91, 38, 72, 13, 27, 11, 85, 5, 22, 20, 19, 8, 17, 11, 15, 13, 23, 14, 17)
> m = mean(sample)
> s = sd(sample)
> outliers = vector()
> j = 0
> for(i in 1:length(sample))
>    if(sample[i] <m - 2*s — sample[i] > m + 2*s) {
>       j = j + 1
>       outliers[j] = sample[i]
>    }
> outliers
[1] 91 85
```

Proposed exercises.

III.1 Write a function (within a script), named *outliers_mean(sample)*, that finds the outliers from a sample using the first method. Test your function on the example we already solved.

III.2 Write a function (within the same script), *outliers_iqr(sample)* that finds the outliers from a sample using the second method (3/2 IQR).

III.3 Use *summary*() and the functions *outliers_mean(sample)*, *outliers_iqr(sample)* on the sample from the file "sample2.txt". The outliers are the same?

RStudio. After editing, the script is saved (**Ctrl+S**) with a name like "my_script.R" and can be loaded with **Code** → **Source File** (**Ctrl+Shift+O**) or from the command line **source(script_file)**

RStudio. Once loaded, you can call any function belonging to this script from the command line: **normal_density(8)** or from the edit panel: we select the lines and use **Ctrl+Enter**; the entire script can be executed with **Ctrl+Alt+R**.

[3] More general the interval is $(M - k \cdot s, M + k \cdot s)$, $k \in \mathbb{R}_+$.
[4] More general the interval is $(Q_1 - k \cdot IQR, Q_3 + k \cdot IQR)$, $k \in \mathbb{R}_+$.