Elementary Statistics Course

Alfredo Sánchez Alberca (asalber@ceu.es)

Feb 2016

Department of Applied Math and Statistics
CEU San Pablo

CEU
Universidad San Pablo

License terms

Contents

1 Intoduction to Statistics

1.1 Statistics as a scientific tool

What is Statistics?

Definition 1 (Statistics)**.** *Statistics* is a branch of Mathematics that deals with data collection, summary, analysis an interpretation.

The role of Statistics is to extract information from data in order to gain knowledge for taking decisions.

Statistics is essential in any scientific or technical discipline which require data handling, especially with large volumes of data, such as Physics, Chemistry, Medicine, Psychology, Economics or Social Sciences.

But, Why is necessary Statistics?

A changing World

Scientists try to study the World. A World with a high variability that makes difficult determining the behaviour of things.

Variability is the reason for Statistics!

Statistics provides a bridge between the real world and the mathematical models that attempt to explain it, providing a methodology to assess the discrepancies between reality and theoretical models.

This makes Statistics an indispensable tool in applied sciences that require design of experiments and data analysis.

1.2 Population and sample

Statistical population

Definition 2 (Population)**.** A *population* is a set of elements defined by one or more features that has all the elements and they alone. Every element of the population is called *individual*.

Definition 3 (Population size)**.** The number of individuals in a population is known as the *population size* and is represented by N.

Sometimes not all the individuals are accessible to study. Then we distinguish between:

Theoretical population: Individuals to which we want extrapolate the study conclusions.

Studied population: Individuals truly accessible in the study.

Drawbacks in the population study

Scientists study a phenomenon in a population to understand it, to get knowledge about it, and so to control it.

But, for a complete knowledge of the population it is necessary to study all his individuals.

However, this is not always possible for several reasons:

- The population size is infinite or too large to study all his individuals.
- The operations that individuals undergo are destructive.
- The cost, both money and time, that would require study all the individuals in the population is not affordable.

Statistics Sample

When it is not possible or convenient to study all the individuals in a population, we study only a subset of them.

Definition 4 (Sample)**.** A *sample* is a subset of the population.

Definition 5 (Sample size)**.** The number of individuals of the sample is called *sample size* and is represented by n.

Usually, the population study is conducted on samples drawn from it.

The sample study only gives an approximate knowledge of the population. But in most cases is *enough*.

Sample size determination

One of the most interesting questions that arise:

How many individuals are required to sample to have an approximate but enough knowledge of the population?

The answer depends of several factors, as the population variability or the desired reliability for extrapolations on the population.

Unfortunately we can't answer that question until the end of the course, but in general, the most individuals have the sample, the more reliable will be the conclusions on the population, but also the study will be longer and more expensive.

Sample size determination

Small sample of pixels of a picture

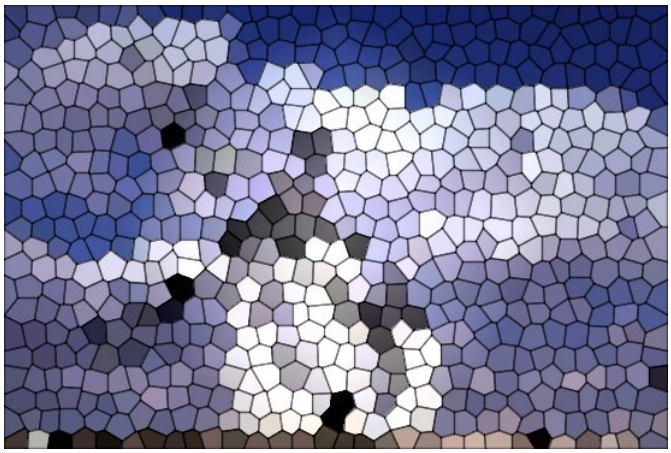
Example. To understand what a sufficient sample size means we can use a picture example. A digital photography consist in a lot of small points called pixels disposed in an big array layout with rows and columns (the more rows and columns, the more resolution has the picture). Here the picture is the population and every pixel is and individual. Every pixel has a colour and it's the variability of colours what forms the picture motif.

How many pixels must we take in a sample in order to find out the motif of a picture?

The answer depends on the variability of colours in the picture. If all the pixels in the picture are of the same colour, only one pixel is required to know the motif. But, if there is a lot of variability in the colours, a large sample size will be required.

The image below contains a small sample of the pixels of a picture.

Could you find out the motif of the picture?



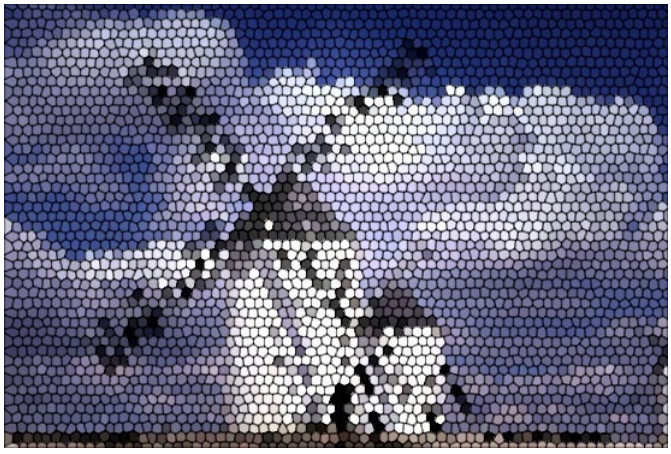
With a small sample size it's difficult to find out the picture motif!!

Sample size determination

Large sample of pixels of a picture

The image below contains a larger sample of pixels.

Could you find out the motif of the picture now?



With a large sample is easier to find out the picture motif!

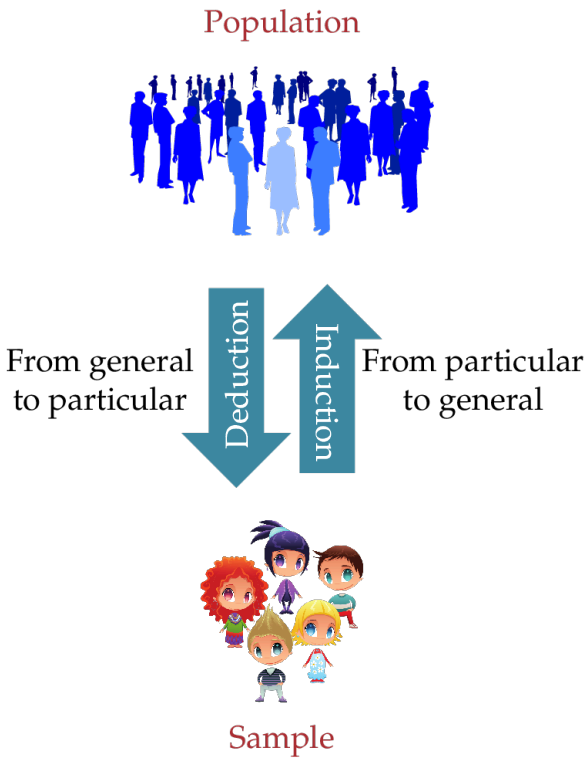
Sample size determination

Whole population of pixels of a picture

And here is the whole population.

It's not required to know all the pixels of a picture to find out its motif!

Types of reasoning



Types of reasoning

Deduction properties: If the premises are true, it guarantees the certainty of the conclusions (that is, if something is true in the population, it is also true in the sample). However, it does not provide new knowledge!

Induction properties: It doesn't guarantee the certainty of the conclusions (if something is true in the sample, it may not be true in the population, so be careful with the extrapolations!). But, ¡it is the only way to generate new knowledge!

Statistics is fundamentally based on inductive reasoning, because it uses the information obtained from samples to draw conclusions about populations.

1.3 Sampling

Sampling

Definition 6 (Sampling)**.** The process of selecting the elements included in a sample is known as *sampling*.

To reflect reliable information about the whole population, the sample must be representative of the population. That means that the sample should reproduce on a smaller scale the population variability.

Our goal is to get a representative sample!

Types of sampling

There exists a lot of sampling methods but all of them can be grouped in two categories:

Random sampling The sample individuals are selected randomly. All the population individuals have the same likelihood of being selected (equiprobability).

Non random sampling: The sample individuals are not selected randomly. Some population individuals have a higher likelihood of being selected than others.

Only random sampling methods avoid the selection bias and guarantee the representativeness of the sample, and therefore, the validity of conclusions.

Non random sampling methods are not suitable to make generalizations because doesn't guarantee the representativeness of the sample. Nevertheless, usually are less expensive and can be used in exploratory studies.

Simple random sampling

The most popular random sampling method is the *simple random sampling*, that has the following properties:

- All the population individuals have the same likelihood of being selected in the sample.
- The individual selection is performed with replacement, that is, each selected individual is returned to the population before selecting the next one. This way the population doesn't change.
- Each individual selection is independent of the others.

The only way of doing a random sampling is to assign a unique identity number to each population individual (conducting a *census*) and performing a random drawing.

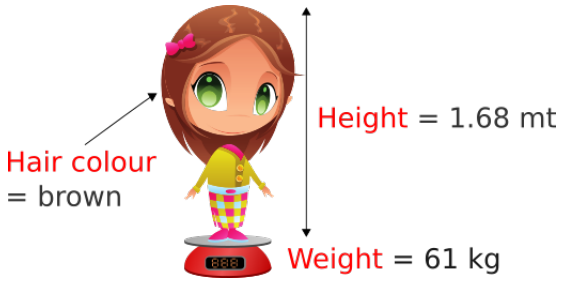
1.4 Statistical variables

Statistical variables and data

In every statistical study we are interested in some properties or characteristics of individuals.

Definition 7 (Statistical variable)**.** A *statistical variable* is a property or characteristic measured in the population individuals.

The *data* is the actual values or outcomes recorded on a statistical variable.



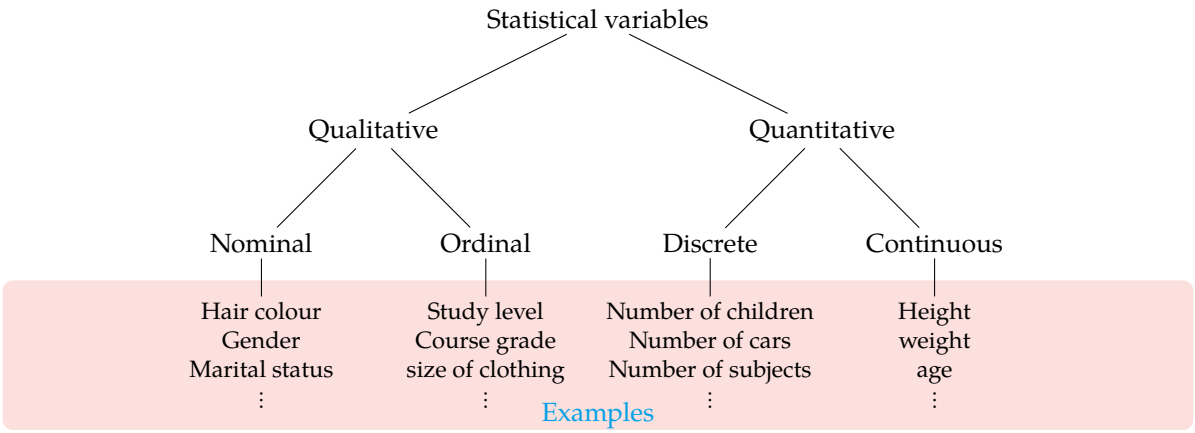
Types of statistical variables

According to the nature of their values and their scale, they can be:

- **Qualitative variables:** They measure non-numeric qualities. They can be:
  - **Nominal:** There is no natural order between its categories. Example: The eyes or hair colour.
  - **Ordinal:** There is a natural order between its categories. Example: The education level.
- **Quantitative variables:** They measure numeric quantities. They can be:
  - **Discrete:** Their values are isolated numbers (usually integers). Example: The number of children or cars in a family.
  - **Continuous:** They can take any value in a real interval. Example: The height, weight or age of a person.

Qualitative and discrete variables are also called *categorical variables* and their values *categories*.

Types of statistical variables



Types of statistical variables

Choosing the appropriate variable

Sometimes a characteristic could be measured in variables of different types.

Example Whether a person smokes or not could be measure in several ways:

- Smokes: yes/no. (Nominal)
- Smoking level: No smoking/unusual/moderate/quite/heavy. (Ordinal)
- Number of cigarettes per day: 0,1,2,…(Discrete)

In those cases quantitative variables are preferable to qualitative, continuous variables are preferable to discrete variables and ordinal variables are preferable to nominal, as they give more information.

Types of statistical variables

According to their role in the study:

- **Independent variables:** Variables that no depends on other variables in the study. Usually they are manipulate in an experiment in order to observe their effect on a dependent variable. They are also known as *predictor variables*.
- **Dependent variables:** Variables that depends on other variables in the study. They are not manipulated in an experiment and are also known as *outcome variables*.

Example In a study on the performance of students in a course, the intelligence of students and the daily study time are independent variables, while the course grade is a dependent variable.

Types of statistical studies

According to their role in the study:

- **Experimental:** When the independent variables are manipulated in order to see the effect that that change have on the dependent variables. **Example** In a study on the performance of students in a test, the teacher manipulates the study time and create two or more groups asking students in each group to study a different number of hours.
- **Non-experimental:** When the independent variables are not manipulated. That not means that it is impossible to do so, but it will either be impractical or unethical to do so. **Example** In a study a researcher could be interested in the effect of smoking over the lung cancer. However, whilst possible, it would be unethical to ask individuals to smoke in order to study what effect this had on their lungs. In this case, the researcher could study two groups of people, one with lung cancer and other without, an observe in each group how many persons smoke or not.

Experimental studies allow to identify a cause and effect between variables while non-experimental studies only allow to identify association or relationship between variables.

The data table

The variables of a study will be measured in each individual of the sample. This will give a data set that usually is arranged in a tabular form known as **data table**.

In this table each column contains the information of a variable and each row contains the information of an individual.

Example

Name	Age	Gender	Weight(Kg)	Height (cm)
José Luis Martínez	18	M	85	179
Rosa Díaz	32	F	65	173
Javier García	24	M	71	181
Carmen López	35	F	65	170
Marisa López	46	F	51	158
Antonio Ruiz	68	M	66	174

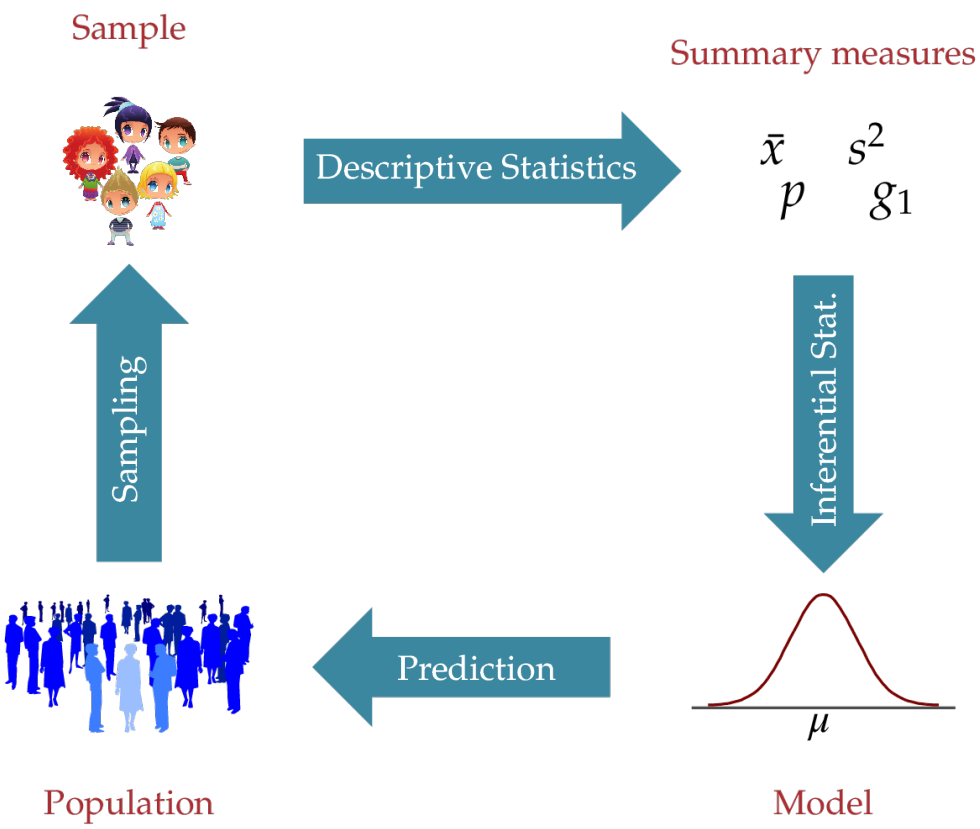
1.5 Phases of a statistical study

Phases of a statistical study

Usually a statistical study goes through the following phases:

1. The study begins with a previous design in which are set the study goals, the population, the variables to measure and the required sample size.
2. Next, the sample is selected from the population and the variables are measured in the individuals of the sample (getting the data table). This is accomplished by *sampling*.
3. The next step consists in describing and summarizing the information of the sample. This is the job of *Descriptive statistics*.
4. Then, the information obtained is projected on a mathematical model that intend to explain what happens in population, and the model is validated. This is accomplished by *Inferential statistics*.
5. Finally, the validated model is used to perform predictions and to draw conclusions on the population.

