Elementary Statistics Manual

Alfredo Sánchez Alberca (asalber@ceu.es)

Sep 2019

Department of Applied Math and Statistics
CEU 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 and 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 requires data handling, especially with large volumes of data, such as Physics, Chemistry, Medicine, Psychology, Economics or Social Sciences.

But, Why is Statistics necessary?

A changing World

Scientists try to study the World. A World with a high variability that makes difficult determining the behavior 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 an or more features that have all the elements and only them. 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 to 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 its individuals.

However, this is not always possible for several reasons:

- The population size is infinite or too large to study all its individuals.
- The operations that individuals undergo are destructive.
- The cost, both in 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 it 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 to the population.

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

Sample size determination

Small sample of pixels of a picture

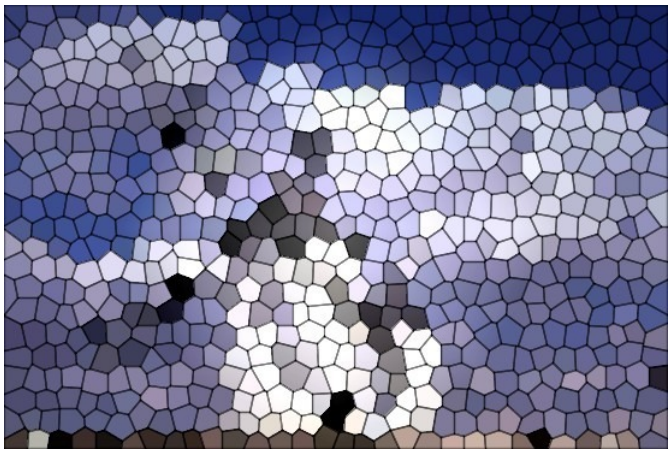
Example. To understand what a sufficient sample size means we can use a picture example. A digital photography consists of a lot of small points called pixels disposed in an big array layout with rows and columns (the more rows and columns, the more resolution the picture has). Here the picture is the population and every pixel is an 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!!

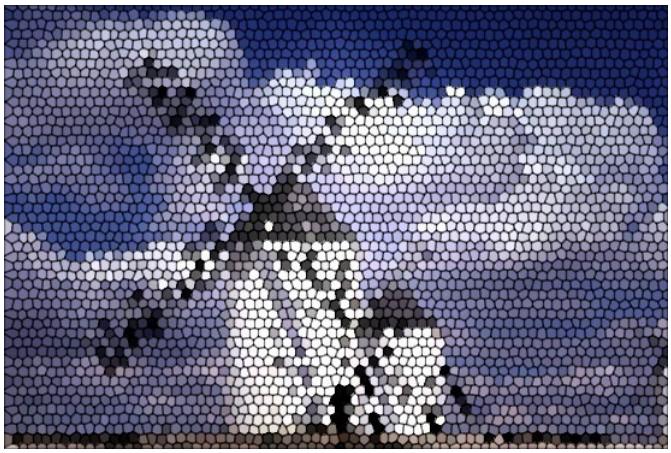
Sample size determination

Large sample of pixels of a picture

Surely you has not been able to guess the motif because the number of pixels picked in the sample is too small to understand the variability of colors in the 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!

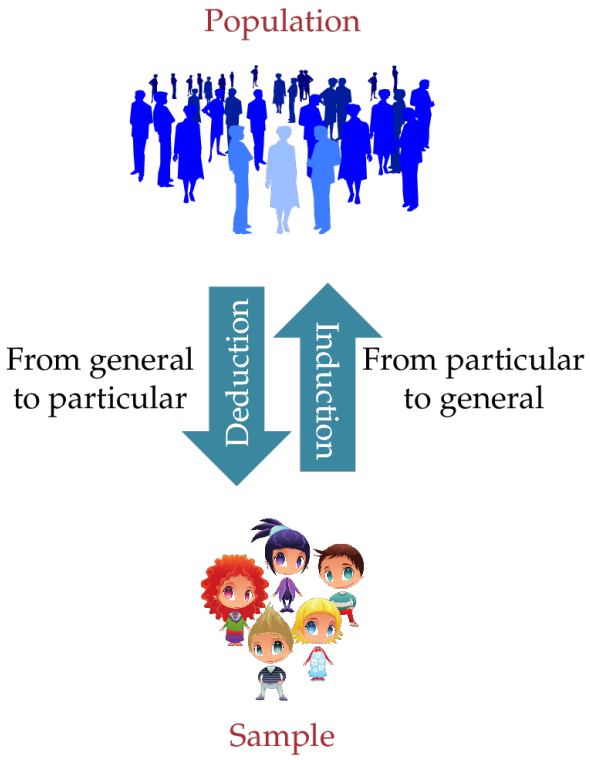
Sample size determination

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 does not 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!). However, 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 exist 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 they do not guarantee the representativeness of the sample. Nevertheless, usually they 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. In this way the population does not 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.

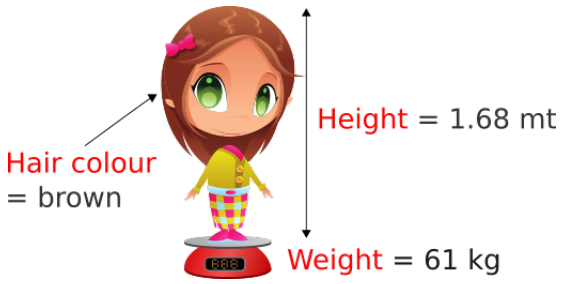
1.4 Statistical variables

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.



