

Unit 1

Descriptive Statistics

ESIGELEC
Instructor: Federico Perea

1

Contents

- Data sets
- Population, sample, random variable
- Graphical representation
- Average
- Standard deviation

2

2



3

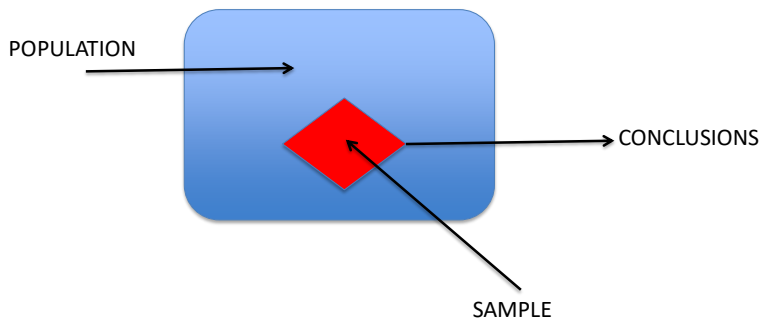
The objective of statistics is twofold:

- Collecting data with relevant information about a given population.
- Analyzing these data in order to extract information out of them

4

Inference (Units 4,5,6)

- We use inferential statistics when conclusions about populations are formed from sample data. Before that, we need some probability concepts!



5

Descriptive statistics

- **Data sets:** Data obtained from observation, pools, experiments, etc.. They are called *sample*, and a sample is extracted from the *population*.
- Data are normally organized as a table or matrix (rows and columns) so that:
 - Each row represents one element of the sample.
 - Each column represents one observed characteristic.

6

An example

Cardata.sf									
	mpg	cylinders	displace	horsepower	accel	year	weight	origin	make
1	43.1	4	90	48	21.5	78	1985	2	Volkswagen R
2	36.1	4	98	66	14.4	78	1800	1	Ford
3	32.8	4	78	52	19.4	78	1985	3	Mazda
4	39.4	4	85	70	18.6	78	2070	3	Datsun
5	36.1	4	91	60	16.4	78	1800	3	Honda
6	19.9	8	260	110	15.5	78	3365	1	Oldsmobile
7	19.4	8	318	140	13.2	78	3735	1	Dodge
8	20.2	8	302	139	12.8	78	3570	1	Mercury
9	19.2	6	231	105	19.2	78	3535	1	Pontiac
10	20.5	6	200	95	18.2	78	3155	1	Chevrolet
11	20.2	6	200	85	15.8	78	2965	1	Ford
12	25.1	4	140	88	15.4	78	2720	1	Ford
13	20.5	6	225	100	17.2	78	3430	1	Plymouth
14	19.4	6	232	90	17.2	78	3210	1	AMC
15	20.6	6	231	105	15.8	78	3380	1	Buick
16	20.8	6	200	85	16.7	78	3070	1	Mercury
17	18.6	6	225	110	18.7	78	3620	1	Dodge
18	18.1	6	258	120	15.1	78	3410	1	AMC
19	19.2	8	305	145	13.2	78	3425	1	Chevrolet
20	17.7	6	231	165	13.4	78	3445	1	Buick
21	18.1	8	302	139	11.2	78	3205	1	Ford
22	17.5	8	318	140	13.7	78	4080	1	Dodge
23	30	4	98	68	16.5	78	2155	1	Chevrolet
24	27.5	4	134	95	14.2	78	2560	3	Toyota
25	27.2	4	119	97	14.7	78	2300	3	Datsun
26	30.9	4	105	75	14.5	78	2230	1	Dodge
27	21.1	4	134	95	14.8	78	2515	3	Toyota
28	23.2	4	156	105	16.7	78	2745	1	Plymouth
29	23.8	4	151	85	17.6	78	2855	1	Oldsmobile

7

- First step: simple analysis of data → Descriptive statistics.
- What for?
 - To observe characteristics.
 - To summarize the information by means of :
 - **Statistics** (a numerical measurement describing some characteristics of a sample)
 - **Graphic representations**

8

- From now on we will use the following keywords:

Population

Random variable

Random sample

Statistical data

9

- **Population:** Is the complete collection of elements (scores, people, measurements, etc.) to be studied.
- **Example :** If you want to draw a study about the result in the following elections in France, the population will be the millions of people with right to vote in France.
- **Example :** If you want to study the quality of certain laptop model, the population would be all laptops of this model.