The statistical cycle


Sample
Summary measures
Descriptive Statistics
$\bar{x}$ s^2
p g_1
Inferential Stat.
Sampling
Prediction
μ
Population
Model

2 Frequency distributions: Tabulation and charts

Descriptive Statistics

Descriptive Statistics is the part of Statistics in charge of representing, analysing and summarizing the information contained in the sample.

After the sampling process, is the next step in every statistical study and usually consists of:

1. Classify, group and sort the data of the sample.
2. Tabulate and plot data according to their frequencies.
3. Calculate numerical measures that summarize the information contained in the sample (*sample statistics*).

It has no inferential power ⇒ Do not generalize to the population!

2.1 Frequency distribution

Sample classification

The study of a statistical variable starts measuring the variable in the individuals of the sample and classifying the values.

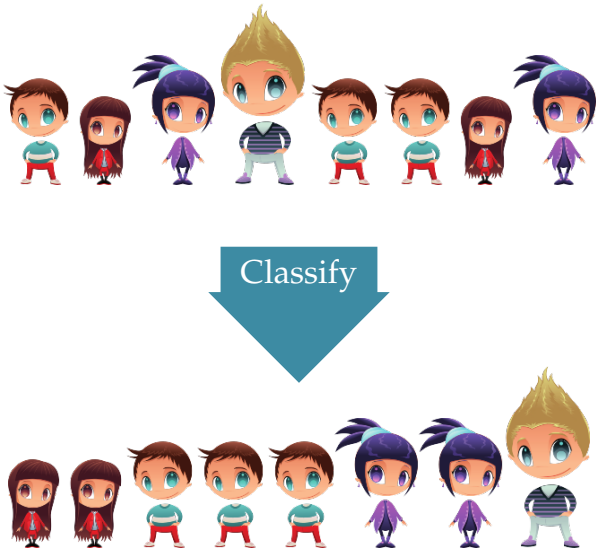
There are two ways of classifying data:

Non-grouping Sort values from lowest to highest value (if there is an order). Used with qualitative variables and discrete variables with few distinct values.

Grouping Group values in intervals (classes) and sort them from lowest to highest intervals. Used with continuous variables and discrete variables with many distinct values.

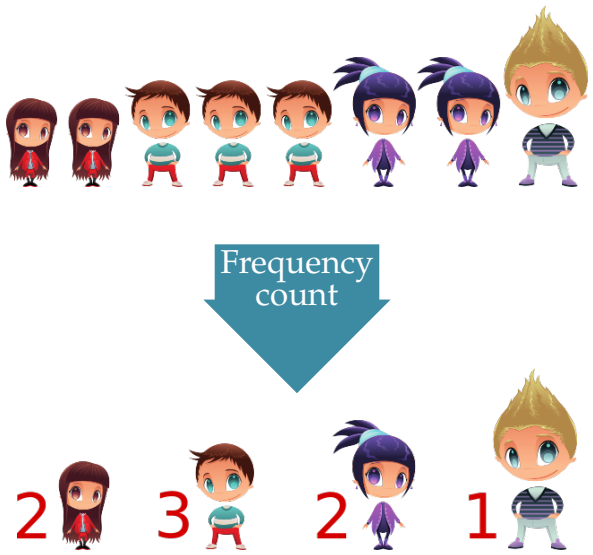
Sample classification

X =Height



Frequency count

X =Height



Sample frequencies

Definition 8 (Sample frequencies)**.** Given a sample of n values of a variable X, for every value x_i of the variable is defined

- **Absolute frequency** n_i: Is the number of times that value x_i appears in the sample.
- **Relative frequency** f_i: Is the proportion of times that value x_i appears in the sample.

$$f_i = \frac{n_i}{n}$$

- **Cumulative absolute frequency** N_i: Is the number of values in the sample less than or equal to x_i.

$$N_i = n_1 + \cdots + n_i$$

- **Cumulative relative frequency** F_i: Is the proportion of values in the sample less than or equal to x_i.

$$F_i = \frac{N_i}{n}$$

Frequency table

The set of values of a variable with their respective frequencies is called **frequency distribution** of the variable in the sample, and it is usually represented as a **frequency table**.

X values	Absolute frequency	Relative frequency	Cumulative absolute frequency	Cumulative relative frequency
x_1	n_1	f_1	N_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i	N_i	F_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	N_k	F_k

Frequency table

Example of quantitative variable and non-grouped data

The number of children in 25 families are:

$$1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, \ 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2$$

The frequency table for the number of children in this sample is

x_i	n_i	f_i	N_i	F_i
0	2	0.08	2	0.08
1	6	0.24	8	0.32
2	14	0.56	22	0.88
3	2	0.08	24	0.96
4	1	0.04	25	1
$\sum$	25	1		

Frequency table

Example of quantitative variable and grouped data

The heights (in cm) of 30 students are:

179, 173, 181, 170, 158, 174, 172, 166, 194, 185, 162, 187, 198, 177, 178, 165, 154, 188, 166, 171, 175, 182, 167, 169, 172, 186, 172, 176, 168, 187.

The frequency table for the height in this sample is

x_i	n_i	f_i	N_i	F_i
$(150, 160]$	2	0.07	2	0.07
$(160, 170]$	8	0.27	10	0.34
$(170, 180]$	11	0.36	21	0.70
$(180, 190]$	7	0.23	28	0.93
$(190, 200]$	2	0.07	30	1
$\sum$	30	1		

Classes construction

Intervals are known as **classes** and the center of intervals as **class marks**.

When grouping data into intervals, the following rules must be taken into account:

- The number of intervals should not be too big nor too small. A usual rule of thumb is to take a number of intervals approximately $\sqrt{n}$ or $\log_2(n)$.
- The intervals must not overlap and must cover the entire range of values. It doesn't matter if intervals are left-open and right-closed or vice versa.
- The minimum value must fall in the first interval and the maximum value in the last.

Frequency table

Example with qualitative variable

The blood type of 30 people are:

A, B, B, A, AB, 0, 0, A, B, B, A, A, A, A, AB, A, A, A, B, 0, B, B, B, A, A, A, 0, A, AB, 0.

The frequency table of the blood type is

x_i	n_i	f_i
0	5	0.16
A	14	0.47
B	8	0.27
AB	3	0.10
$\sum$	30	1

Why there are not cumulative frequencies?

2.2 Frequency distribution graphs

Frequency distribution graphs

Usually the frequency distribution is also displayed graphically.

Depending on the type of variable and if data has been grouped or not, there are different types of charts:

- Bar chart
- Histogram
- Line chart
- Pie chart

Bar chart

A **bar chart** consists in a set of bars, one for every value or category of the variable, plotted on a coordinate system.

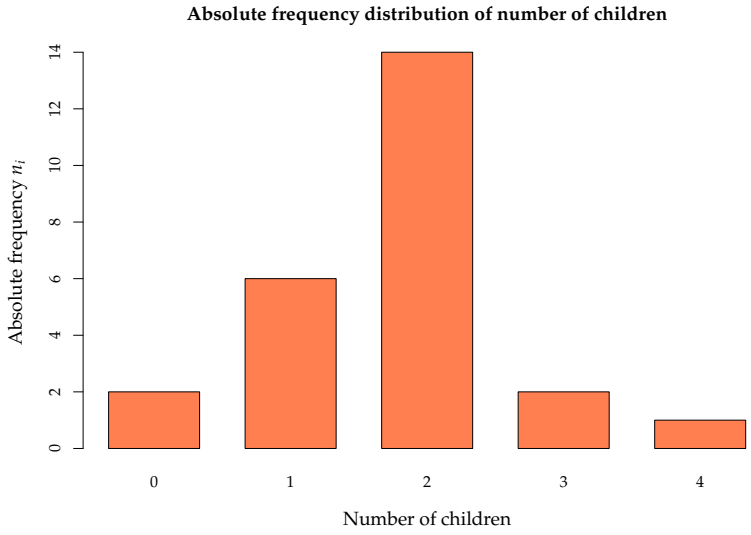
Usually the values or categories of the variable are represented on the x-axis, and the frequencies on the y-axis. For each value or category of the variable, a bar is draw to the height of its frequency. The width of the bar is not important but bars should be clearly separated among them.

Depending on the type of frequency represented in the y-axis we get different types of bar charts.

Sometimes a polygon, known as **frequency polygon**, is plotted joining the top of every bar with straight lines.

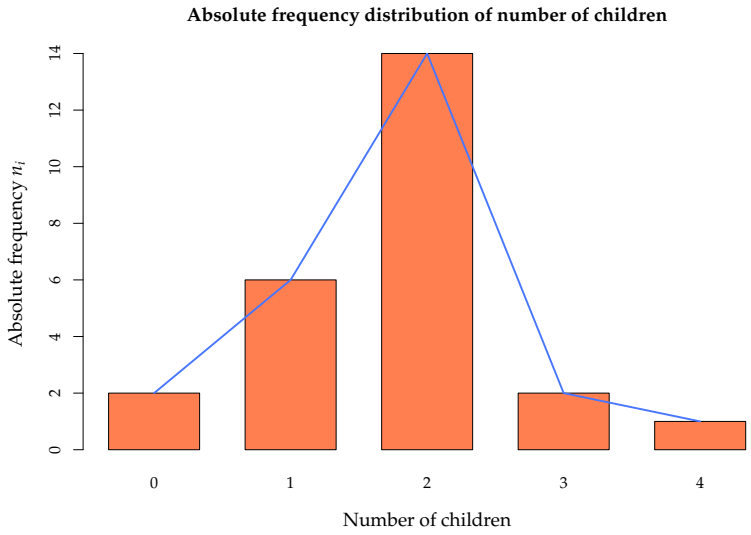
Absolute frequency bar chart

Non-grouped data



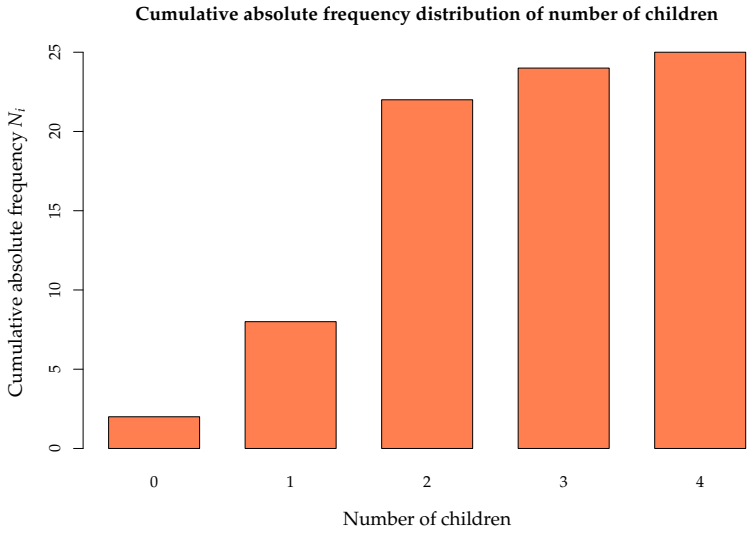
Absolute frequency line chart or polygon

Non-grouped data



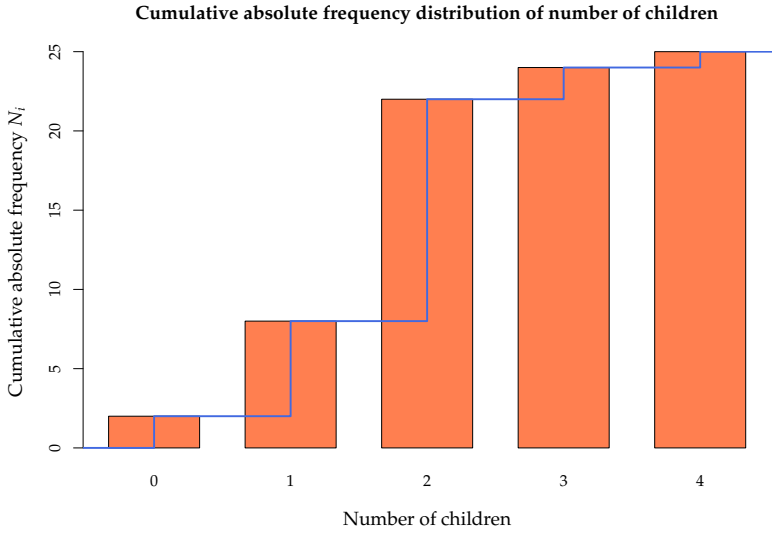
Cumulative absolute frequency bar chart

Non-grouped data



Cumulative absolute frequency line chart or polygon

Non-grouped data



Histogram

A **histogram** is similar to a bar chart but for grouped data.

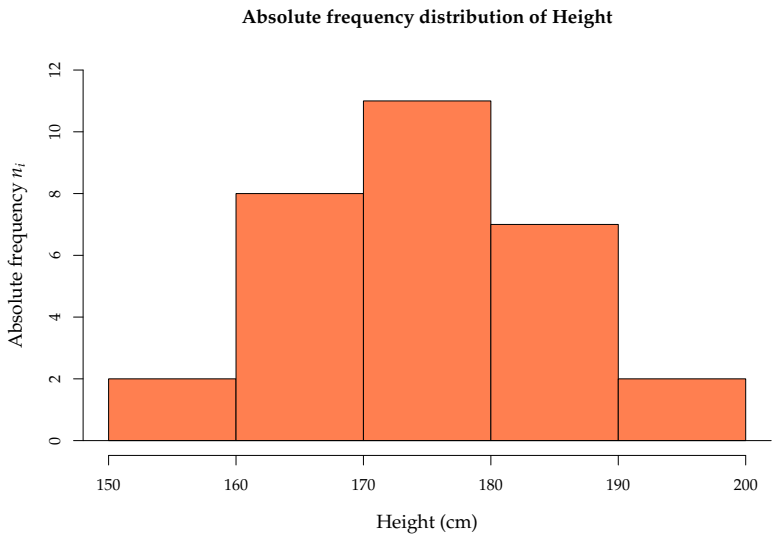
Usually the classes or grouping intervals are represented on the x-axis, and the frequencies on the y-axis. For each class, a bar is draw to the height of its frequency. Contrary to bar charts, the width of bars coincides with the width of classes, and there are no space between two consecutive bars.

Depending on the type of frequency represented in the y-axis we get different types of histograms.

Sometimes a polygon, known as **frequency polygon**, is plotted joining the top of every bar.

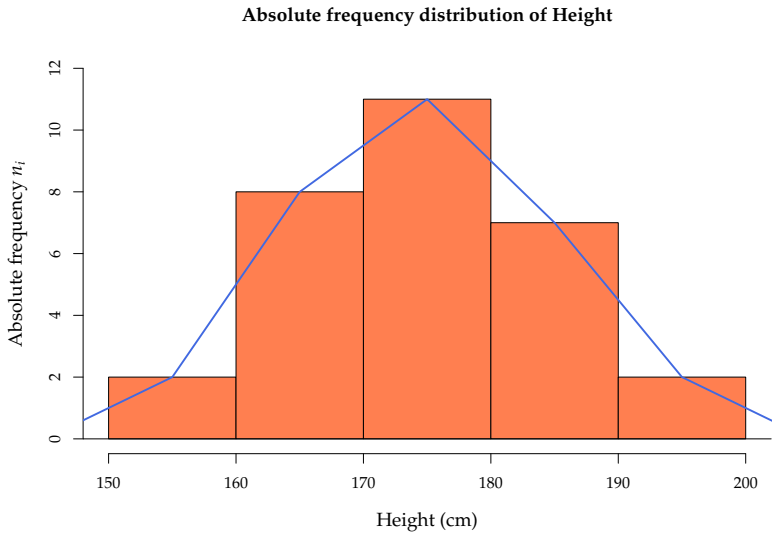
Absolute frequency histogram

Grouped data



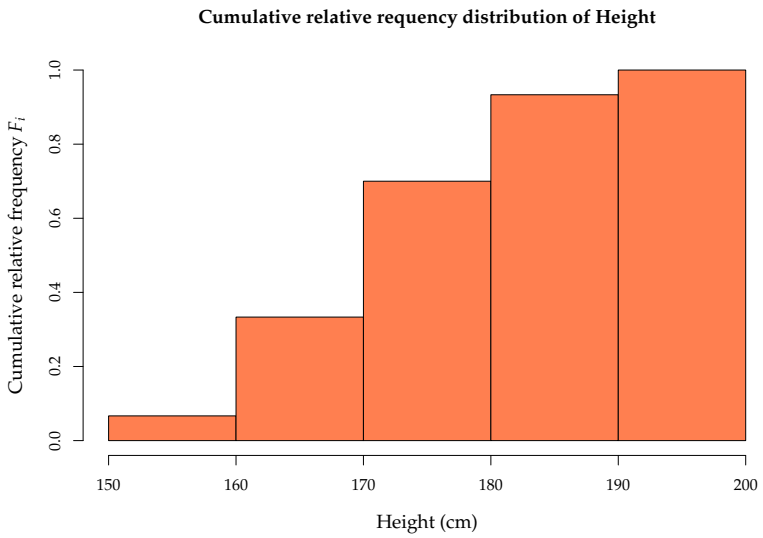
Absolute frequency histogram

Grouped data



Cumulative relative frequency histogram

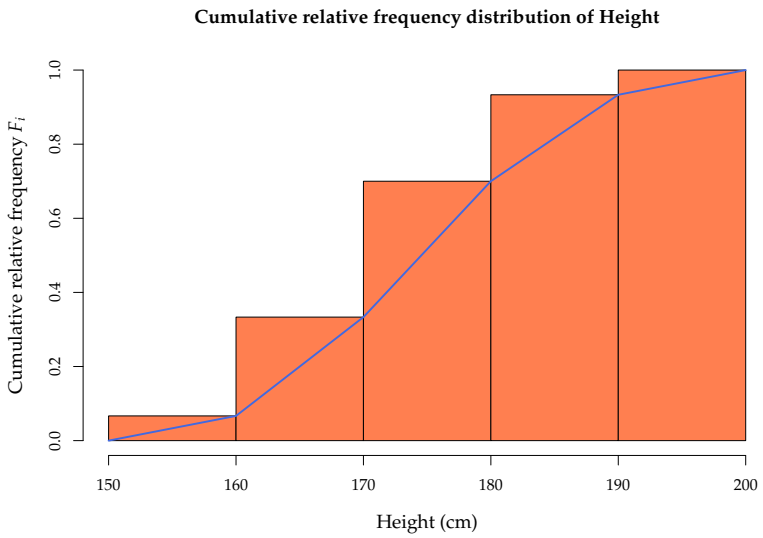
Grouped data



The cumulative frequency polygon (for absolute or relative frequencies) is known as **ogive**.