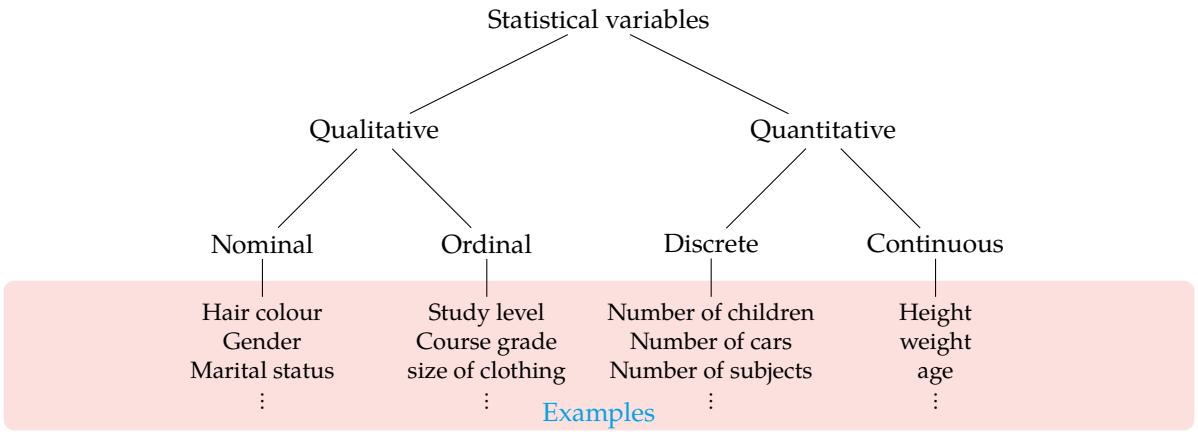
Types of statistical variables

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 hair colour or the gender.
  - **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 do not depend 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 depend 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

- **Experimental:** When the independent variables are manipulated in order to see the effect that that change has on the dependent variables.
  Example In a study on the performance of students in a test, the teacher manipulates the methodology and creates two or more groups following different methodologies.
- **Non-experimental:** When the independent variables are not manipulated. That does not mean 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

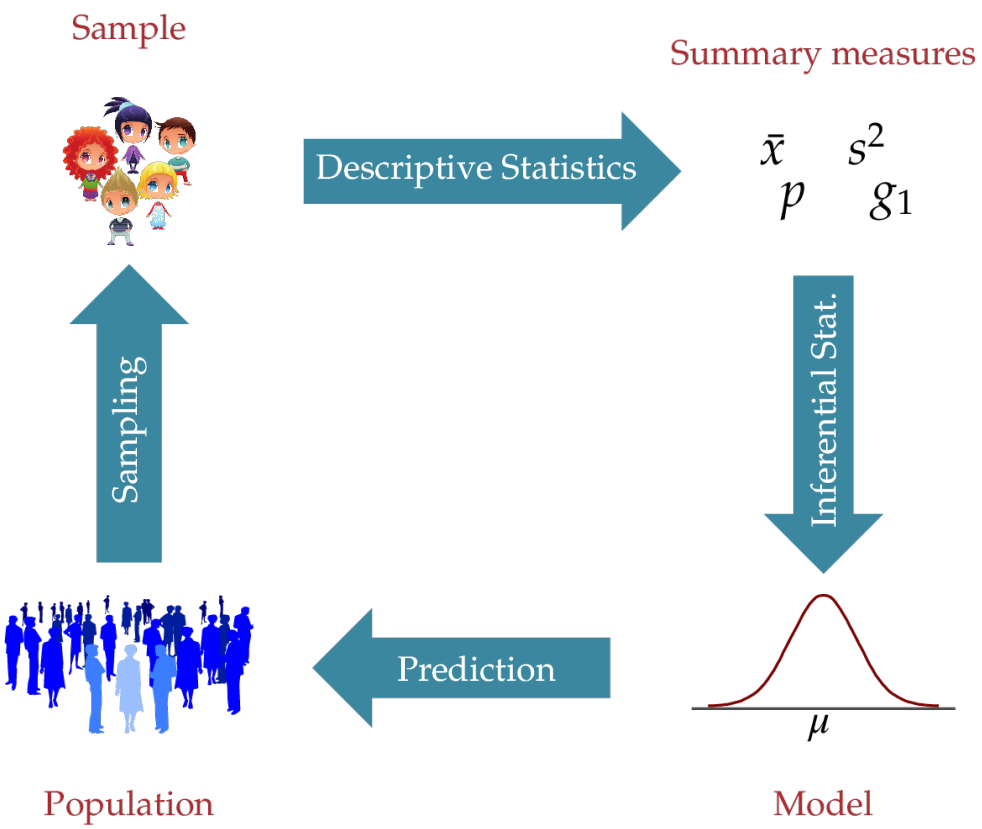
1.5 Phases of a statistical study

Phases of a statistical study

Usually a statistical study goes through the following phases:

1. The study begins with a previous design in which the study goals, the population, the variables to measure and the required sample size are set.
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 intends 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, analyzing and summarizing the information contained in the sample.

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

1. To classify, group and sort the data of the sample.
2. To tabulate and plot data according to their frequencies.
3. To 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 by measuring the variable in the individuals of the sample and classifying the values.

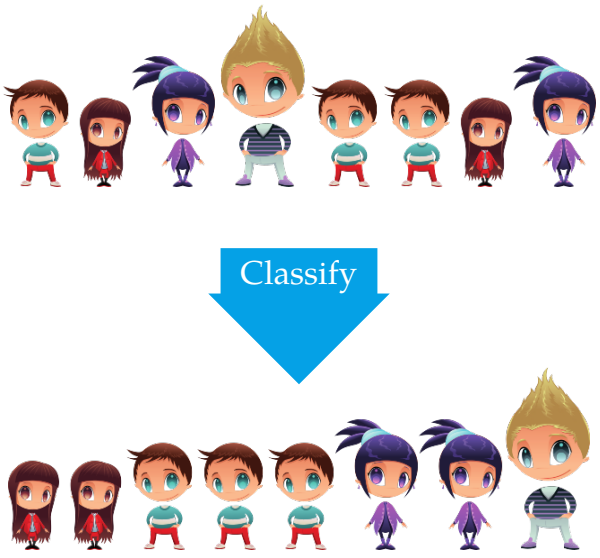
There are two ways of classifying data:

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

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

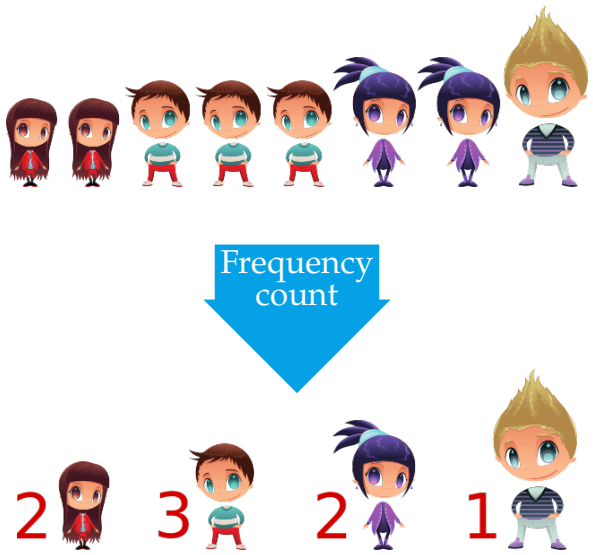
Sample classification

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 we define

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

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

- **Cumulative absolute frequency** N_i: 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: 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 does not 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 types 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 whether 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 of a set of bars, one for every value or category of the variable, plotted on a coordinate system.