Elements : People, computers, etc.. All of them form the population. The individuals that form the population.

10

Randomness

- **Random experiment** : Is a process that, when it is repeated, generates the different elements of the population. The result is, in principle, unknown!
- **Random variable(RV)** : A characteristic that associates a single numerical value with each outcome of a random experiment. They can be qualitative or quantitative.
- **Example:** Random experiment: roll a dice. Random variable: number obtained

11

Discrete vs. Continuous

- **Discrete or continuous?**
 - Score when rolling a dice
 - Number of defective units in a production chain
 - Height of a person
 - Eye color of a person
 - Life time of a computer
 - Weight of a chair
 - Score obtained by a student in an exam
 - Number of defective screws in a box
 - Width of a screw box

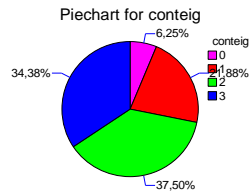
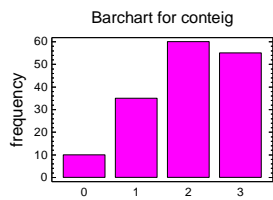
12

Sample

- **Random sample** : In general, it is not possible to study all the elements in the population
 - Infinite populations (or too many elements)
 - Economical reasons
- In statistics we always work with a subset of the population. This subset is the **sample**. The sample must **represent** the population, so it is possible to extrapolate the conclusions obtained in the sample to the complete population (remember that we are interested in studying the population, not the sample).
- One way to obtain a representative sample is by using randomness, that is, by choosing a random sample.
- **Statistical data** : When a random sample is selected from a population, and characteristics (random variables) are observed, we have a set of statistical data.
- Usually, we denote a sample as a collection of numbers: x_1, x_2, \dots, x_n (sample size n)

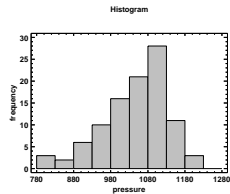
13

Graphical representation



show data in graphical representation

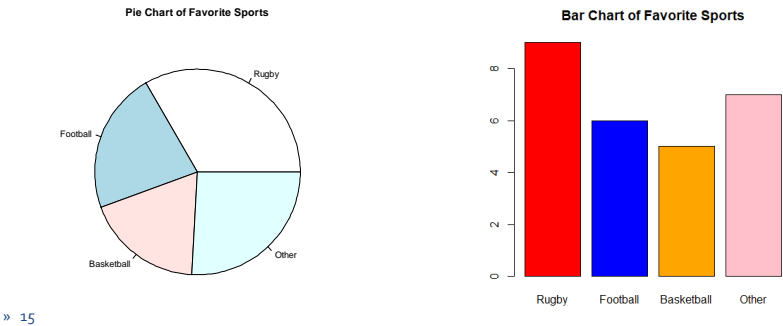
firstly!



Variables with few possible values: barcharts, piecharts,...
Variables with many possible values: histograms,...
And many others. Which one to use? Depends on the sample

14

- **Exercise :** When asking a group of people about their favorite sport, 9 of them said Rugby, 6 of them Football, 5 of them Basketball, and the other 7 chose other answer. Draw a piechart and a barchart for this exercise. (*Unit1_Piechart.r*)



Measures

- Graphics are useful but limited.
- Data can be summarized numerically, so they can be more easily represented and compared.
- We will use **parameters** (numerical measurements describing some characteristic of a population) and **statistics** (numerical measurements describing some characteristics of a sample).
- Three main types: LOCATION, DISPERSION, SHAPE

The mean

- **Mean:** Also known as average, is the Location parameter most commonly used.
- It corresponds with the idea of “distributing in equal shares”.
- Calculus: $\bar{x} = \frac{x_1 + \dots + x_n}{n}$

» 17

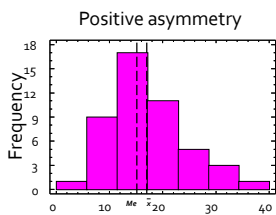
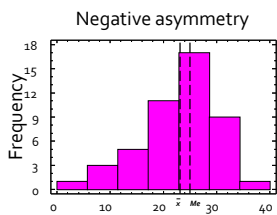
The median

- **Median :** In presence of asymmetry or outliers, it is more recommended than the mean.
- It follows the idea of a “central value”.
- How to compute it.: **Me** or \tilde{x} =

- If n is odd: **Central value** (the one in position $(n+1)/2$).
 - If n is even: **Average of the two central values** (those in positions $n/2$ and $n/2 + 1$).
- Note that data must be sorted in increasing order!