Cumulative relative frequency line chart or ogive

Grouped data



Observe that in the ogive we join the top right corner of bars with straight lines, instead of the top center, cause we don't reach the accumulated frequency of the class until the end of the interval.

Pie chart

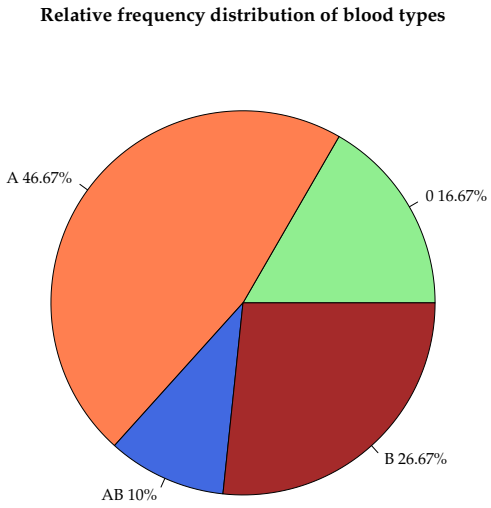
A **pie chart** consists in a circle divided in slices, one for every value or category of the variable. Each slice is called **sector** and its angle or area is proportional to the frequency of the corresponding value or category.

Pie charts can represent absolute or relative frequencies, but not cumulative frequencies, and are used with nominal qualitative variables. For ordinal qualitative or quantitative variables is better to use

bar charts or histograms, cause it's easier to perceive differences in one dimension (length of bars) than in two dimensions (areas of sectors).

Pie chart

Nominal variables



Outliers

One of the main problems in samples are **outliers**, that are values very different from the rest of values of the sample.

It's important to find out outliers before doing any analysis, cause outliers usually distort the results.

They always appears in the ends of the distribution, and can be find out easily with a box and whiskers chart (as be showed later).

Outliers management

With big samples outliers have less importance and can be left in the sample.

With small samples we have several options:

- Remove the outlier if it's an error.
- Replace the outlier by the lower or higher value in the distribution that is not an outlier if it's not an error and the outlier doesn't fit the theoretical distribution.
- Leave the outlier if it's not an error, and change the theoretical model to fit it to outliers.

3 Sample statistics

Sample statistics

The frequency table and charts summarize and give an overview of the distribution of values of the studied variable in the sample, but it's difficult to describe some aspects of the distribution from it, as for example, which are the most representative values of the distribution, how is the spread of data, which data could be considered outliers, how is the symmetry of the distribution.

To describe those aspects of the sample distribution more specific numerical measures, called **sample statistics**, are used.

According to the aspect of the distribution that they study, there are different types of statistics:

Measures of locations: They measure the values where data are concentrated or that divide the distribution into equal parts.

Measures of dispersion: They measure the spread of data.

Measures of shape: They measure the symmetry and kurtosis of the distribution.

3.1 Location statistics

Location statistics

There are two groups:

Central location measures: They measure the values where data are concentrated, and that usually are in the centre of the distribution. These values are the values that best represents the sample data. The most important are:

- Arithmetic mean
- Median
- Mode

Non-central location measures: They divide the sample data into equals parts. The most important are:

- Quartiles.
- Deciles.
- Percentiles.

Arithmetic mean

Definition 9 (Sample arithmetic mean $\bar{x}$)**.** The *sample arithmetic mean* of a variable X is the sum of observed values in the sample divided by the sample size

$$\bar{x} = \frac{\sum x_i}{n}$$

From the frequency table can be calculated with the formula

$$\bar{x} = \frac{\sum x_i n_i}{n} = \sum x_i f_i$$

In most cases the arithmetic mean is the value that best represent the observed values in the sample.

Watch out! It can not be calculated with qualitative variables.

Arithmetic mean calculation

Example with non-grouped data

Using the data of the sample with the number of children of families, the arithmetic mean is

$$\bar{x} = \frac{1+2+4+2+2+2+3+2+1+1+0+2+2}{25} + \\ + \frac{0+2+2+1+2+2+3+1+2+2+1+2}{25} = \frac{44}{25} = 1.76 \text{ children.}$$

or using the frequency table

x_i	n_i	f_i	$x_i n_i$	$x_i f_i$
0	2	0.08	0	0
1	6	0.24	6	0.24
2	14	0.56	28	1.12
3	2	0.08	6	0.24
4	1	0.04	4	0.16
$\sum$	25	1	44	1.76

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{44}{25} = 1.76 \text{ children} \qquad \bar{x} = \sum x_i f_i = 1.76 \text{ children.}$$

This is the number of children that best represent the families in the sample.

Arithmetic mean calculation

Example with grouped data

Using the data of the sample of student heights, the arithmetic mean is

$$\bar{x} = \frac{179 + 173 + \cdots + 187}{30} = 175.07 \text{ cm.}$$

or using the frequency table and taking the class marks as x_i,

X	x_i	n_i	f_i	$x_i n_i$	$x_i f_i$
$(150, 160]$	155	2	0.07	310	10.33
$(160, 170]$	165	8	0.27	1320	44.00
$(170, 180]$	175	11	0.36	1925	64.17
$(180, 190]$	185	7	0.23	1295	43.17
$(190, 200]$	195	2	0.07	390	13
$\sum$		30	1	5240	174.67

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{5240}{30} = 174.67 \text{ cm} \qquad \bar{x} = \sum x_i f_i = 174.67 \text{ cm.}$$

Observe that when the mean is calculated from the table the result differs a little from the real value, cause the values used in the calculations are the class marks instead of the actual values.

Weighted mean

In some cases the values of the sample have different importance. In that case the importance or *weight* of each value of the sample must be taken into account when calculating the mean.

Definition 10 (Sample weighted mean $\bar{x}_w$)**.** Given a sample of values $x_1, \ldots, x_n$ where every value x_i has a weight w_i, the *weighted mean* of variable X is the sum of the product of each value by its weight, divided by sum of weights

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i}$$

From the frequency table can be calculated with the formula

$$\bar{x}_w = \frac{\sum x_i w_i n_i}{\sum w_i}$$

Weighted mean calculation

Assume that a student wants to calculate a representative measure o its performance in a course. The grade and the credits of every subjects are

Subject	Credits	Grade
Maths	6	5
Economics	4	3
Chemistry	8	6

The arithmetic mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{5+3+6}{3} = 4.67 \text{ points},$$

However, this measure does not represent well the performance of the student, as not all the subjects have the same importance and require the same effort to pass. Subjects with more credits require more work and must have more weight in the calculation of the mean.

In this case is better to use the weighted mean, using the credits as the weights of grades, as a representative measure of the student effort

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i} = \frac{5 \cdot 6 + 3 \cdot 4 + 6 \cdot 8}{6+4+8} = \frac{90}{18} = 5 \text{ points}.$$

Median

Definition 11 (Sample median *Me*)**.** The *sample median* of a variable X is the value that is in the middle of the ordered sample.

The median divides the sample distribution into two equal parts, that is, there are the same number of values above and below the median. Therefore, it has cumulative frequencies $N_{Me} = n/2$ y $F_{Me} = 0.5$.

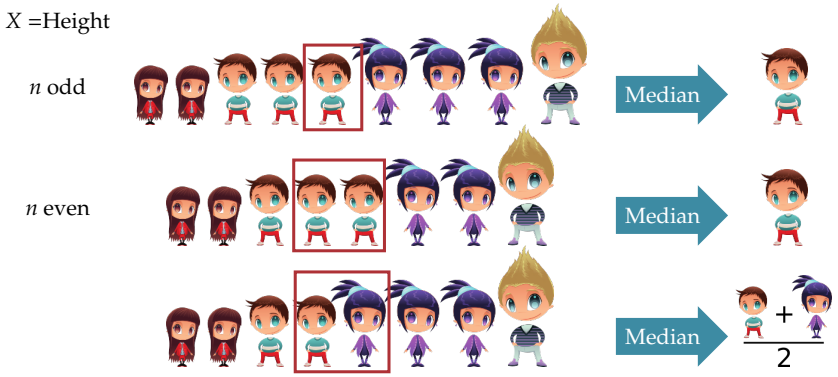
Watch out! It can not be calculated for nominal variables.

Median calculation

Non-grouped data

With non-grouped data, there are two possibilities:

- Odd sample size: The median is the value in the position $\frac{n+1}{2}$.
- Even sample size: The median is the average of values in positions $\frac{n}{2}$ and $\frac{n}{2}+1$.



Median calculation

Example with non-grouped data

Using the data of the sample with the number of children of families, the sample size is 25, that is odd, and the median is the value in the position $\frac{25+1}{2} = 13$ of the sorted sample.

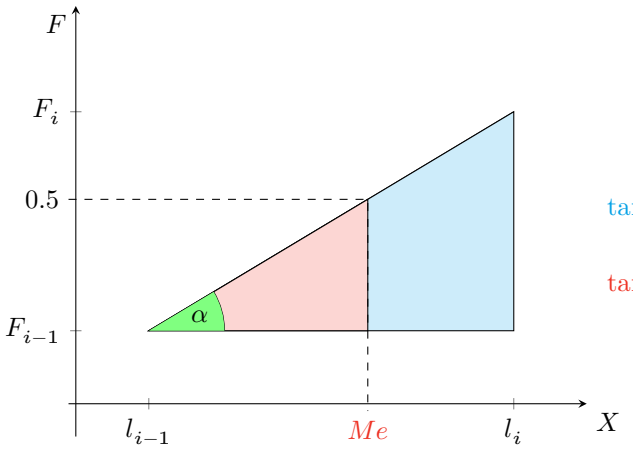
$$0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, \boxed{2}, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 4$$

and the median is 2 children.

With the frequency table, the median is the lowest value with a cumulative absolute frequency greater than or equal to 13, or with a cumulative relative frequency greater than or equal to 0.5.

x_i	n_i	f_i	N_i	F_i
0	2	0.08	2	0.08
1	6	0.24	8	0.32
2	14	0.56	22	0.88
3	2	0.08	24	0.96
4	1	0.04	25	1
$\sum$	25	1		

Median calculation for grouped data



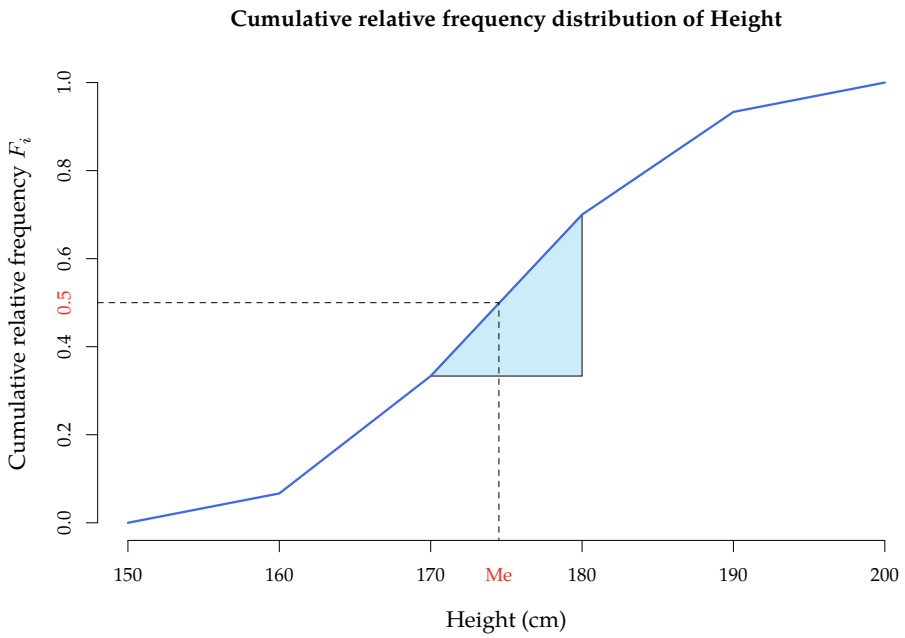
$$\tan(\alpha) = \frac{F_i - F_{i-1}}{l_i - l_{i-1}}$$

$$\tan(\alpha) = \frac{0.5 - F_{i-1}}{Me - l_{i-1}}$$

$$Me = l_i + \frac{0.5 - F_{i-1}}{F_i - F_{i-1}}(l_i - l_{i-1}) = l_i + \frac{0.5 - F_{i-1}}{f_i}a_i$$

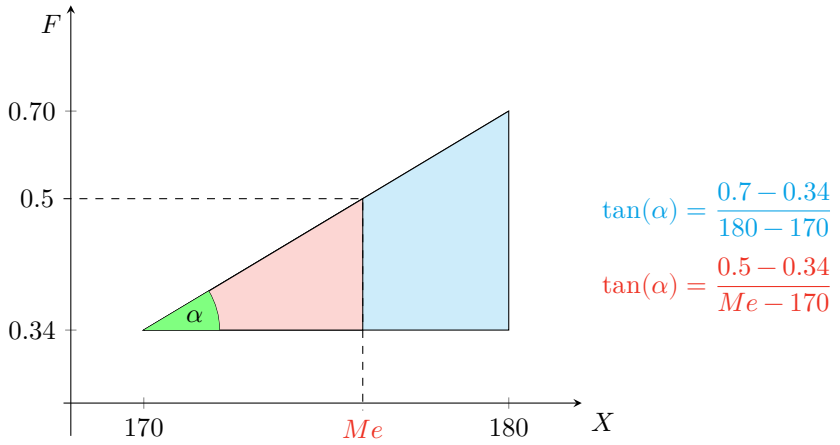
Median calculation for grouped data

Example



Median calculation for grouped data

Example



$$Me = 170 + \frac{0.5 - 0.34}{0.7 - 0.34}(180 - 170) = 170 + \frac{0.16}{0.36}10 = 174.54 \text{ cm}$$

Mode

Definition 12 (Sample Mode *Mo*)**.** The *sample mode* of a variable X is the most frequent value in the sample.

With grouped data the *modal class* is the class with the highest frequency.

It can be calculated for all types of variables (qualitative and quantitative).

Some distributions can have more than one mode

Mode calculation

Using the data of the sample with the number of children of families, the value with the highest frequency is 2, that is the mode $Mo = 2$ children.

x_i	n_i
0	2
1	6
2	14
3	2
4	1

Using the data of the sample of student heights, the class with the highest frequency is $(170, 180]$ that is the modal class $Mo = (170, 180]$.

X	n_i
$(150, 160]$	2
$(160, 170]$	8
$(170, 180]$	11
$(180, 190]$	7
$(190, 200]$	2

Which central tendency statistic should I use?

In general, when all the central tendency statistics can be calculated, is advisable to use them as representative values in the following order:

1. Mean. Mean takes more information from the sample than the others, as it takes into account the magnitude of data.
2. Median. Median takes less information than mean but more than mode, as it takes into account the order of data.
3. Mode. Mode is the measure that fewer information takes from the sample, as it only takes into account the absolute frequency of values.

But, *be careful with outliers*, as the mean can be distorted by them. In that case is better to use the median as the value most representative.