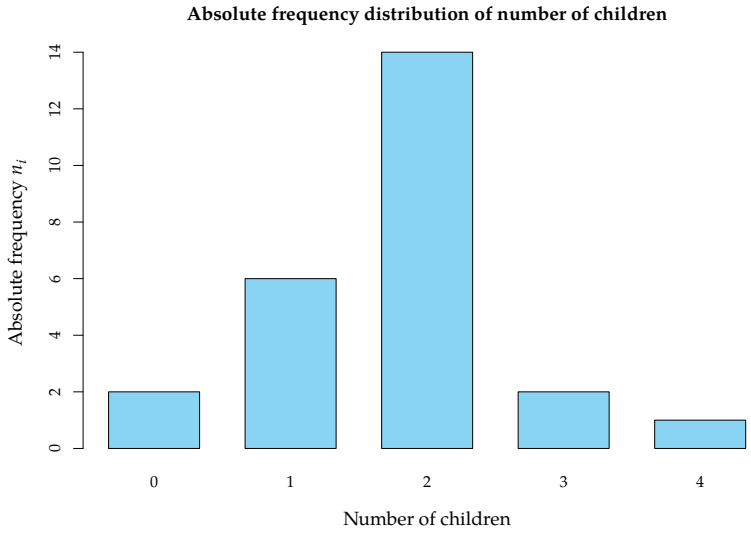
Usually the values or categories of the variable are represented on the x-axis, and the frequencies on the y-axis. For each value or category of the variable, a bar is draw to the height of its frequency. The width of the bar is not important but bars should be clearly separated among them.

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.

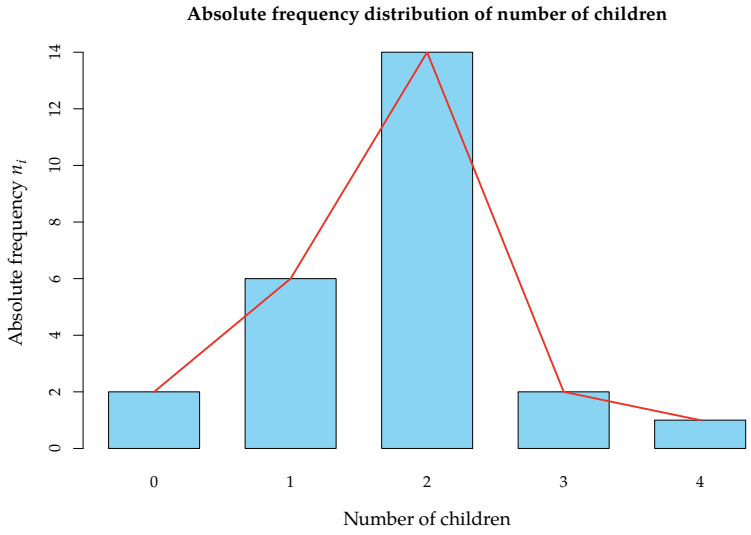
Absolute frequency bar chart

Non-grouped data



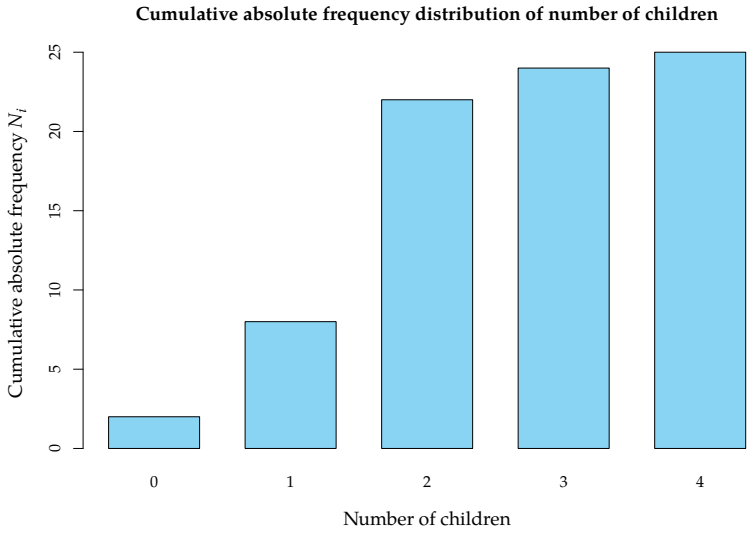
Absolute frequency line chart or polygon

Non-grouped data



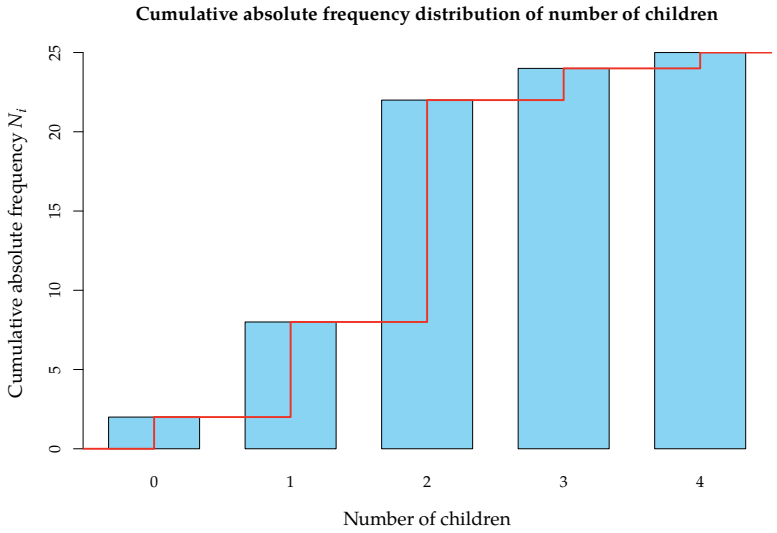
Cumulative absolute frequency bar chart

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.