» 18

- **Median**: It is the value leaving 50% of data above, and 50% of data below.
- It is more “stable” (**robust**) than the mean, in the sense that wrong data and outliers affect it less than they affect the mean:



» 19

- **Quartiles** : the three points that divide the data set into four equal groups.
 - **Q1** : Value that leaves 25% below it, and 75% above it.
 - **Q2** : Value that cuts data set in half
 - **Q3** : Value that leaves 75% below it, and 25% above it.

» 20

- $Q2$: Coincides with the **median**.
- The “central” 50% of data are between $Q1$ and $Q3$.
- How to calculate $Q1$ and $Q3$:
 - $Q1 \approx$ median of the “first half” of data.
 - $Q3 \approx$ median of the “second half” of data.

» 21

- Alternatively, $Q1$ and $Q3$ can be calculated as follows:

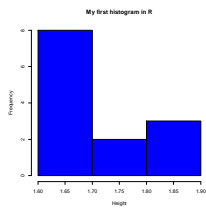
- Sort the observations in increasing order.
- Unless n is multiple of 4, $Q1$ is the value in position $n/4$ (rounding up to the nearest larger integer if necessary). If n is multiple of 4, $Q1$ is the median of the first $n/2$ data.
- $Q3$ is calculated “symmetrically”.

» 22

- **Exercise :** Calculate the **mean**, the **median** and the **quartiles** of the heights of the following randomly chosen students in ESIGELEC. Also draw a histogram with classes of length 0.1. Do the exercise both by hand and using R (*Unit1_Heights.r*)

1,64 | 1,66 | 1,74 | 1,86 | 1,87 | 1,69 | 1,65 |
1,71 | 1,68 | 1,70 | 1,62 | 1,84 | 1,61

SOL: $\bar{x} = 1.71$; Me = 1.69; $Q_1 = 1.65$; $Q_3 = 1.74$



» 23

- **Percentiles :** They generalize the quartiles.
- **Percentile p (or p th percentile)** is the point leaving below some $p\%$ of data.
- It gives the proportion of data below and above a given value.

- **Example:** (Used by pediatricians) “Your baby is in the 70th height percentile, and 50th weight percentile.”

» 24

Dispersion measures

- **Famous quote:** “Statistics is a science that shows that if my neighbor has two cars and I none, we both have one.”
— George Bernard Shaw (1856–1950), Irish playwright and a co-founder of the London School of Economics

- The mean represents “proportional sharing”, but...

- **Example:** What is the average score in an exam if half of the students got 10 and the other half 0? How about if all of them got 5?
- **Example :** You are about to jump in a lake from a high rock. You know that the average depth of the lake is 1.40 m. Would you jump?

» 25

- Location parameters do not give me information about how similar or different data are.
- Are my data close to each other? Or on the contrary, is there much dispersion?
- We will study the following parameters and statistics related to dispersion:

Range isn't robust(stable), influence of one element is important sometimes

Range	Interquartile range <small>more advisable Q3-Q1</small>
Variance	Standard deviation

» 26

- **Maximum** : x_{\max} = maximum observed value
- **Minimum** : x_{\min} = minimum observed value
- **Range**.

$$x_{\max} - x_{\min}$$

- The range gives us information about the difference between the two most separated data.

- **Exercise** : Give a main drawback of this measure

» 27

- **Interquartile range** . $IQR()$

$$IR = Q3 - Q1$$

- IR gives information about the dispersion found in “central” values.
- It is “**robust**” in the sense that outliers do not affect it.