For example, if a sample of number of children of 7 families is

$$0, 0, 1, 1, 2, 2, 15$$

$$\bar{x} = 3 \text{ children} \quad \text{and} \quad Me = 1 \text{ children}$$

Which measure represent better the number of children in the sample?

Non-central location measures

The non-central location measures or *quantiles* divide the sample distribution in equal parts.

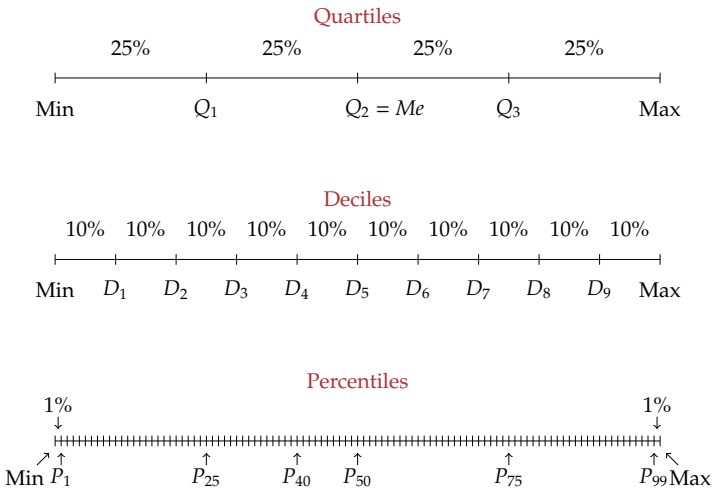
The most used are:

Quartiles: Divide the distribution into 4 equal parts. There are 3 quartiles: Q_1 (25% accumulated) , Q_2 (50% accumulated), Q_3 (75% accumulated).

Deciles: Divide the distribution into 10 equal parts. There are 9 deciles: D_1 (10% accumulated) ,..., D_9 (90% accumulated).

Percentiles: Divide the distribution into en 100 equal parts. There are 99 percentiles: P_1 (1% accumulated),..., P_{99} (99% accumulated).

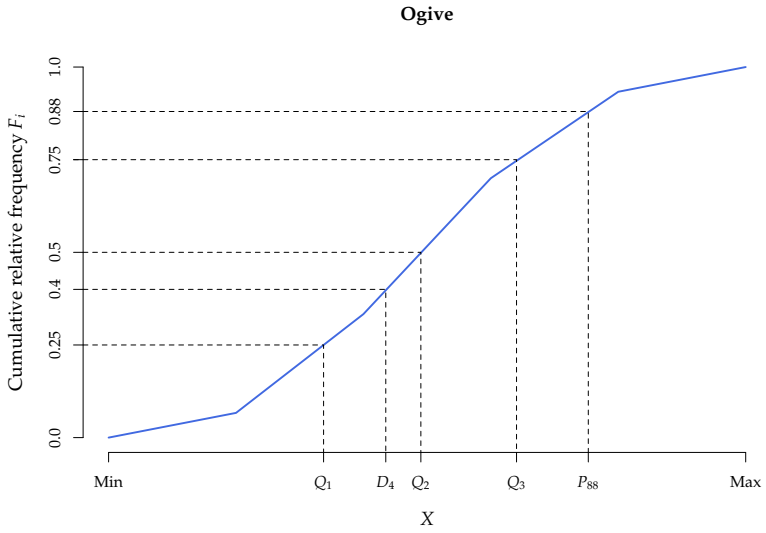
Quantiles



Observe that there is a correspondence between quartiles, deciles and percentiles. For example, first quartile coincide with percentile 25, and fourth decile coincides with the percentile 40.

Quantiles calculation

Quantiles are calculated in a similar way to the median. The only difference lies in the cumulative relative frequency that correspond to every quantile.



Quantile calculation

Example with non-grouped data

Using the data of the sample with the number of children of families, the cumulative relative frequencies were

x_i	F_i
0	0.08
1	0.32
2	0.88
3	0.96
4	1

$$
\begin{aligned}
F_{Q_1} = 0.25 &\Rightarrow C_1 = 1 \text{ children},\\
F_{Q_2} = 0.5 &\Rightarrow C_2 = 2 \text{ children},\\
F_{Q_3} = 0.75 &\Rightarrow C_3 = 2 \text{ children},\\
F_{D_4} = 0.4 &\Rightarrow D_3 = 2 \text{ children},\\
F_{P_{92}} = 0.92 &\Rightarrow P_{92} = 3 \text{ children}.
\end{aligned}
$$

3.2 Dispersion statistics

Dispersion statistics

Dispersion or *spread* refers to the variability of data. So, dispersion statistics measure how the data values are scattered in general, or with respect to a central location measure.

For quantitative variables, the most important are:

- Range
- Interquartile range

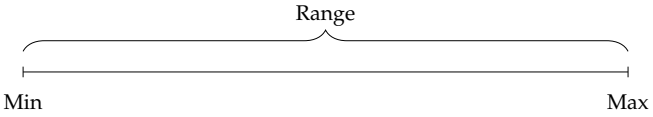
- Variance
- Standard deviation
- Coefficient of variation

Range and interquartile range

Definition 13 (Sample range)**.** The *sample range* of a variable X is the difference between the the maximum and the minimum value in the sample.

$$\text{Range} = \max_{x_i} - \min_{x_i}$$

The range measure the largest variation among the sample data. However, it's very sensitive to outliers, as they appear at the ends of the distribution, and for that reason is rarely used.

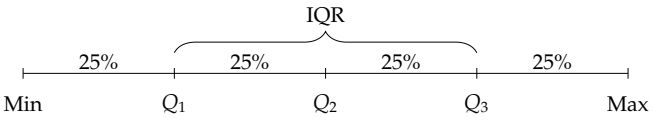


Range and interquartile range

The following measure avoid the problem of outliers and is much more used.

Definition 14 (Sample interquartile range)**.** The *sample interquartile range* of a variable X is the difference between the third and the first sample quartiles.

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



The interquartile range measures the spread of the 50% central data.

Box plot

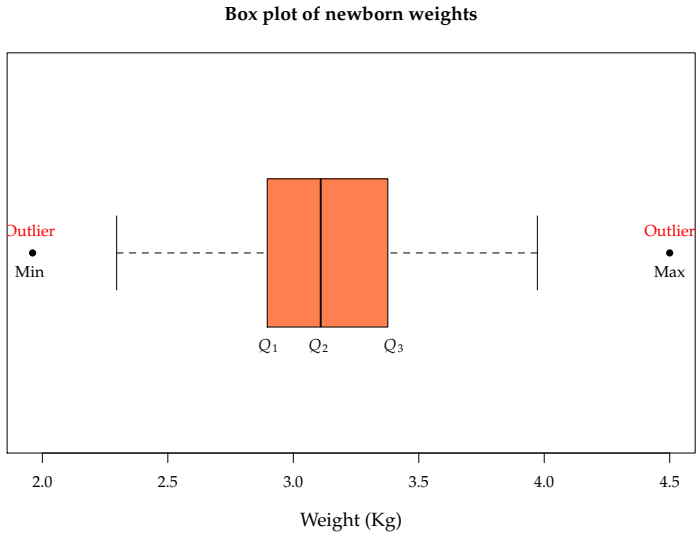
The dispersion of a variable in a sample can be graphically represented with a **box plot**, that represent five descriptive statistics (minimum, quartiles and maximum) known as the *five-numbers*. It consist in a box, drawn from the lower to the upper quartile, that represent the interquartile range, and two segments, known as the lower and the upper *whiskers*. Usually the box is split in two with the median.

This chart is very helpful as it serves to many purposes:

- It serves to measure the spread of data as it represent the range and the interquartile range.
- It serves to detect outliers, that are the values outside the interval defined by the whiskers.
- It serves to measure the symmetry of distribution, comparing the length of the boxes and whiskers above and below the median.

Box plot

Example with newborn weights



Box plot construction

To create a box plot follow the steps below

1. Calculate the quartiles.
2. Draw a box from the lower to the upper quartile.
3. Split the box with the median or second quartile.
4. For the whiskers calculate first two values called *fences*

$$f_1 = Q_1 - 1.5\,\text{IQR}$$
$$f_2 = Q_3 + 1.5\,\text{IQR}$$

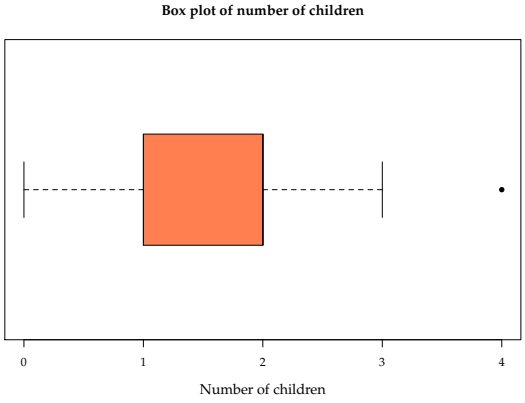
   The fences define the interval where data are considered normal. Any value outside that interval is considered an outlier. For the lower whisker draw a segment from the lower quartile to the lower value in the sample grater than or equal to f_1, and for the upper whisker draw a segment from the upper quartile to the highest value in the sample lower than or equal to f_2.

5. Finally, if there are some outlier, draw a dot in every outlier.

Box plot construction

Example of number of children

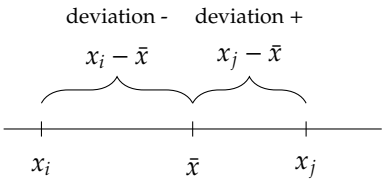
1. Calculate the quartiles: $Q_1 = 1$ children and $Q_2 = Q_3 = 2$ children
2. Draw the box.
3. Calculate the fences $f_1 = 1 - 1.5 * 1 = -0.5$ and $f_2 = 2 + 1.5 * 1 = 3.5$.
4. Draw the whiskers: $w_1 = 0$ children and $w_2 = 3$ children.
5. Draw the outliers: 4 children.



Deviations from the mean

Another way of measuring spread of data is with respect to a central tendency measure, as for example the mean.

In that case, it's measured the distance from every value in the sample to the mean, that is called **deviation from the mean.**



If deviations are big, the mean is less representative than when they are small.

Variance and standard deviation

Definition 15 (Sample variance s^2)**.** The *sample variance* of a variable X is the average of squared deviations from the mean.

$$s^2 = \frac{\sum (x_i - \bar{x})^2 n_i}{n} = \sum (x_i - \bar{x})^2 f_i$$

It can also be calculated with the formula

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \sum x_i^2 f_i - \bar{x}^2$$

The variance has the units of the variable squared, and to ease their interpretation it's common to calculate its square root.

Definition 16 (Sample standard deviation s)**.** The *sample standard deviation* of a variable X is the square root of the variance.

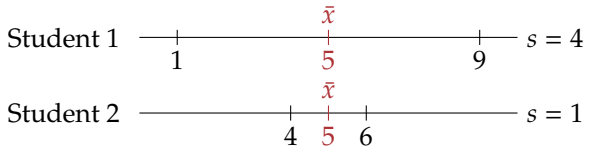
$$s = +\sqrt{s^2}$$

Variance and standard deviation interpretation

Both variance and standard deviation measures the spread of data around the mean. When the variance or the standard deviation are small, the sample data are concentrated around the mean, and the mean is a good representative measure. In contrast, when variance or the standard deviation are high, the sample data are far from the mean, and the mean doesn't represent so good.

Standard deviation small $\Rightarrow$ *Mean is representative*
Standard deviation big $\Rightarrow$ *Mean is unrepresentative*

Example The following samples contains the grades of 2 students in 2 subjects



Which mean is more representative?

Variance and standard deviation calculation

Example with non-grouped data

Using the data of the sample with the number of children of families, and adding a new column to the frequency table with the squared values,

x_i	f_i	$x_i^2 f_i$
0	0.08	0.00
1	0.24	0.24
2	0.56	2.24
3	0.08	0.72
4	0.04	0.64
$\sum$		3.84

$$s^2 = \sum x_i^2 f_i - \bar{x}^2 = 3.84 - 1.76^2 = 0.7424 \text{ children}^2,$$

and the standard deviation is $s = \sqrt{0.7424} = 0.8616$ children.

Compared to the range, that is 4 children, the standard deviation is not very large, so we can conclude that the dispersion of the distribution is small and consequently the mean, $\bar{x} = 1.76$ children, represents quite well the number of children of families of the sample.

Variance and standard deviation calculation

Example with grouped data

Using the data of the sample with the heights of students and grouping heights in classes, the calculation is the same but using the class marks.

X	x_i	n_i	$x_i^2 n_i$
$(150, 160]$	155	2	48050
$(160, 170]$	165	8	217800
$(170, 180]$	175	11	336875
$(180, 190]$	185	7	239575
$(190, 200]$	195	2	76050
$\sum$		30	918350

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \frac{918350}{30} - 174.67^2 = 102.06 \text{ cm}^2,$$

and the standard deviation is $s = \sqrt{102.06} = 10.1$ cm.

This value is quite small compared to the range of the variable, that goes from 150 to 200 cm, therefore the distribution of heights has little dispersion and the mean is very representative.

Coefficient of variation

Both, variance and standard deviation, have units and that makes difficult to interpret them, specially when comparing distributions of variables with different units.

For that reason it's also common to use the following dispersion measure that has no units.

Definition 17 (Sample coefficient of variation cv)**.** The *sample coefficient of variation* of a variable X is the quotient between the sample standard deviation and se absolute value of the sample mean.

$$cv = \frac{s}{|\bar{x}|}$$

The coefficient of variation measures the relative dispersion of data around the sample mean.

As it has no units, it's easier to interpret: The higher the it is the higher the relative dispersion with respect to the mean and less representative is the mean.

The coefficient of variation it's very helpful to compare dispersion in distributions of different variables, even if variables have different units.

Watch out! It makes no sense when the mean is 0 or close to 0.

Coefficient of variation

Example

In the sample of the number of children, where the mean was $\bar{x} = 1.76$ and the standard deviation was $s = 0.8616$ children, the coefficient of variation is

$$cv = \frac{s}{|\bar{x}|} = \frac{0.8616}{|1.76|} = 0.49.$$

In the sample of heights, where the mean was $\bar{x} = 174.67$ cm and the standard deviation was $s = 10.1$ cm, the coefficient of variation is

$$cv = \frac{s}{|\bar{x}|} = \frac{10.1}{|174.67|} = 0.06.$$

This means that the relative dispersion in the heights distribution is lower than in the number of children distribution, and consequently the mean of height is most representative than the mean of number of children.

3.3 Shape statistics

Shape statistics

They are measures that describe the shape of the distribution.

In particular, the most important aspects are:

Symmetry: It measures the symmetry of the distribution with respect to the mean. The statistics most used is the *coefficient of skewness*.

Kurtosis: It measures the length of tails or the peakness of distribution. The statistics most used is the *coefficient of kurtosis*.

Coefficient of skewness

Definition 18 (Sample coefficient of skewness g_1)**.** The *sample coefficient of skewness* of a variable X is the average of the deviations of values from the sample mean to cube, divided by the standard deviation to cube.

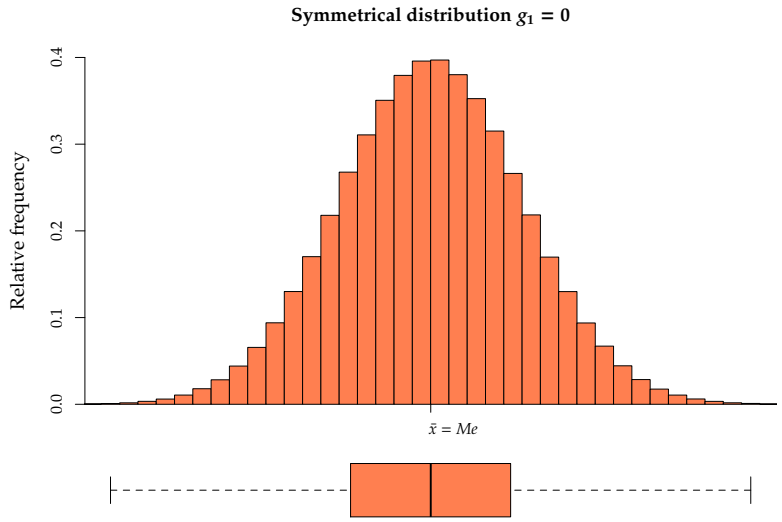
$$g_1 = \frac{\sum (x_i - \bar{x})^3 n_i/n}{s^3} = \frac{\sum (x_i - \bar{x})^3 f_i}{s^3}$$

It measures the symmetry or skewness of the distribution, that is, how many values in the sample are above or below the mean and how far from it.

- $g_1 = 0$ indicates that there are the same number of values in the sample above and below the mean and equally deviated from it, and the distribution is symmetrical.
- $g_1 < 0$ indicates that there are more values above the mean than below it, but the values below are further from it, and the distribution is right-skewed (it has longer tail to the right).
- $g_1 > 0$ indicates that there are more values below the mean than above it, but the values above are further from it, and the distribution is left-skewed (it has longer tail to the left).