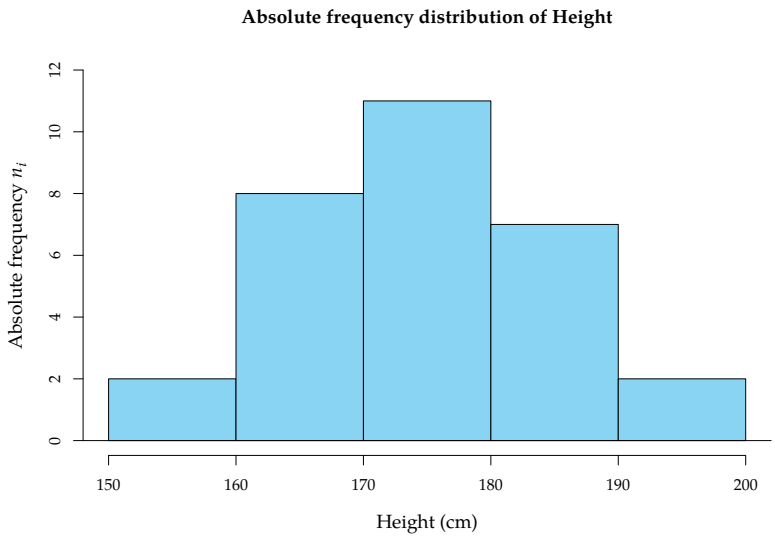
Usually the classes or grouping intervals are represented on the x-axis, and the frequencies on the y-axis. For each class, a bar is draw to the height of its frequency. Contrary to bar charts, the width of bars coincides with the width of classes, and there are no space between two consecutive bars.

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.