» 28

- **Exercise :** Calculate the **range** and the **interquartile range** of the data about student's heights. *SOL:* $Range = 0.26$; $I.R. = 0.09$ (*Unit1_Heights.r*)

- Ranges give some valuable information, they are easily calculated, but...
- Could we use them to calculate how far the observations from the mean on average are?
- **Variance and standard deviation** give this information.

» 29

always positif

- **Variance (sample) .**

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

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

- In some textbooks you will find it as **quasi-variance**, also denoted as s^2_{n-1} or $(s')^1$.

» 30

- Instead of the variance, it is more common to use its square root, because it is expressed in the same units as the data.
- **Standard deviation (sample)** . Calculus:

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

» 31

Describing data: summary

There are many measures to describe a sample: mean, median, quartiles, percentiles, standard deviation, range, ...

In this course you will use, mainly, two of them:

- The sample mean (known): $\bar{x} = \frac{x_1 + \dots + x_n}{n}$; *mean()*
- The sample standard deviation (known): *sd()*

$$S = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}};$$

32

- At times, it is useful to have a dispersion measure that does not depend on the units in which data are measured, that is, a *dimensionless* data.
- Coefficient of variation .

$$CV = \frac{s}{\bar{x}} 100\%$$

» 33

- In short : Depending on the data, what parameters should you use?

	Symmetric data without outliers	Asymmetric data or presence of outliers
Location	Mean	Median
Dispersion	Standard deviation	Interquartile range

- OUTLIER: an observation that lies an *abnormal* distance from other values in the sample (e.g., more than 1.5 times the interquartile range far from the nearest quartile)

» 34

Shape measures

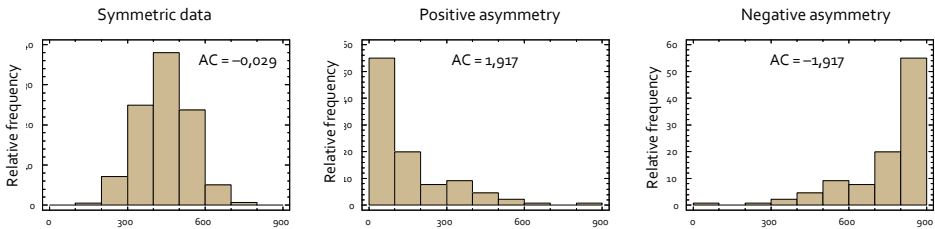
- Asymmetry coefficient (also called *Skewness coefficient*) and Kurtosis coefficient are the most commonly used: *skewness()*, *kurtosis()*
- Both together allow us to check whether or not our data follow a “Gaussian” or “bell-shaped” curve (Normal distribution).
- **Asymmetry.** Calculus:

$$AC = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)s^3}$$

» 35

- Asymmetry: Under the hypothesis of normality, data should be symmetrical (i.e. skewness should be equal to zero)

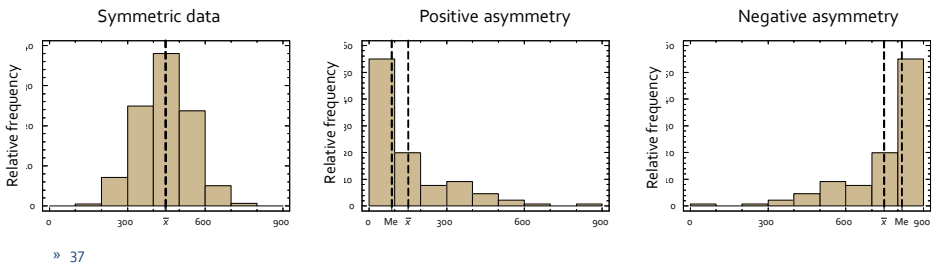
- $AC \approx 0$: Symmetric data
 - $AC > 0$: Positive asymmetry (right tailed)
 - $AC < 0$: Negative asymmetry (left tailed)



» 36

- Generally we have:

- Symmetric data $\Leftrightarrow \bar{x} \sim Me$
- Positive asymmetry $\Leftrightarrow \bar{x} > Me$
- Negative asymmetry $\Leftrightarrow \bar{x} < Me$



» 37

- **Kurtosis coefficient:** It measures the “tailedness” in the data
- The reference is the Gaussian curve.
- Calculus in R:

$$KU = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_j \frac{(x_j - \bar{x})^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