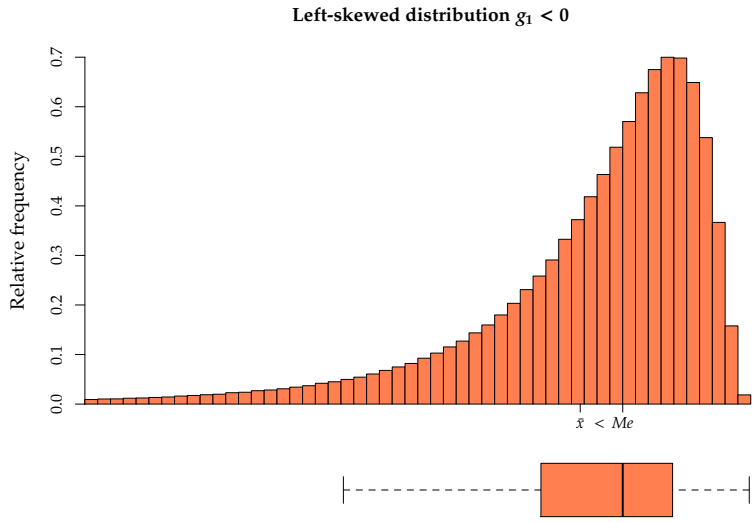
Coefficient of skewness

Example of symmetrical distribution



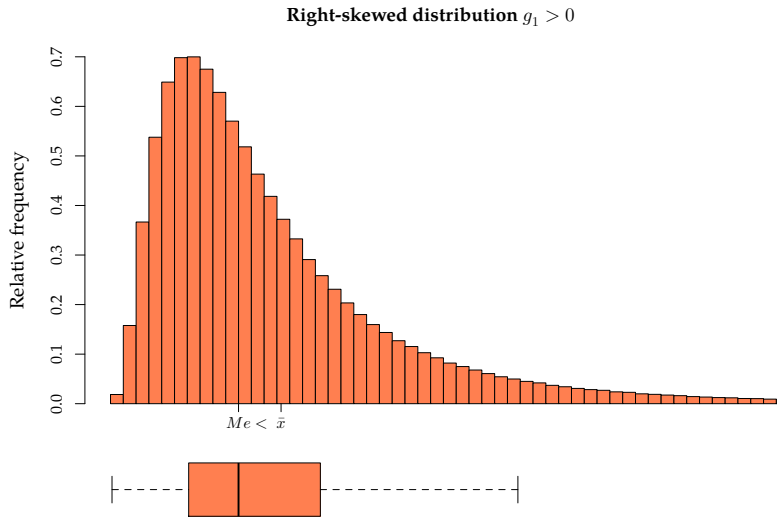
Coefficient of skewness

Example of left-skewed distribution



Coefficient of skewness

Example of right-skewed distribution



Coefficient of skewness calculation

Example with grouped data

Using the frequency table of the sample with the heights of students and adding a new column with the deviations to the mean $\bar{x} = 174.67$ cm to cube, we get

X	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^3 n_i$
$(150, 160]$	155	2	-19.67	-15221.00
$(160, 170]$	165	8	-9.67	-7233.85
$(170, 180]$	175	11	0.33	0.40
$(180, 190]$	185	7	10.33	7716.12
$(190, 200]$	195	2	20.33	16805.14
$\sum$		30		2066.81

$$g_1 = \frac{\sum (x_i - \bar{x})^3 n_i / n}{s^3} = \frac{2066.81/30}{10.1^3} = 0.07.$$

As it is close to 0, that means that the distribution of heights is fairly symmetrical.

Coefficient of kurtosis

Definition 19 (Sample coefficient of kurtosis g_2)**.** The *sample coefficient of kurtosis* of a variable X is the average of the deviations of values from the sample mean to the fourth power, divided by the standard deviation to the fourth power and minus 3.

$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i / n}{s^4} - 3 = \frac{\sum (x_i - \bar{x})^4 f_i}{s^4} - 3$$

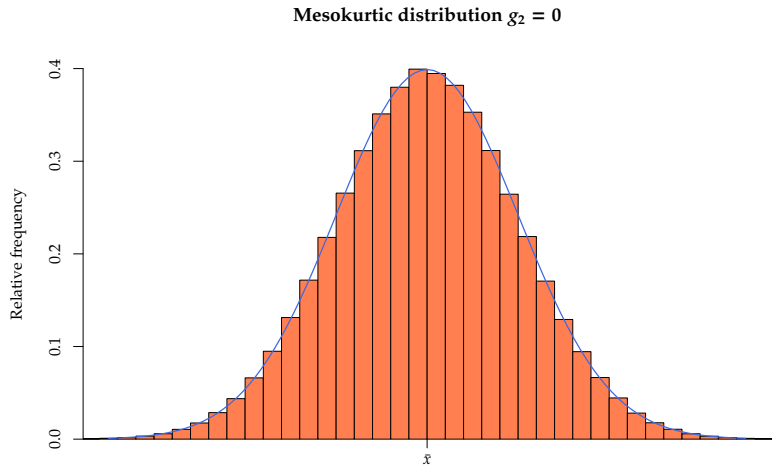
The coefficient of kurtosis measures the the length of tails or the peakness of distribution with respect to normal (bell-shaped) distribution of reference.

- $g_2 = 0$ indicates that the distribution has the same tails and peakedness than a normal distribution (*mesokurtic*).

- $g_2 < 0$ indicates that the distribution has longer tails and lower peakedness than a normal distribution (*platykurtic*).
- $g_2 > 0$ indicates that the distribution has shorter tails and higher peakedness than a normal distribution (*leptokurtic*).

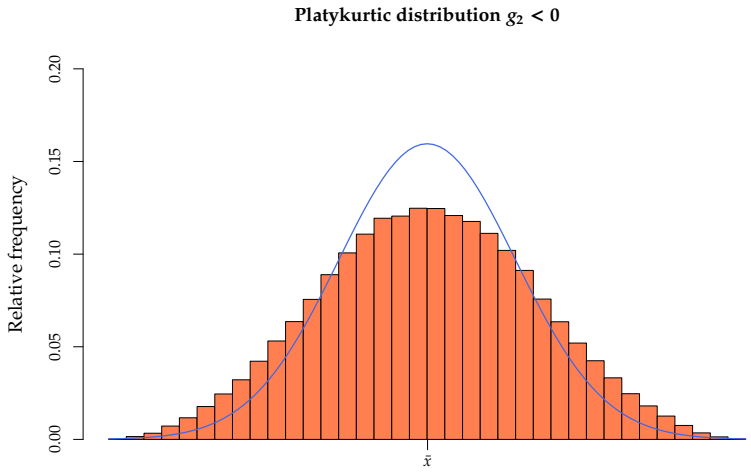
Coefficient of kurtosis

Example of mesokurtic distribution



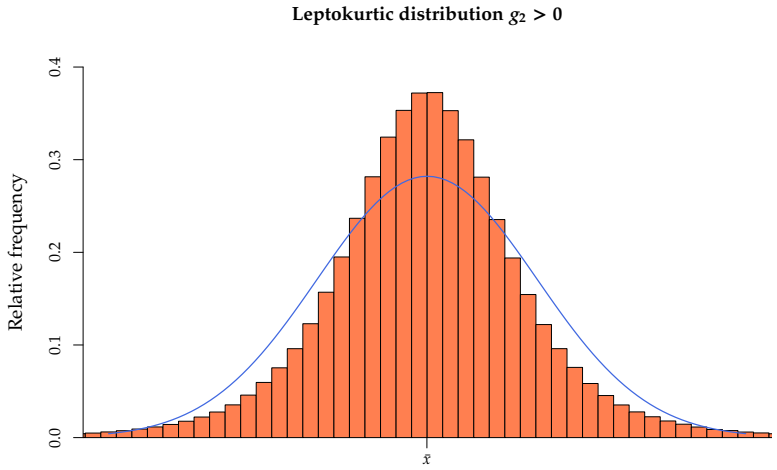
Coefficient of kurtosis

Example of platykurtic distribution



Coefficient of kurtosis

Example of leptokurtic distribution



Coefficient of kurtosis

Example with grouped data

Using the frequency table of the sample with the heights of students and adding a new column with the deviations to the mean $\bar{x} = 174.67$ cm to the fourth power, we get

X	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^4 n_i$
$(150, 160]$	155	2	-19.67	299396.99
$(160, 170]$	165	8	-9.67	69951.31
$(170, 180]$	175	11	0.33	0.13
$(180, 190]$	185	7	10.33	79707.53
$(190, 200]$	195	2	20.33	341648.49
$\sum$		30		790704.45

$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i / n}{s^4} - 3 = \frac{790704.45/30}{10.1^4} - 3 = -0.47.$$

As it is a negative value but not too far from 0, that means that the distribution of heights is a little bit platykurtic.

Interpretation

As we will see in the chapters of inferential statistics, many of the statistical test can only be applied to normal (bell-shaped) populations.

Normal distributions are symmetrical and mesokurtic, and therefore, they have both the coefficients of symmetry and kurtosis 0. So, a way of checking if a sample comes from a normal population is looking how far are the coefficients of skewness and kurtosis from 0.

In general, the normality of population is rejected when g_1 or g_2 are outside the interval $[-2, 2]$.

In that case, is common to apply a transformation to the variable to correct non-normality.

3.4 Variable transformations

Variable transformations

In many cases, the raw sample data are transformed to correct non-normality of distribution or just to get a more appropriate scale.

For example, if we are working with heights in metres and a sample contains the following values:

$$1.75 \text{ m}, 1.65 \text{ m}, 1.80 \text{ m},$$

it's possible to avoid decimals multiplying by 100, that is, changing from metres to centimetres:

$$175 \text{ cm}, 165 \text{ cm}, 180 \text{ cm},$$

And it's also possible to reduce the magnitude of data subtracting the minimum value in the sample, in this case 165 cm:

$$10 \text{ cm}, 0 \text{ cm}, 15 \text{ cm},$$

It's obvious that these data are easier to work with than the original ones. In essences, what it's been done is to apply the following transformation o data:

$$Y = 100X - 165$$

Linear transformations

One of the most common transformations is a *linear transformation*:

$$Y = a + bX.$$

For a linear transformation the mean and the standard deviation of the transformed variable are

$$\bar{y} = a + b\bar{x},$$
$$s_y = |b|s_x$$

Additionally, the coefficient of kurtosis doesn't change and the coefficient of skewness changes only the sign if b is negative.

Standardization and standard scores

One of the most common linear transformations is the *standardization*.

Definition 20 (Standardized variable and standard scores)**.** The *standardized variable* of a variable X is the variable that result of subtracting the mean from X and dividing by the standard deviation

$$Z = \frac{X - \bar{x}}{s_x}.$$

For each value x_i of the sample, the *standard score* is the value that results of applying the standardization transformation

$$z_i = \frac{x_i - \bar{x}}{s_x}.$$

The standard score is the number of standard deviations a value is above or below the mean, and it's useful to avoid the dependency of the variable from its measurement units.

The standardized variable always have mean 0 and standard deviation 1.

$$\bar{z} = 0 \qquad s_z = 1$$

Standardization and standard scores

Example

The grades of 5 students in 2 subjects are

Student:	1	2	3	4	5		
X :	2	5	4	8	6	$\bar{x} = 5$	$s_x = 2$
Y :	1	9	8	5	2	$\bar{y} = 5$	$s_y = 3.16$

Did the fourth student get the same performance in subject X than the third student in subject Y?

It might seem that both students had the same performance in every subject because they have the same degree, but in order to get the performance of every student relative to the group of students, the dispersion of grades in every subject must be considered. For that reason is better to use the standard score as a measure of relative performance.

X :	-1.5	0	-0.5	1.5	0.5
Y :	-1.26	1.26	0.95	0	-0.95

That is, the student with an 8 in X is 1.5 times the standard deviation below the mean of X, while the student with an 8 in Y is only 0.95 times the standard deviation below the mean of Y. Therefore, the first student had a higher performance in X than the second in Y.

Standardization and standard scores

Example

Following with the previous example and considering both subjects,

which is the best student?

If we only consider the sum of grades

Student:	1	2	3	4	5
X :	2	5	4	8	6
Y :	1	9	8	5	2
$\sum$	3	14	12	13	8

the best student is the second one.

But if the relative performance is considered, taking the standard scores

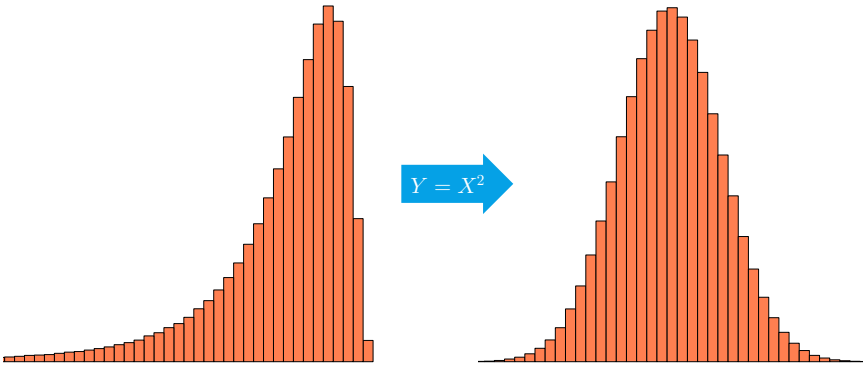
Student:	1	2	3	4	5
X :	-1.5	0	-0.5	1.5	0.5
Y :	-1.26	1.26	0.95	0	-0.95
$\sum$	-2.76	1.26	0.45	1.5	-0.45

the best student is the fourth one.

Non-linear transformations

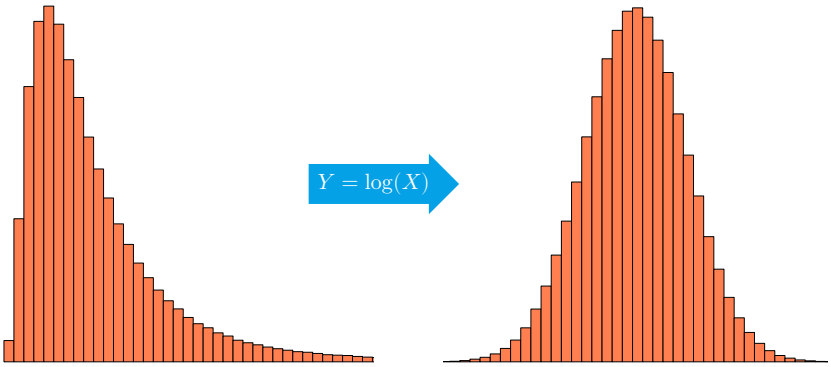
Non-linear transformations are also common to correct non-normality of distributions.

The square transformation $Y = X^2$ compresses small values and expand large values. So, it's used to correct left-skewed distributions.



Non-linear transformation

The square root transformation $Y = \sqrt{x}$, the logarithmic tranformation $Y = \log X$ and the inverse transformation $Y = 1/X$ compresses large values and expand small values. So, they are used to correct right-skewed distributions.



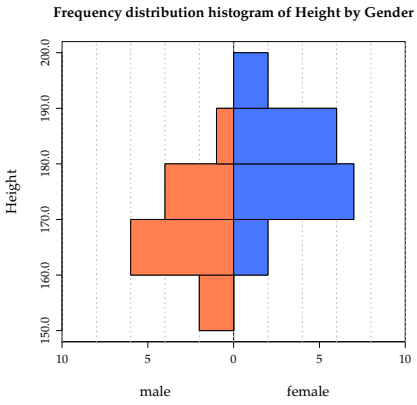
Factors

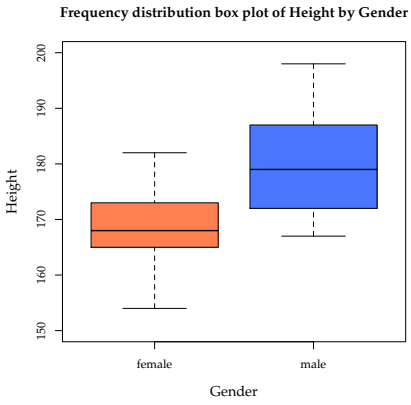
Sometimes is interesting to describe the frequency distribution of the main variable for different subsamples corresponding to the categories of another variable that is known as **classificatory variable** or **factor**.

Example Dividing the sample of heights by gender we get two subsamples

Females	173, 158, 174, 166, 162, 177, 165, 154, 166, 182, 169, 172, 170, 168.
Males	179, 181, 172, 194, 185, 187, 198, 178, 188, 171, 175, 167, 186, 172, 176, 187.

Comparing distributions for the levels of a factor


Frequency distribution histogram of Height by Gender
Height
150.0
160.0
170.0
180.0
190.0
200.0
10
5
0
5
10
male
female


Frequency distribution box plot of Height by Gender
Height
150
160
170
180
190
200
female
male
Gender

4 Regression and correlation

Relations among variables

In the last chapter we saw how to describe the distribution of a single variable in a sample. However, in most cases, studies require to describe several variables that are often related. For instance, a nutritional study should consider all the variables that could be related to the weight, as height, age, gender, smoking, diet, physic exercise, etc.