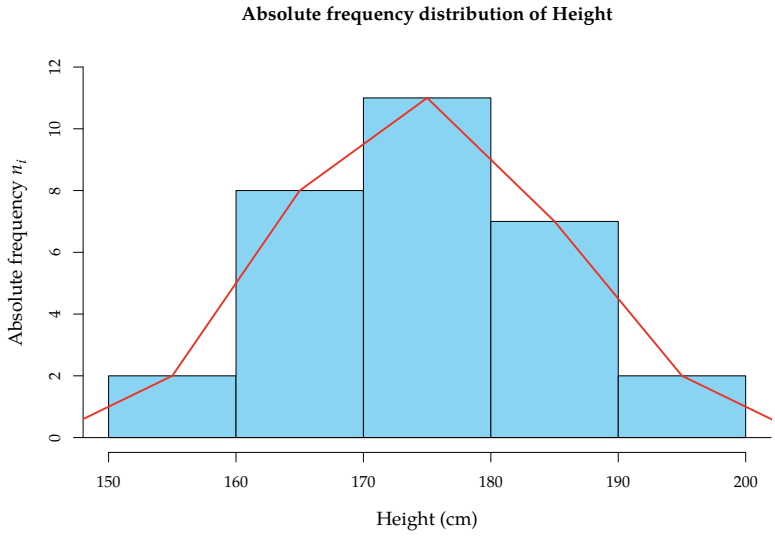
Absolute frequency histogram

Grouped data



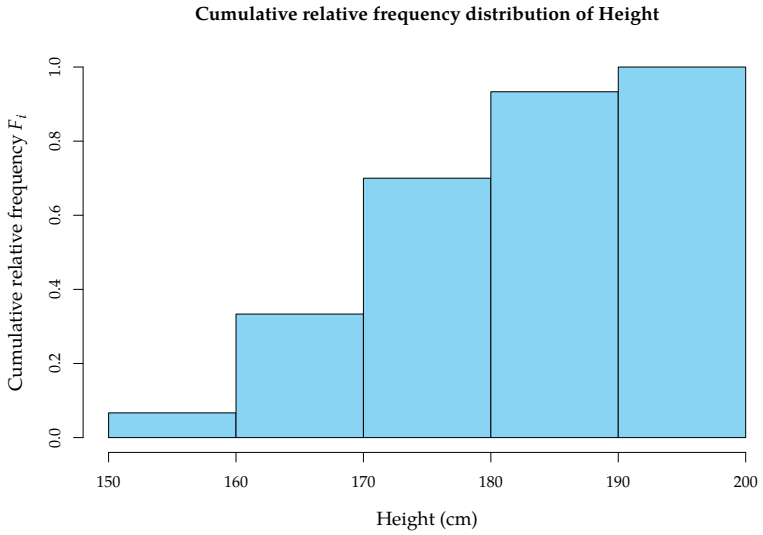
Absolute frequency polygon

Grouped data



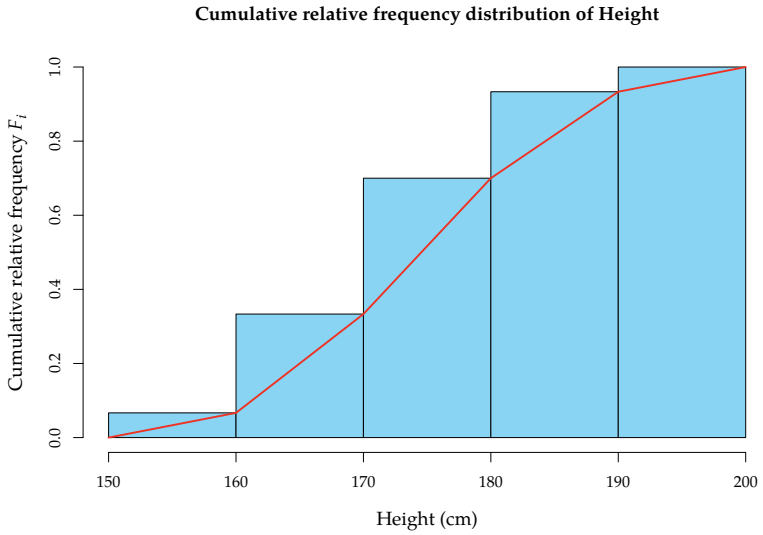
Cumulative relative frequency histogram

Grouped data



Cumulative relative frequency line chart or ogive

Grouped data



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

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

Pie chart

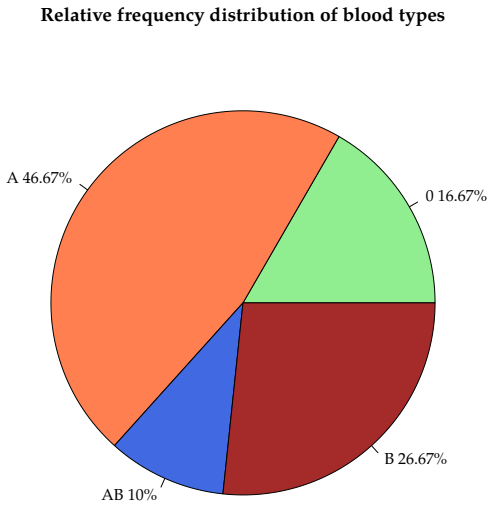
A **pie chart** consists of a circle divided in slices, one for every value or category of the variable. Each slice is called a **sector** and its angle or area is proportional to the frequency of the corresponding value or category.

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, because it is easier to perceive differences in one dimension (length of bars) than in two dimensions (areas of sectors).