- Different software may compute this coefficient in different ways

» 38

- “tailedness” degree :
- In R, under the hypothesis of normality, data should have kurtosis equal to 3

- $KU \approx 3$: “Normal” data (bell-shaped)
- $KU > 3$: More acute peak around the mean
- $KU < 3$: Wider and lower peak around the mean

» 39

Box-and-Whisker plot

- The **Box-and-Whisker** plot allows you to represent the main features about location and dispersion. *boxplot()*
- **Procedure** : Draw a box and two whiskers.
- **Box**:
 - Left side : $Q1$
 - Right side : $Q3$
 - Vertical line : $Q2$ (Median)
 - Point or cross: **Mean** (optional)

$\left. \begin{matrix} Q1 \\ Q3 \end{matrix} \right\} \Rightarrow \text{Width: Interquartile range}$

» 40

- **Whiskers:**
 - The maximum length of each whisker will be 1.5 times the box width.
 - Each whisker stops in the last value that DOES NOT exceed such length.
 - Those values that are further than the whiskers, if any, are represented by dots, and called “outliers”.
- **Necessary computations :**
 - Quartiles ($Q1$, $Q2$, $Q3$) and mean (\bar{x})
 - Interquartile range (IR)
 - $1,5 \cdot IR$ (to determine whether or not there are outliers)

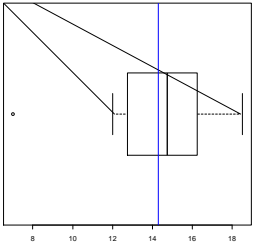
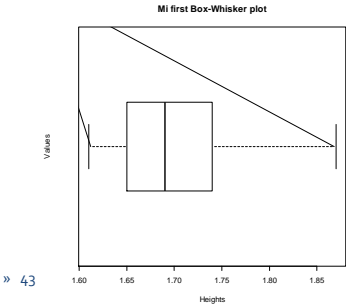
» 41

- The Box-and-Whisker box allows you to detect :
 - Asymmetry
 - Wrong data
 - Outliers
 - Differences between groups
- A look at the plot gives you:
 - Quartiles, median (and the mean, if it is depicted)
 - Interquartile range \Rightarrow 50% “central”
 - If observations are “symmetrically” distributed or not

» 42

• **Exercise :** Depict the Box-and-Whisker plot for the heights exercise.
(Unit1_Heights.r)

• **Exercise :** From the following data
12 | 14 | 14,5 | 17 | 13,5 | 18,5 | 16 | 15,5 | 15 | 7 | 12 | 16,5
draw the corresponding Box-and-Whisker plot. (Unit1_BoxWhisker.r)

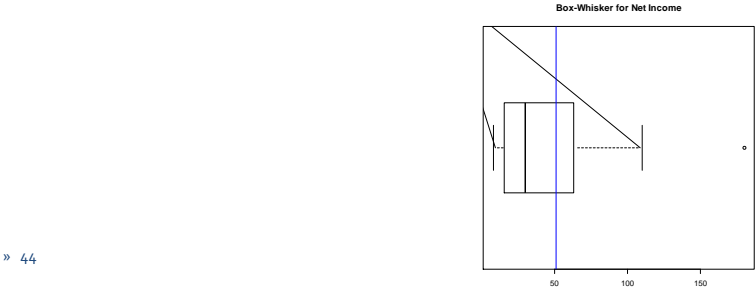


• **Exercise :** The following data represent the net incomes (in thousands of euros) of a sample constituted by 11 companies of a given sector:

25 | 110 | 42 | 10 | 8 | 180 | 70 | 14 | 56 | 17 | 30

a) What type of random variable do we have?
b) Depict the Box-and-Whisker plot.
c) What statistics would be appropriate to describe location and dispersion of data?

SOL: Continuous, Median and IR (as data are asymmetric with outliers).
(Unit1_Netincome.r)



Extra information: Subsets in R

- Sometimes it is useful to define one variable only for a subset of the sample.
- In R: *new.variable <-subset(variable,condition)*
- For example, define a new variable *Vble.restr* that only takes the values of another variable *Vble* if *Vble2* is equal to a certain *Value*
Vble.restr <- subset(Vble, Vble2 == Value)

45