To understand a phenomenon that involve several variables is not enough to study every variable by its own. We have to study all the variables together to describe how they interact and the type of relation among them.

Usually in a *dependency study* there is a **dependent variable** Y that it is supposed to be influenced by a set of variables $X_1, \dots, X_n$ known as **independent variables**. The simpler case is a *simple dependency study* when there is only one independent variable, that is the case covered in this chapter.

4.1 Joint frequency distribution

Joint frequencies

To study the relation between two variables X and Y, we have to study the joint distribution of the **two-dimensional variable** (X, Y), whose values are pairs (x_i, y_j) where the first element is a value of X and the second a value of Y.

Definition 21 (Joint sample frequencies)**.** Given a sample of n values and a two-dimensional variable (X, Y), for every value of the variable (x_i, y_j) is defined

- **Absolute frequency** n_{ij}: Is the number of times that the pair (x_i, y_j) appears in the sample.
- **Relative frequency** f_{ij}: Is the proportion of times that the pair (x_i, y_j) appears in the sample.

$$f_{ij} = \frac{n_{ij}}{n}$$

Watch out! For two-dimensional variables it make no sense cumulative frequencies.

Joint frequency distribution

The values of the two-dimensional variable with their frequencies is known as **joint frequency distribution**, and is represented in a **joint frequency table**.

$X \backslash Y$	y_1	$\cdots$	y_j	$\cdots$	y_q
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}

Joint frequency distribution

Example with grouped data

The height (in cm) and weight (in kg) of a sample of 30 students is:

(179,85), (173,65), (181,71), (170,65), (158,51), (174,66), (172,62), (166,60), (194,90), (185,75), (162,55), (187,78), (198,109), (177,61), (178,70), (165,58), (154,50), (183,93), (166,51), (171,65), (175,70), (182,60), (167,59), (169,62), (172,70), (186,71), (172,54), (176,68),(168,67), (187,80).

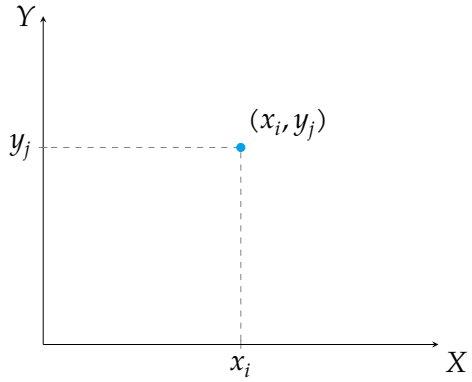
The joint frequency table is

X/Y	$[50,60)$	$[60,70)$	$[70,80)$	$[80,90)$	$[90,100)$	$[100,110)$
$(150,160]$	2	0	0	0	0	0
$(160,170]$	4	4	0	0	0	0
$(170,180]$	1	6	3	1	0	0
$(180,190]$	0	1	4	1	1	0
$(190,200]$	0	0	0	0	1	1

Scatter plot

The joint frequency distribution can be represented graphically with a **scatter plot**, where data is displayed as a collections of points on a XY coordinate system.

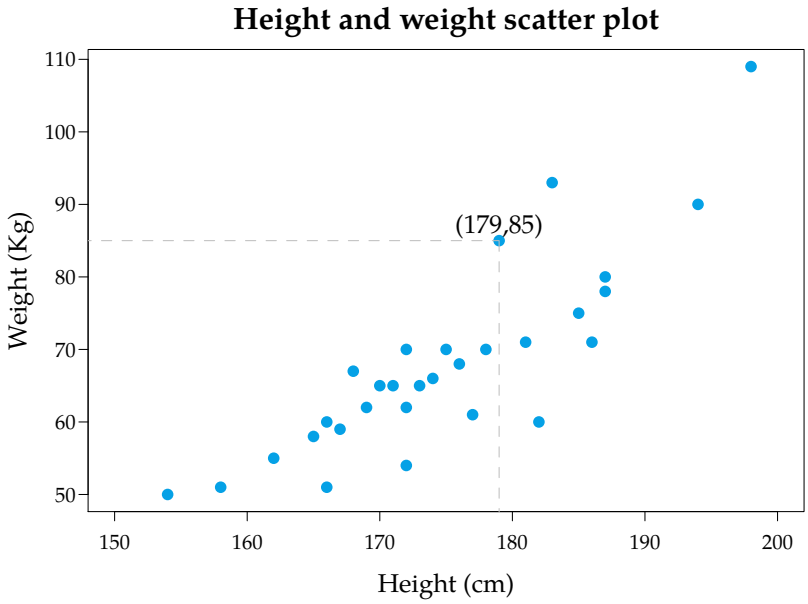
Usually the independent variable is represented in the X axis and the dependent variable in the Y axis. For every data pair (x_i, y_j) in the sample a dot is drawn on the plane with those coordinates.



The result is a set of points that usually is known as a *point cloud*.

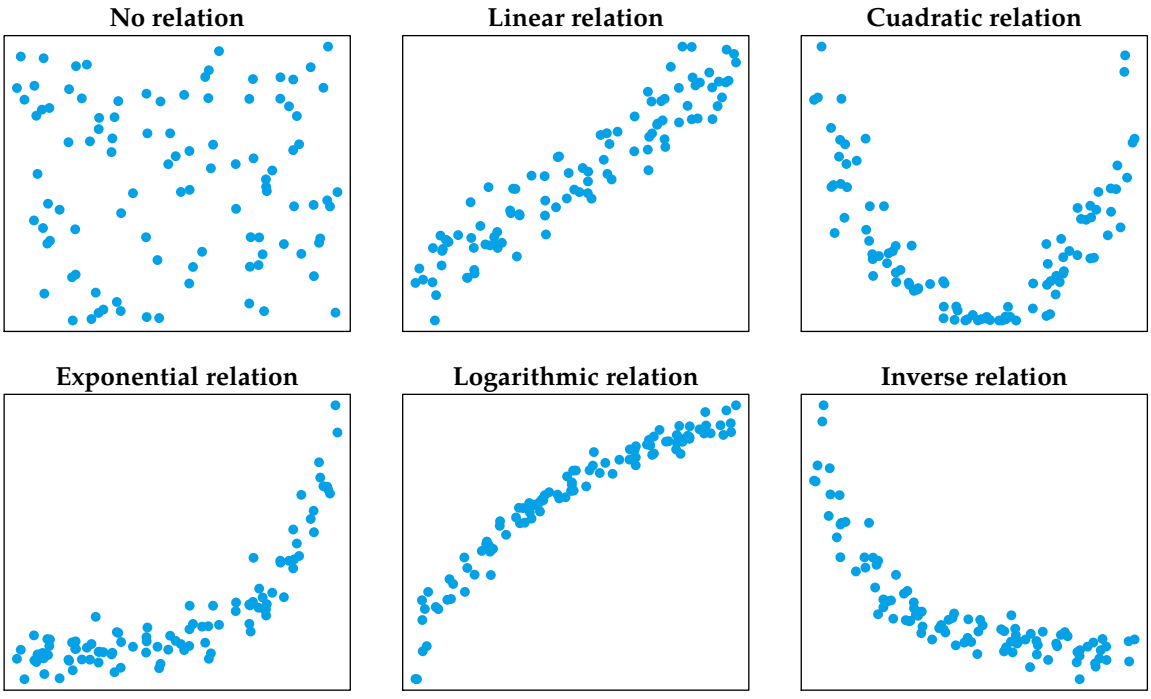
Scatter plot

Example of heights and weights



Scatter plot interpretation

The shape of the point cloud in a scatter plot gives information about the type of relation between the variables.

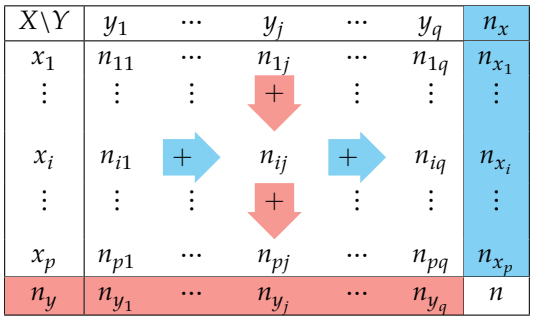


Marginal frequency distributions

The frequency distributions of each variable of the two-dimensional variable are known as **marginal frequency distributions**.

We can get the marginal frequency distributions from the joint frequency table summing frequencies

by rows and columns.



Marginal frequency distributions

Example of heights and weights

The marginal frequency distributions for the previous sample of heights and weights are

X/Y	$[50,60)$	$[60,70)$	$[70,80)$	$[80,90)$	$[90,100)$	$[100,110)$	n_x
$(150,160]$	2	0	0	0	0	0	2
$(160,170]$	4	4	0	0	0	0	8
$(170,180]$	1	6	3	1	0	0	11
$(180,190]$	0	1	4	1	1	0	7
$(190,200]$	0	0	0	0	1	1	2
n_y	7	11	7	2	2	1	30

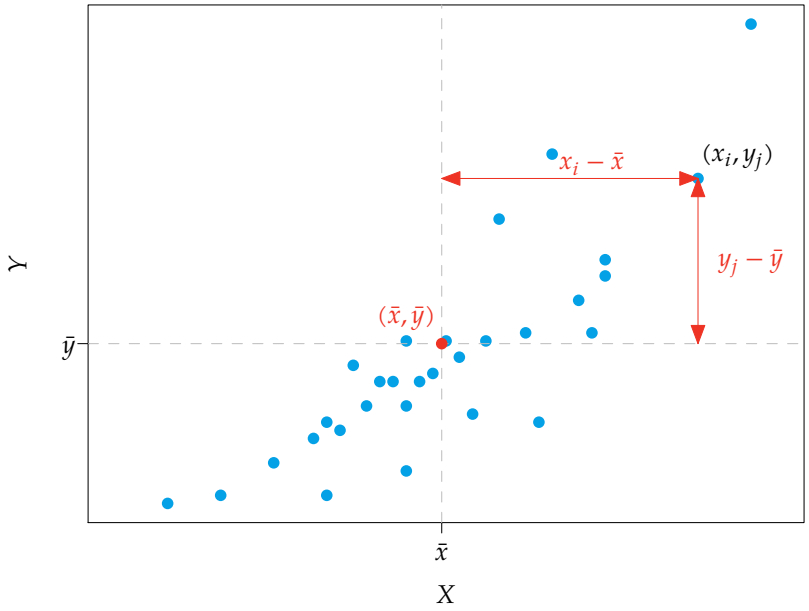
and the corresponding statistics are

$$\bar{x} = 174.67 \text{ cm} \qquad s_x^2 = 102.06 \text{ cm}^2 \qquad s_x = 10.1 \text{ cm}$$
$$\bar{y} = 69.67 \text{ Kg} \qquad s_y^2 = 164.42 \text{ Kg}^2 \qquad s_y = 12.82 \text{ Kg}$$

4.2 Covariance

Deviations from the means

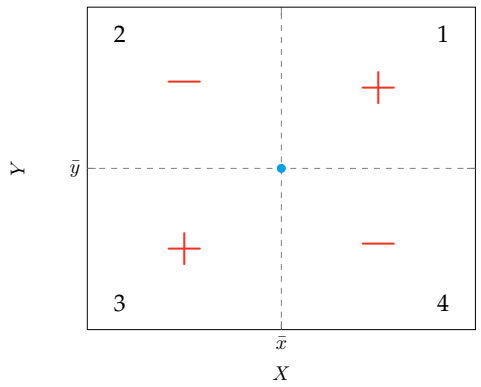
To study the relation between two variables, we have to analyze the joint variation of them.



Sign of deviations from the mean

Dividing the point cloud of the scatter plot in 4 quadrants centered in the mean point $(\bar{x}, \bar{y})$, the sign of deviations from the mean is:

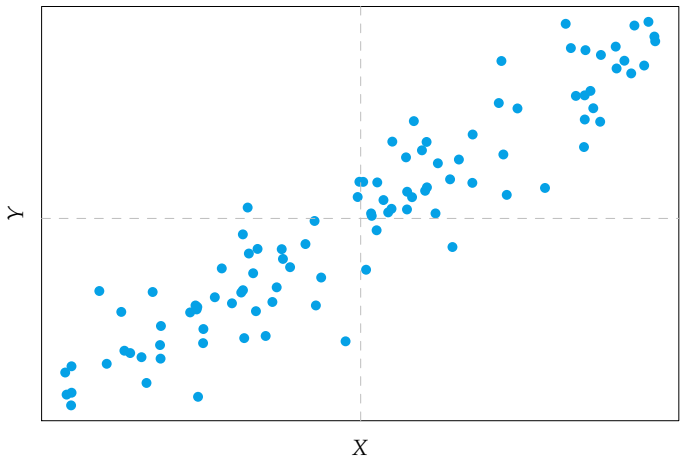
Quadrant	$(x_i - \bar{x})$	$(y_j - \bar{y})$	$(x_i - \bar{x})(y_j - \bar{y})$
1	+	+	+
2	−	+	−
3	−	−	+
4	+	−	−



Sign of the product of deviations from the mean

If there is an *increasing linear* relationship between the variables, most of the points will fall in quadrants 1 and 3, and the sum of the products of deviations from the mean will be positive.

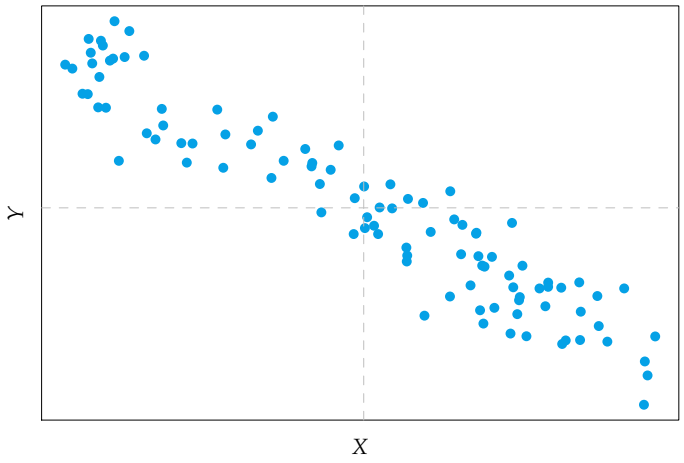
Increasing linear relation



$$\sum (x_i - \bar{x})(y_j - \bar{y}) = +$$

If there is an *decreasing linear* relationship between the variables, most of the points will fall in quadrants 2 and 4, and the sum of the products of deviations from the mean will be negative.

Decreasing linear relation



$$\sum (x_i - \bar{x})(y_j - \bar{y}) = -$$

Covariance

Using the products of deviations from the mean we get the statistic

Definition 22 (Sample covariance)**.** The *sample covariance* of a two-dimensional variable (X, Y) is the average of the products of deviations from the respective means.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_j - \bar{y}) n_{ij}}{n}$$

It can also be calculated using the formula

$$s_{xy} = \frac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y}.$$

The covariance measures the linear relation between two variables:

- If $s_{xy} > 0$ there exists an increasing linear relation.
- If $s_{xy} < 0$ there exists a decreasing linear relation.
- If $s_{xy} = 0$ there is no linear relation.

Covariance calculation

Example of heights and weights

Using the joint frequency table of the sample of heights and weights

X/Y	$[50,60)$	$[60,70)$	$[70,80)$	$[80,90)$	$[90,100)$	$[100,110)$	n_x
$(150,160]$	2	0	0	0	0	0	2
$(160,170]$	4	4	0	0	0	0	8
$(170,180]$	1	6	3	1	0	0	11
$(180,190]$	0	1	4	1	1	0	7
$(190,200]$	0	0	0	0	1	1	2
n_y	7	11	7	2	2	1	30

$$\bar{x} = 174.67 \text{ cm} \qquad \bar{y} = 69.67 \text{ Kg}$$

the covariance is

$$s_{xy} = \frac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y} = \frac{155 \cdot 55 \cdot 2 + 165 \cdot 55 \cdot 4 + \cdots + 195 \cdot 105 \cdot 1}{30} - 174.67 \cdot 69.67 =$$
$$= \frac{368200}{30} - 12169.26 = 104.07 \text{ cm} \cdot \text{Kg},$$

This means that there is a increasing linear relation between the weight and the height.

4.3 Regression

Regression

In most cases the goal of a dependency study is not only to detect a relation between two variables, but also to express that relation with a mathematical function,

$$y = f(x)$$

in order to predict the dependent variable for every value of the independent one.

The part of Statistics in charge of constructing such a function is **regression**, and the function is kwnon as **regression function** or **regression model**.

Simple regression models

Depending on the type of function there are a lot of types of regression models The most common are shown in the table below.

Model	Equation
Linear	$y = a + bx$
Quadratic	$y = a + bx + cx^2$
Cubic	$y = a + bx + cx^2 + dx^3$
Potential	$y = a \cdot x^b$
Exponential	$y = e^{a+bx}$
Logarithmic	$y = a + b \log x$
Inverse	$y = a + \frac{b}{x}$
Sigmoidal	$y = e^{a+\frac{b}{x}}$

The model choice depends on the shape of the points cloud in the scatter plot.