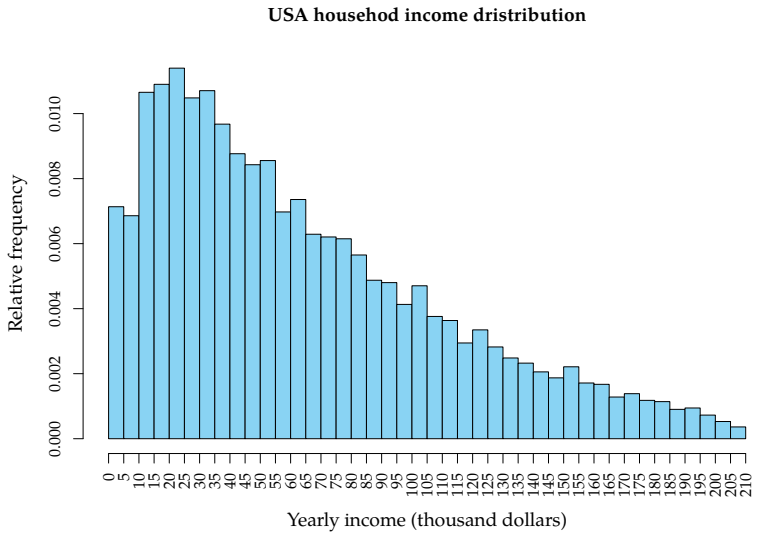
Pie chart

Nominal variables



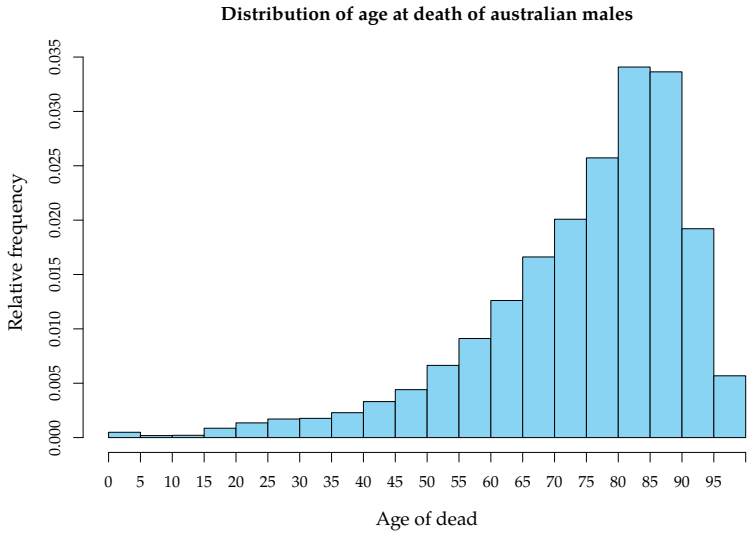
Distribution shapes

Household incomes



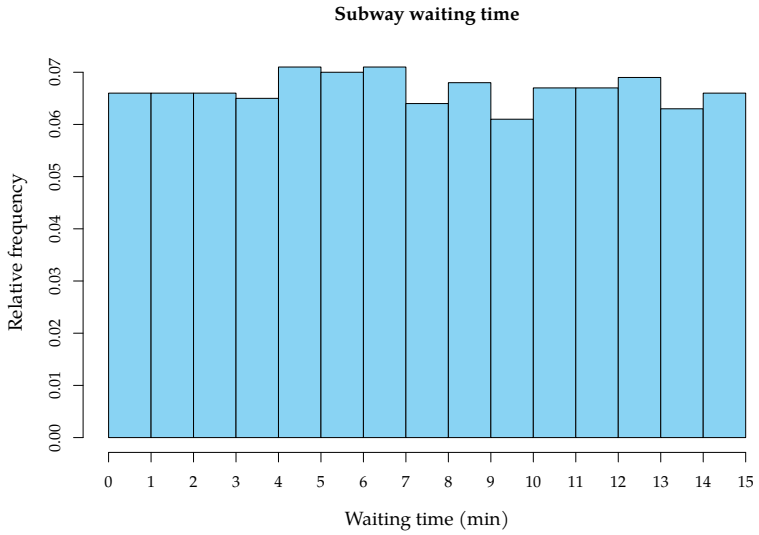
Distribution shapes

Age at death



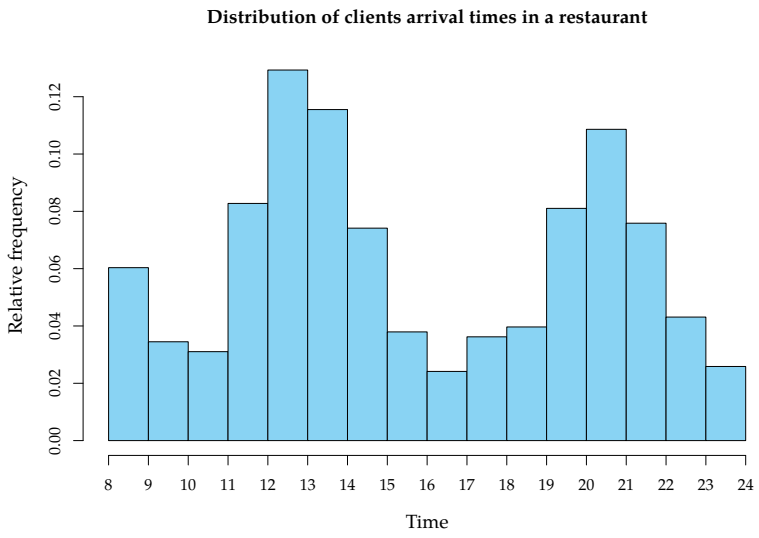
Distribution shapes

Subway waiting time



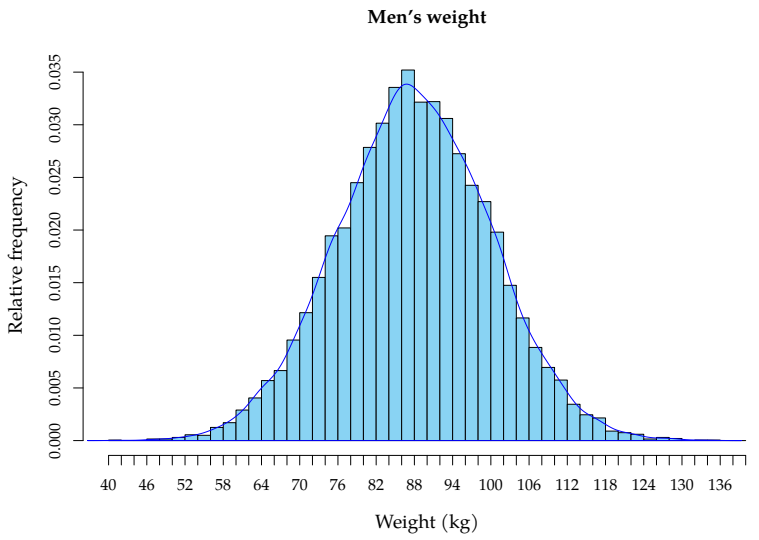
Distribution shapes

Arrival time of clients of a restaurant



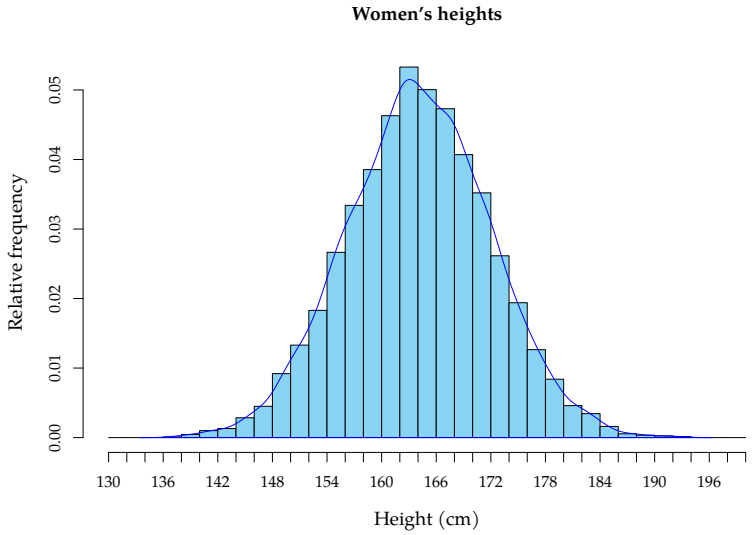
Bell shaped distribution

Women's weight



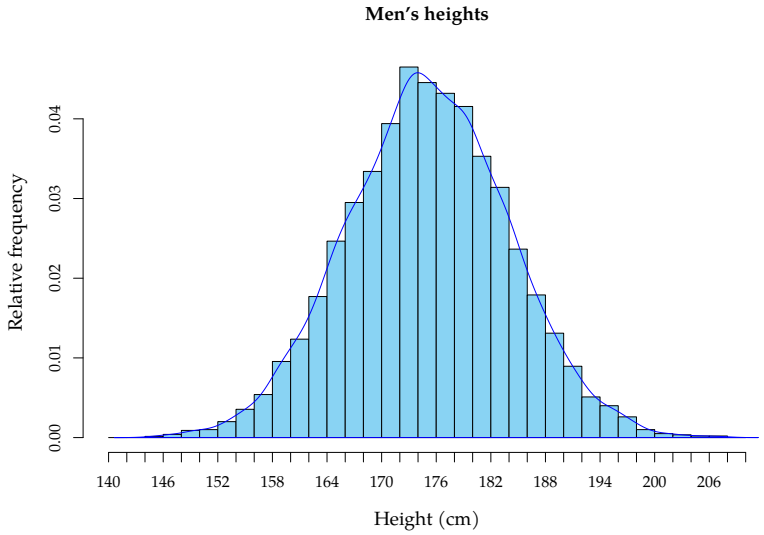
Bell shaped distribution

Women's height



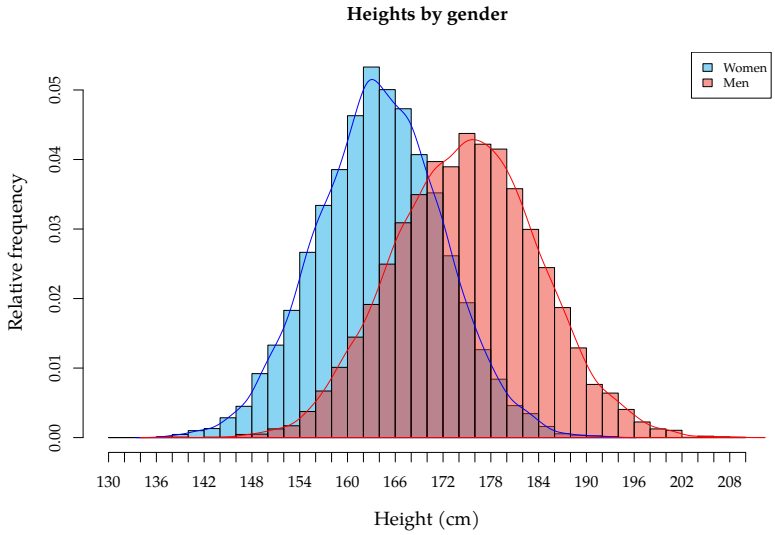
Bell shaped distribution

Men's height



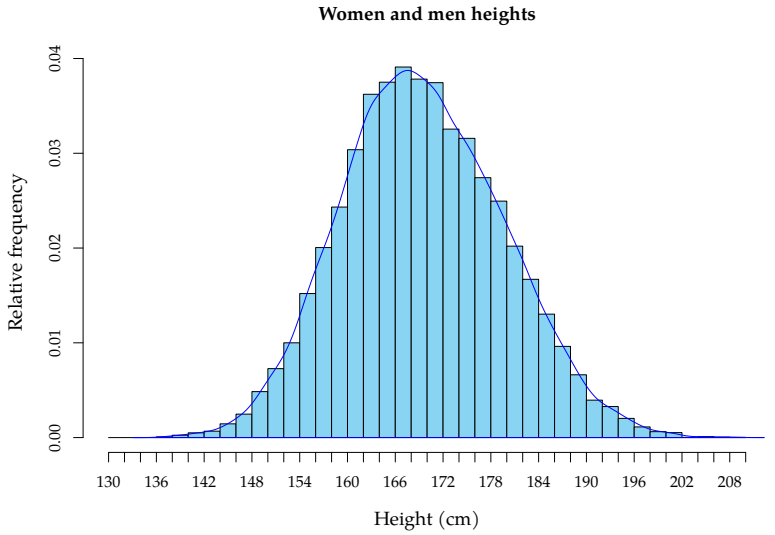
Bell shaped distribution

Height by gender



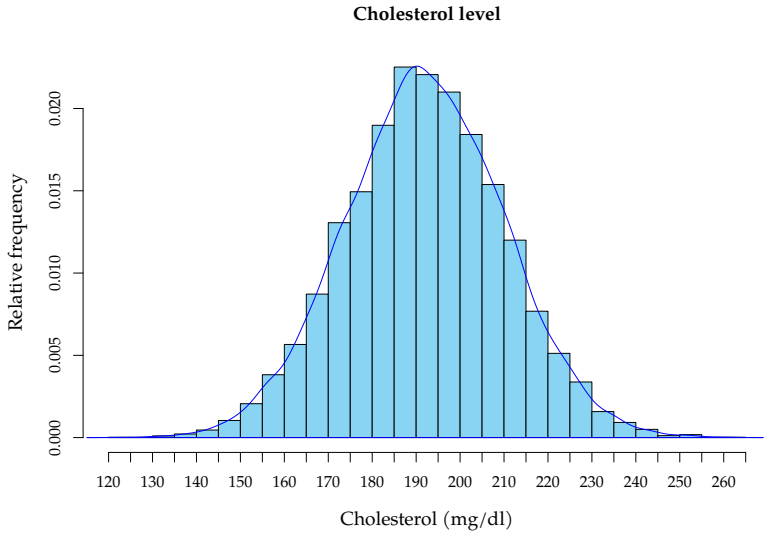
Bell shaped distribution

Heights



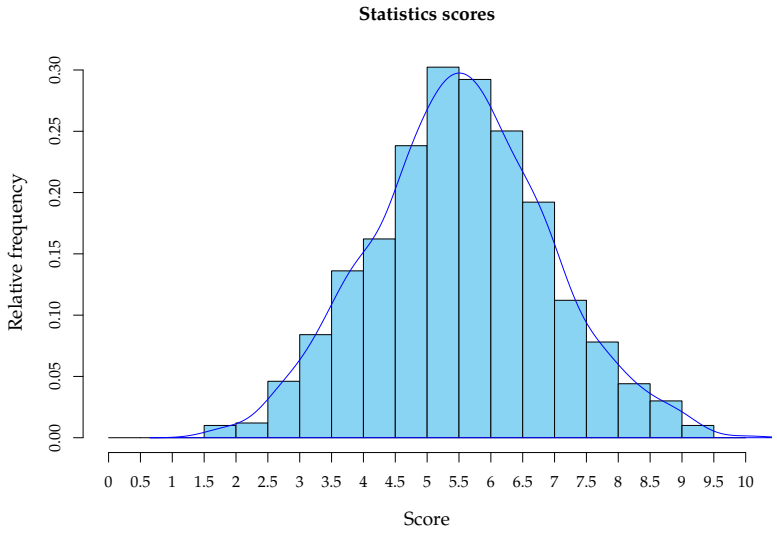
Bell shaped distribution

Cholesterol



Bell shaped distribution

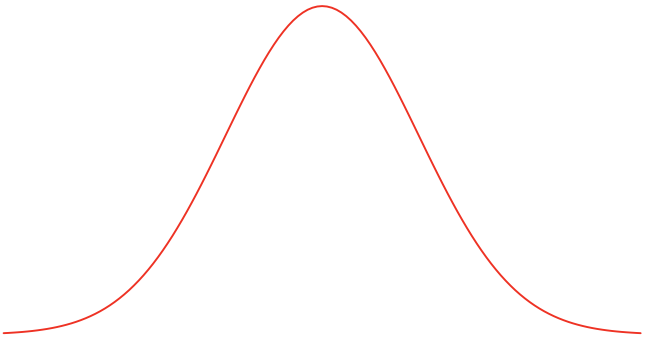
Statistics scores



Normal distribution

The bell shaped distribution appears so frequently in Nature that is known as the **Normal distribution** or **Gaussian distribution**.