Residuals or predictive errors

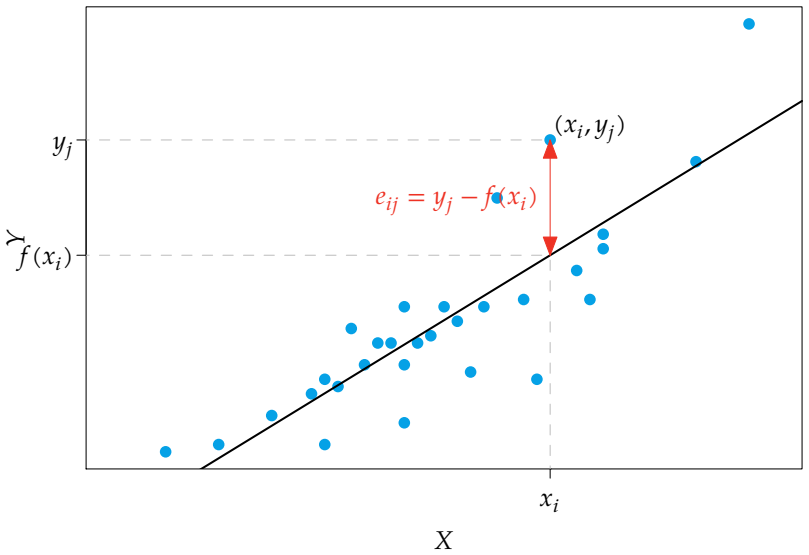
Once chosen the type of regression model, we have to determine which function of that family explains better the relation between the dependent and the independent variables, that is, the function that predict better the dependent variable.

That function is the function that minimize the distances from the observed values for Y in the sample to the predicted values of the regression function. These distances are known as *residuals* or *predictive errors*.

Definition 23 (Residuals or predictive errors)**.** Given a regression model $y = f(x)$ for a two-dimensional variable (X, Y), the *residual* or *predictive error* for every pair (x_i, y_j) of the sample is the difference between the observed value of the dependent variable y_j and the predicted value of the regression function for x_i,

$$e_{ij} = y_j - f(x_i).$$

Residuals or predictive errors on Y



Least squares fitting

A way to get the regression function is the *least squares method*, that determine the function that minimize the squared residuals.

$$\sum e_{ij}^2.$$

For a linear model $f(x) = a + bx$, the sum depends on two parameters, the intercept a, and the slope b of the straight line,

$$\theta(a,b) = \sum e_{ij}^2 = \sum (y_j - f(x_i))^2 = \sum (y_j - a - bx_i)^2.$$

This reduces the problem to determine the values of a and b that minimize this sum.

4.4 Regression line

Least squares fitting of a linear model

To solve the minimization problem, we have to set to zero the partial derivatives with respect to a and b.

$$\frac{\partial \theta(a,b)}{\partial a} = \frac{\partial \sum (y_j - a - bx_i)^2}{\partial a} = 0$$
$$\frac{\partial \theta(a,b)}{\partial b} = \frac{\partial \sum (y_j - a - bx_i)^2}{\partial b} = 0$$

And solving the equation system, we get

$$a = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} \qquad b = \frac{s_{xy}}{s_x^2}$$

This values minimize the residuals on Y and give us the optimal linear model.

Regression line

Definition 24 (Regression line)**.** Given a sample of a two-dimensional variable (X, Y), the *regression line* of Y on X is

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}).$$

The regression line of Y on X is the straight line that minimize the predictive errors on Y, therefore is the linear regression model that gives better predictions of Y.

Regression line calculation

Example of heights and weights

Using the previous sample of heights (X) and weights (Y) with the following statistics

$$\begin{array}{lll} \bar{x} = 174.67 \text{ cm} & s_x^2 = 102.06 \text{ cm}^2 & s_x = 10.1 \text{ cm} \\ \bar{y} = 69.67 \text{ Kg} & s_y^2 = 164.42 \text{ Kg}^2 & s_y = 12.82 \text{ Kg} \\ & s_{xy} = 104.07 \text{ cm} \cdot \text{ Kg} & \end{array}$$

the regression line of weight on height is

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}) = 69.67 + \frac{104.07}{102.06}(x - 174.67) = -108.49 + 1.02x$$

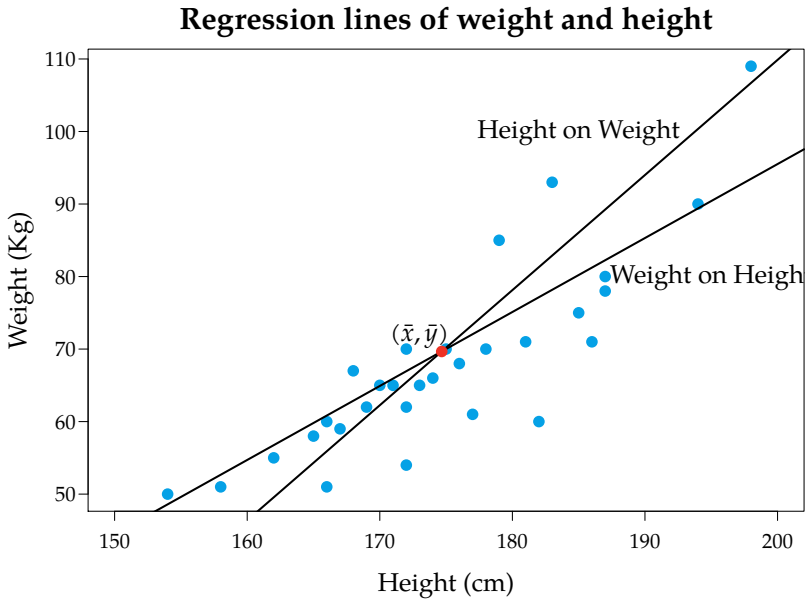
And the regression line of height on weight is

$$x = \bar{x} + \frac{s_{xy}}{s_y^2}(y - \bar{y}) = 174.67 + \frac{104.07}{164.42}(y - 69.67) = 130.78 + 0.63y$$

Observe that the regression lines are different!

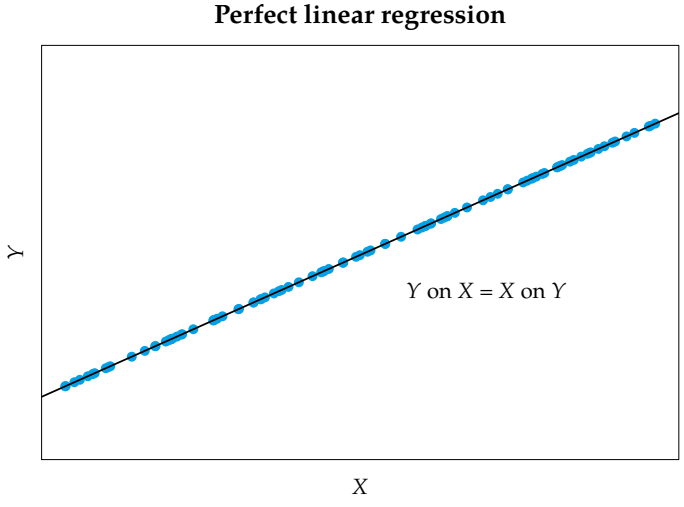
Regression lines

Example of heights and weights



Relative position of the regression lines

Usually, the regression line of Y on X and the regression line of X on Y are not the same, but they always intersect in the mean point $(\bar{x}, \bar{y})$. If there is a perfect linear relation between the variables, then both regression lines are the same, as that line makes both X-residuals and Y-residuals zero.

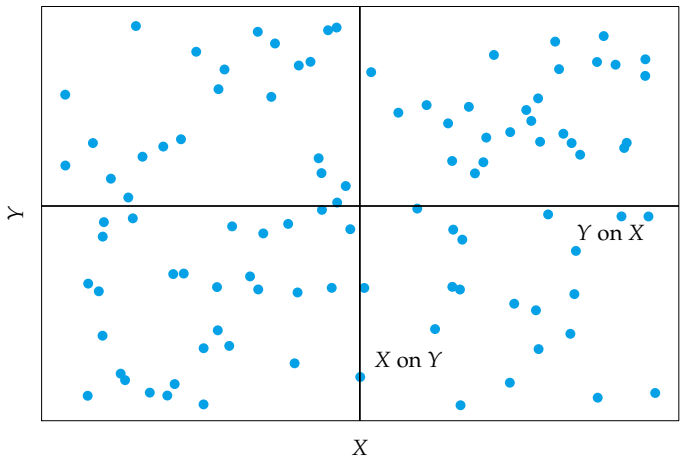


If there is no linear relation between the variables, then both regression lines are constant and equals to the respective means,

$$y = \bar{y}, \qquad x = \bar{x},$$

So, they intersect perpendicularly.

Non-linear regression



Regression coefficient

The most important parameter of a regression line is the slope.

Definition 25 (Regression coefficient b_{yx})**.** Given a sample of a two-dimensional variable (X, Y), the *regression coefficient* of the regression line of Y on X is its slope,

$$b_{yx} = \frac{s_{xy}}{s_x^2}$$

The regression coefficient has always the same sign than the covariance.

It measures how the dependent variable changes in relation to the independent one according to the regression line. In particular, it gives the number of units that the dependent variable increases or decreases for every unit that increases the independent variable.

Regression coefficient

Example of heights and weights

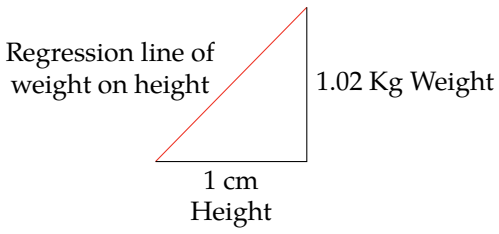
In the sample of heights and weights, the regression line of weight on height was

$$y = -108.49 + 1.02x.$$

Thus, the regression coefficient of weight on height is

$$b_{yx} = 1.02\text{Kg/cm}.$$

That means that, according to the regression line of weight on height, the weight will increase 1.02 Kg for every cm that increases the height.



Regression predictions

Example of heights and weights

Usually the regression models are used to predict the dependent variable for some values of the independent variable.

Watch out! To get the best predictions of a variable you have to use the regression line where that variable plays the dependent variable role.

Thus, in the sample of heights and weights, to predict the weight of a person with a height of 180 cm, we have to use the regression line of weight on height,

$$y = -108.49 + 1.02 \cdot 180 = 75.11 \text{ Kg}.$$

But to predict the height of a person with a weight of 79 Kg, we have to use the regression line of height on weight,

$$x = 130.78 + 0.63 \cdot 79 = 180.55 \text{ cm}.$$

However, how reliable are these predictions?

4.5 Correlation

Correlation

Once we have a regression model, in order to see if it is a good predictive model we have to assess the goodness of fit of the model and the strength of the of relation set by it. The part of Statistics in charge of this is **correlation**.

The correlation study the residuals of a regression model: the smaller the residuals, the greater the goodness of fit, and the stronger the relation set by the model.

Residual variance

To measure the goodness of fit of a regression model is common to use the *residual variance*.

Definition 26 (Sample residual variance s^2_{ry})**.** Given a regression model $y = f(x)$ de of a two-dimensional variable (X, Y), its *sample residual variance* is the average of the squared residuals,

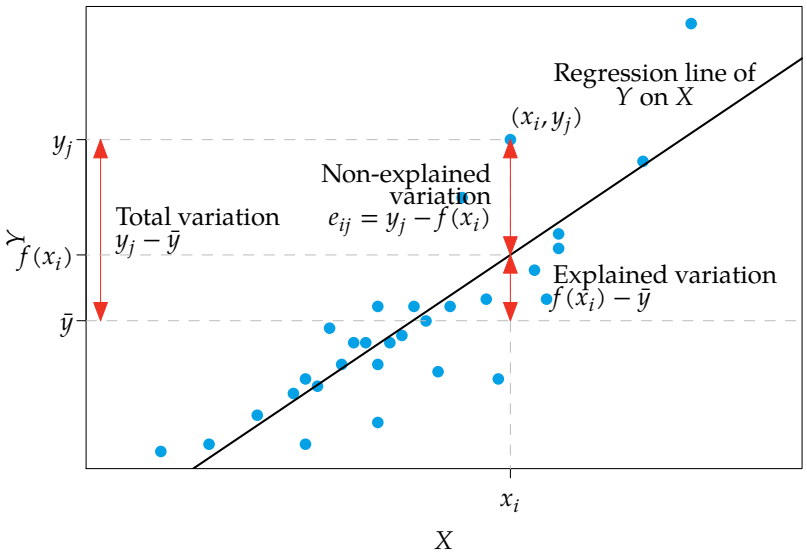
$$s^2_{ry} = \frac{\sum e^2_{ij} n_{ij}}{n} = \frac{\sum (y_j - f(x_i))^2 n_{ij}}{n}.$$

The greater the residuals, the greater the residual variance and the smaller the goodness of fit.

When the linear relation is perfect, the residuals are zero and the residual variance is zero. Conversely, when there are no relation, the residuals coincide with deviations from the mean, and the residual variance is the same than the variance of the dependent variable.

$$0 \leq s^2_{ry} \leq s^2_y$$

Explained and non-explained variation



4.6 Correlation coefficients

Coefficient of determination

From the residual variance is possible to define another correlation statistic easier to interpret.

Definition 27 (Sample coefficient of determination r^2)**.** Given a regression model $y = f(x)$ of a two-dimensional variable (X, Y), its *coefficient of determination* is

$$r^2 = 1 - \frac{s_{ry}^2}{s_y^2}$$

As the residual variance ranges from 0 to s_y^2,

$$0 \leq r^2 \leq 1$$

The greater r^2 is, the greater the goodness of fit of the regression model, and more reliable will be its predictions. In particular,

- If $r^2 = 0$ then there is no relation as set by the regression model.
- If $r^2 = 1$ then the relation set by the model is perfect.

Linear coefficient of determination

When the regression model is linear, the residual variance is

$$\begin{aligned}
s_{ry}^2 &= \sum e_{ij}^2 f_{ij} = \sum (y_j - f(x_i))^2 f_{ij} = \sum \left(y_j - \bar{y} - \frac{s_{xy}}{s_x^2}(x_i - \bar{x}) \right)^2 f_{ij} = \\
&= \sum \left((y_j - \bar{y})^2 + \frac{s_{xy}^2}{s_x^4}(x_i - \bar{x})^2 - 2\frac{s_{xy}}{s_x^2}(x_i - \bar{x})(y_j - \bar{y}) \right) f_{ij} = \\
&= \sum (y_j - \bar{y})^2 f_{ij} + \frac{s_{xy}^2}{s_x^4} \sum (x_i - \bar{x})^2 f_{ij} - 2\frac{s_{xy}}{s_x^2} \sum (x_i - \bar{x})(y_j - \bar{y}) f_{ij} = \\
&= s_y^2 + \frac{s_{xy}^2}{s_x^4} s_x^2 - 2\frac{s_{xy}}{s_x^2} s_{xy} = s_y^2 - \frac{s_{xy}^2}{s_x^2}.
\end{aligned}$$

and the coefficient of determination is

$$r^2 = 1 - \frac{s_{ry}^2}{s_y^2} = 1 - \frac{s_y^2 - \frac{s_{xy}^2}{s_x^2}}{s_y^2} = 1 - 1 + \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{s_{xy}^2}{s_x^2 s_y^2}.$$

Linear coefficient of determination calculation

Example of heights and weights

In the sample of heights and weights, we had

$$\begin{array}{ll} \bar{x} = 174.67 \text{ cm} & s_x^2 = 102.06 \text{ cm}^2 \\ \bar{y} = 69.67 \text{ Kg} & s_y^2 = 164.42 \text{ Kg}^2 \\ s_{xy} = 104.07 \text{ cm} \cdot \text{ Kg} & \end{array}$$

Thus, the linear coefficient of determination is

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{(104.07 \text{ cm Kg})^2}{102.06 \text{ cm}^2 \cdot 164.42 \text{ Kg}^2} = 0.65.$$

This means that the linear model of weight on height explains the 65% of the variation of weight, and the linear model of height on weight also explains 65% of the variation of height.

Correlation coefficient

Definition 28 (Sample correlation coefficient)**.** Given a sample of a two-dimensional variable (X, Y), the *sample correlation coefficient* is the square root of the linear coefficient of determination, with the sign of the covariance,

$$r = \sqrt{r^2} = \frac{s_{xy}}{s_x s_y}.$$

As r^2 ranges from 0 to 1, r ranges from -1 to 1,

$$-1 \leq r \leq 1$$

The correlation coefficient measures the strength of the linear association but also its direction (increasing or decreasing):

- If $r = 0$ then there is no linear relation.
- Si $r = 1$ then there is a perfect increasing linear relation.
- Si $r = -1$ then there is a perfect decreasing linear relation.

Correlation coefficient

Example of heights and weights

In the sample of heights and weights, we had

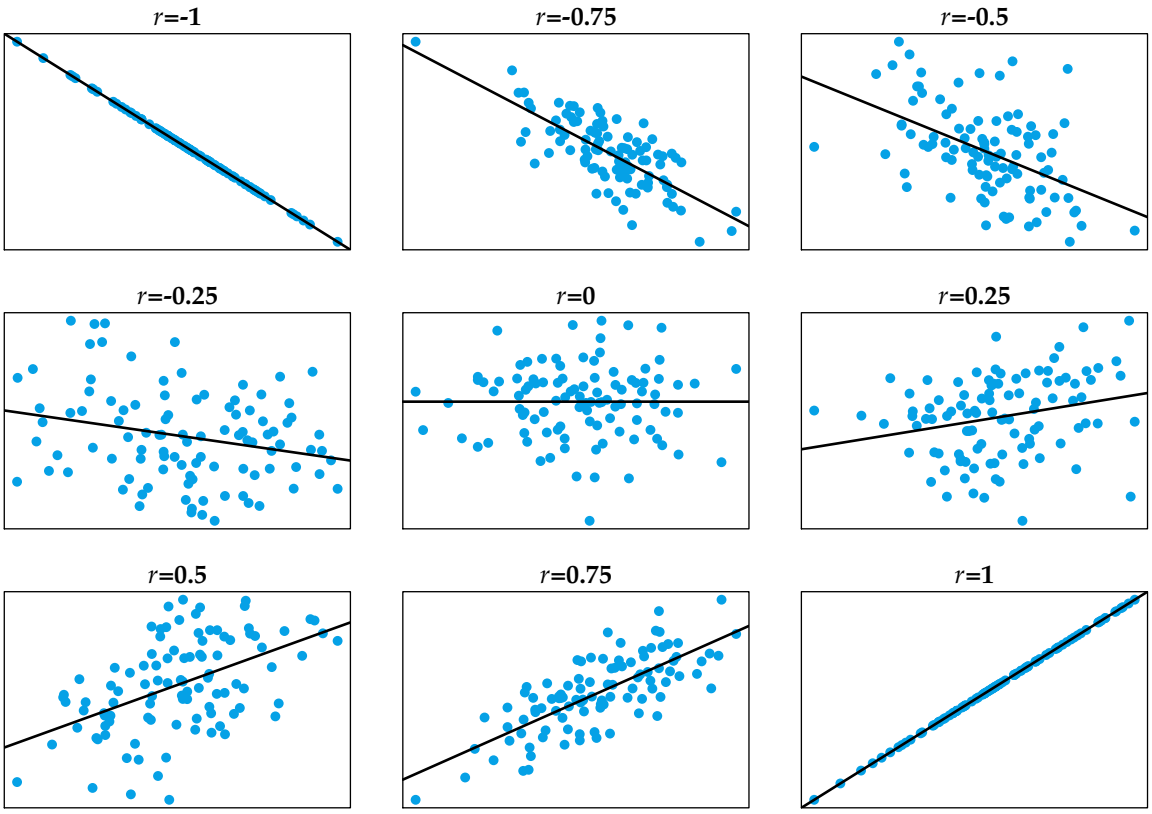
$$\begin{array}{ll} \bar{x} = 174.67 \text{ cm} & s_x^2 = 102.06 \text{ cm}^2 \\ \bar{y} = 69.67 \text{ Kg} & s_y^2 = 164.42 \text{ Kg}^2 \\ s_{xy} = 104.07 \text{ cm} \cdot \text{ Kg} & \end{array}$$

Thus, the correlation coefficient is

$$r = \frac{s_{xy}}{s_x s_y} = \frac{104.07 \text{ cm Kg}}{10.1 \text{ cm} \cdot 12.82 \text{ Kg}} = +0.8.$$

This means that there is a rather strong linear increasing relation between height and weight.

Different correlations



Reliability of regression predictions

The coefficient of determination explains the goodness of fit of a regression model, but there are other factors that influence the reliability of regression predictions:

- The coefficient of determination: The greater r^2, the greater the goodness of fit and the more reliable the predictions.
- The variability of the population distribution: The greater the variation, the more difficult to predict and the less reliable the predictions.
- The sample size: The greater the sample size, the more information we have and the more reliable the predictions.

In addition, we have to take into account that a regression model is only valid for the range of values observed y the sample. That means that, as we don't have any information outside that range, we must not do predictions for values far from that range.

4.7 Non-linear regression

Non-linear regression

The fit of a non-linear regression can be also done by least square fitting method.

However, in some cases the fitting of a non-linear model can be reduced to the fitting of a linear model applying a simple transformation to the variables of the model.

Transformations of non-linear regression models

- **Logarithmic model**: A logarithmic model $y = a + b\log x$ can be transformed in a linear model with the change $t = \log x$:
$$y = a + b\log x = a + bt.$$
- **Exponential model**: An exponential model $y = e^{a+bx}$ can be transformed in a linear model with the change $z = \log y$:
$$z = \log y = \log(e^{a+bx}) = a + bx.$$
- **Potential model**: A potential model $y = ax^b$ can be transformed in a linear model with the changes $t = \log x$ and $z = \log y$:
$$z = \log y = \log(ax^b) = \log a + b\log x = a' + bt.$$
- **Inverse model**: An inverse model $y = a + b/x$ can be transformed in a linear model with the change $t = 1/x$:
$$y = a + b(1/x) = a + bt.$$
- **Sigmoidal model**: A sigmoidal model $y = e^{a+b/x}$ can be transformed in a linear model with the changes $t = 1/x$ and $z = \log y$:
$$z = \log y = \log(e^{a+b/x}) = a + b(1/x) = a + bt.$$

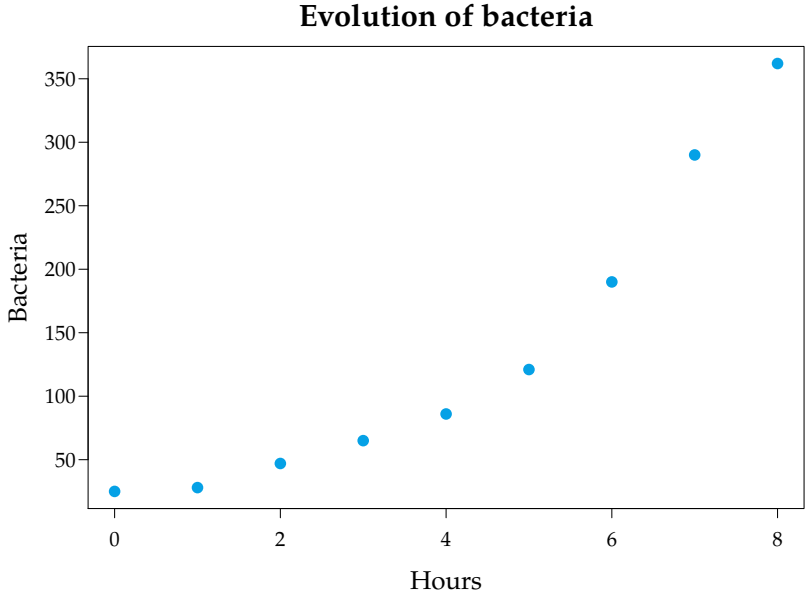
Exponential relation

Example of evolution of a bacterial culture

The number of bacteria in a culture evolves with time according to the table below.

Hours	Bacteria
0	25
1	28
2	47
3	65
4	86
5	121
6	190
7	290
8	362

The scatter plot of the sample is showed below.



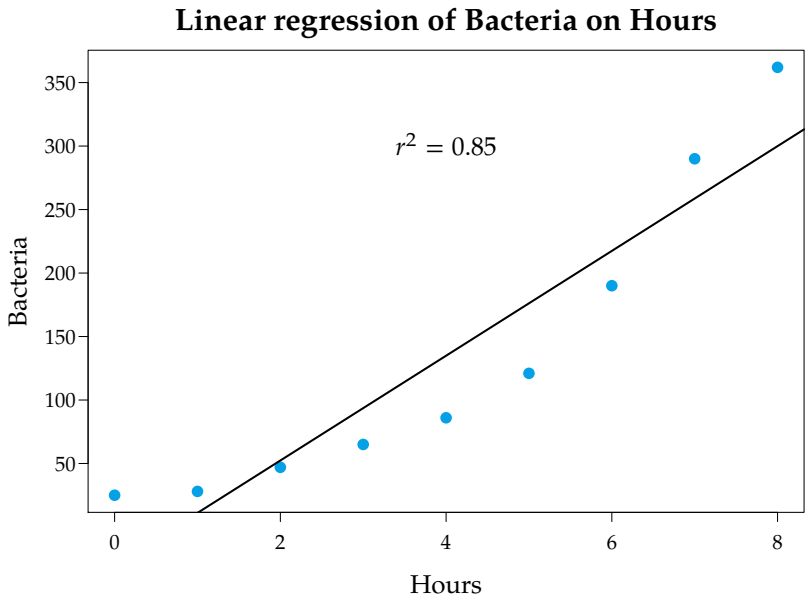
Exponential relation

Example of evolution of a bacterial culture

Fitting a linear model we get

Hours	Bacteria
0	25
1	28
2	47
3	65
4	86
5	121
6	190
7	290
8	362

$$\text{Bacteria} = -30.18 + 41,27\,\text{Hours}, \text{ with } r^2 = 0.85.$$



Is a good model?

Fitting an exponential regression model

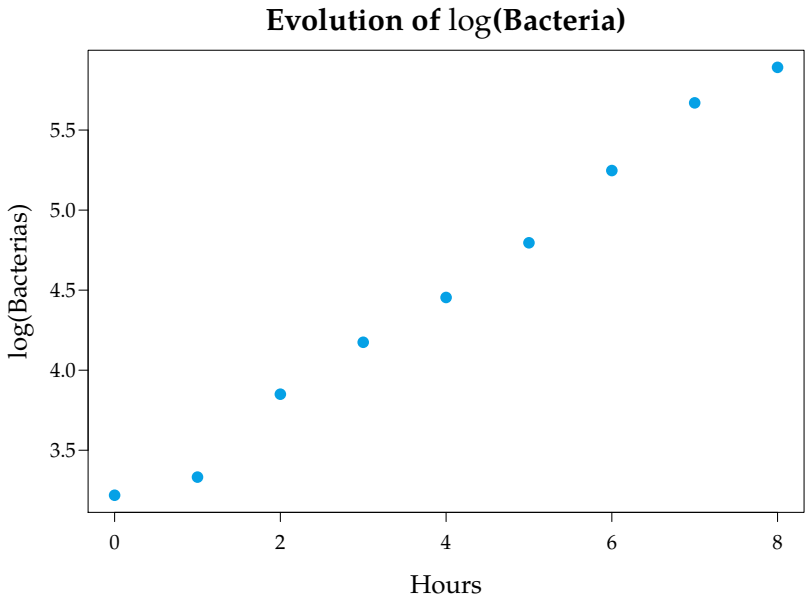
Example of evolution of a bacterial culture

Although the linear model is not bad, according to the shape of the point cloud of the scatter plot, an exponential model looks more suitable.

To construct an exponential model $y = e^{a+bx}$ we can apply the transformation $z = \log y$, that is,

applying a logarithmic transformation to the dependent variable.

Hours	Bacteria	log(Bacteria)
0	25	3.22
1	28	3.33
2	47	3.85
3	65	4.17
4	86	4.45
5	121	4.80
6	190	5.25
7	290	5.67
8	362	5.89



Fitting an exponential regression model

Example of evolution of a bacterial culture

Now it only remains to compute the regression line of the logarithm of bacteria on hours,

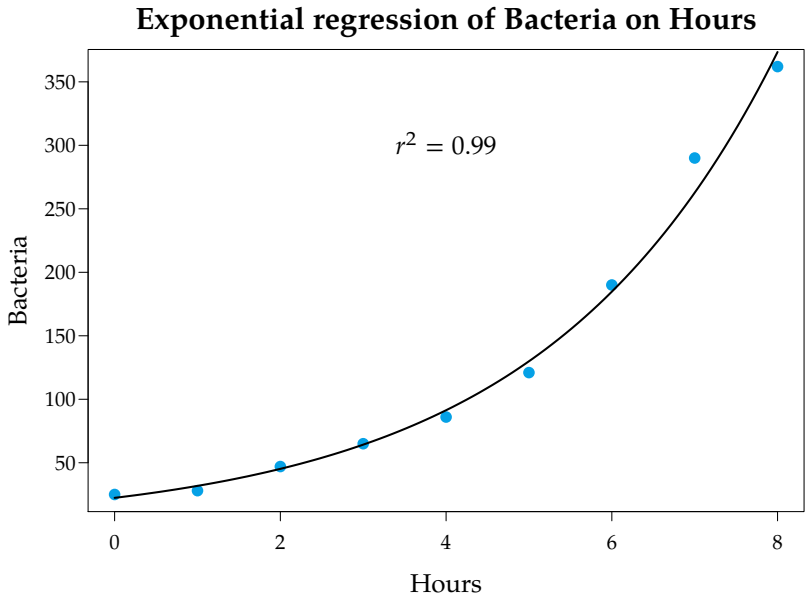
$$\log(\text{Bacteria}) = 3.107 + 0.352\,\text{Horas},$$

and, undoing the change of variable,

$$\text{Bacteria} = e^{3.107+0.352\,\text{Hours}},$$

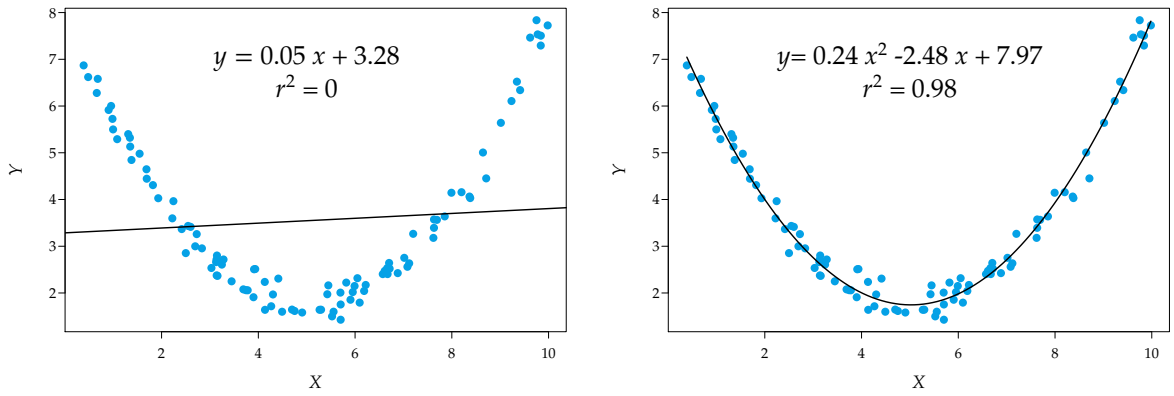
with $r^2 = 0.99$.

Thus, the exponential model fits much better than the linear model.



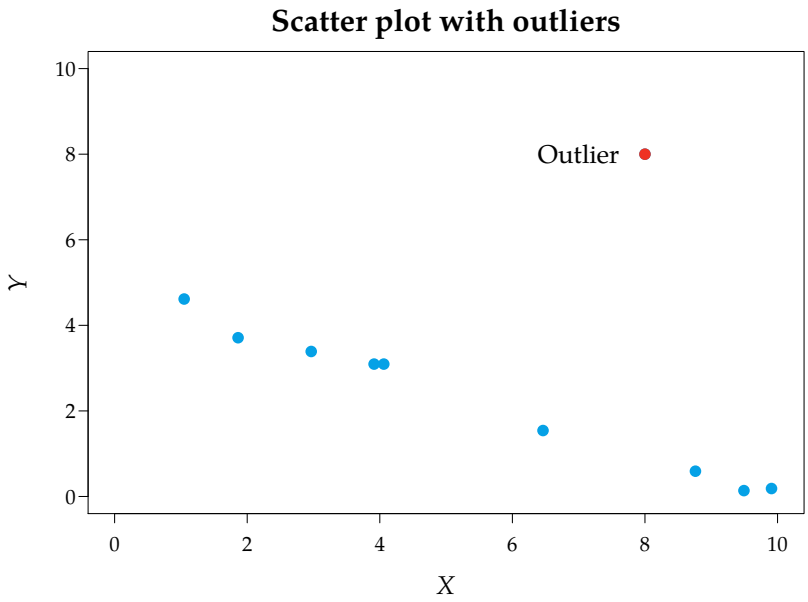
Lack of fit doesn't mean independence

It's important to note that every regression model has its own coefficient of determination. Thus, a coefficient of determination near zero means that there is no relation as set by the model, but *that doesn't mean that the variables are independent*, cause it could be a different type of relation.



Outliers in regression

Outliers in regression studies are points that clearly doesn't follow the tendency of the rest of points, even if the values of the pair are not outliers for every variable separately.



Outliers influence in regression

Outliers in regression studies can provoke drastic changes in the regression models.