Gauss bell



Outliers

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

It is important to find out outliers before doing any analysis, because outliers usually distort the results.

They always appears in the ends of the distribution, and can be found out easily with a box and whiskers chart (as be show 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 is an error.
- Replace the outlier by the lower or higher value in the distribution that is not an outlier if it is not an error and the outlier does not fit the theoretical distribution.
- Leave the outlier if it is 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 is 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 "tailedness" of the distribution.

3.1 Location statistics

Location statistics

There are two groups:

Central location measures: They measure the values where data are concentrated, usually at 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

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

It can be calculated from the frequency table 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, because 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

A student wants to calculate a representative measure of his/her 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 it is better to use the weighted mean, using the credits as weights.

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

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 $F_{Me} = 0.5$.

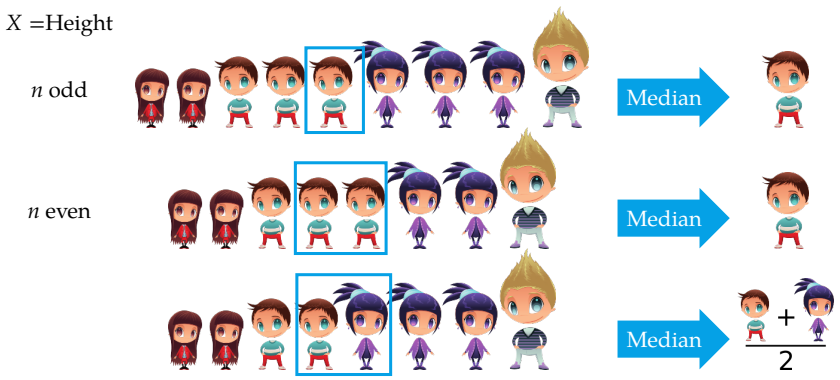
Watch out! It can not be calculated for nominal variables.

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.

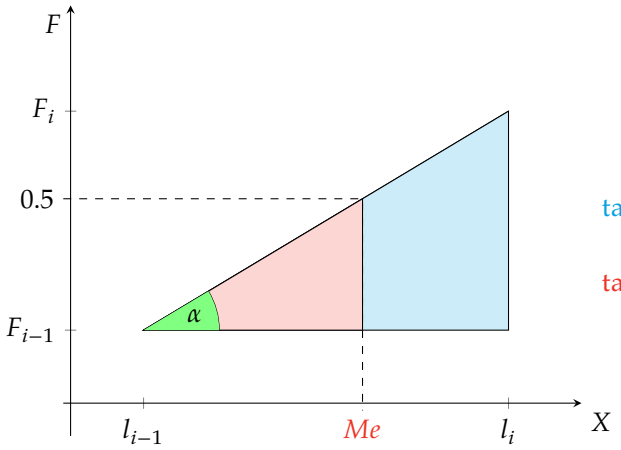
$$0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, \boxed{2}, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 4$$

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



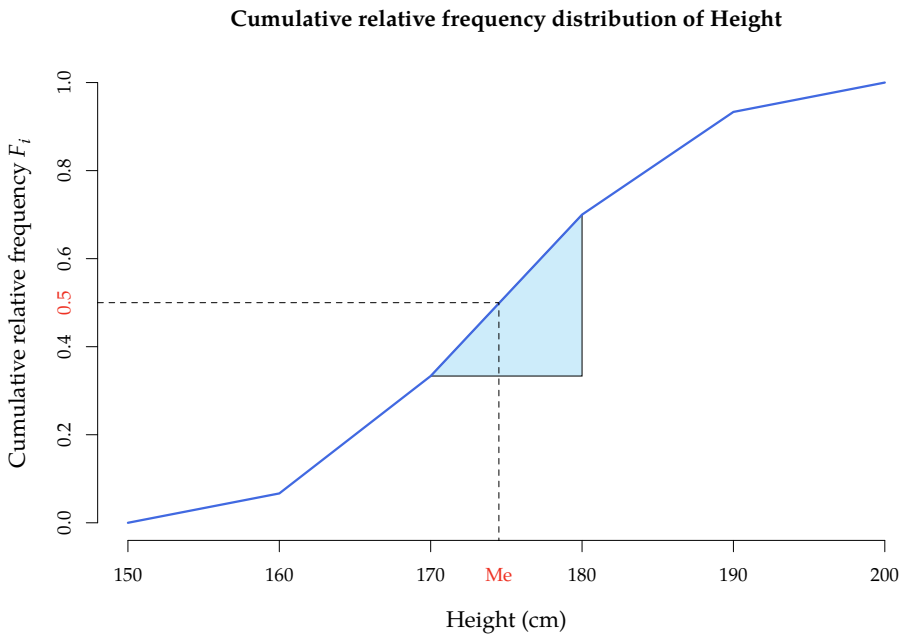
$$\tan(\alpha) = \frac{F_i - F_{i-1}}{l_i - l_{i-1}}$$

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

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

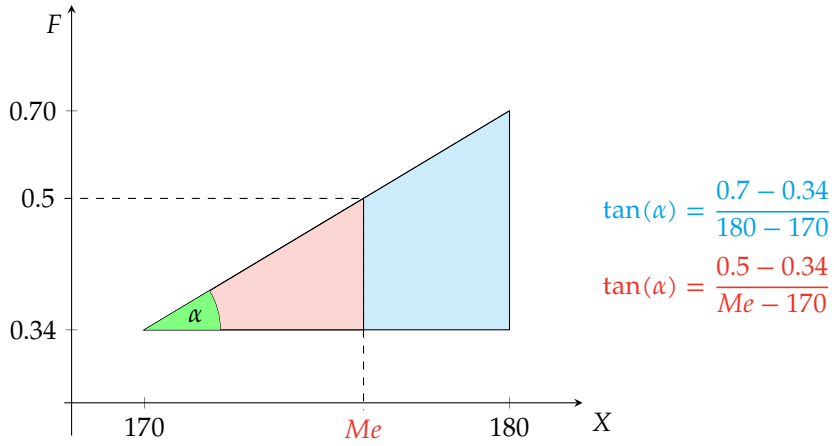
Median calculation for grouped data

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).

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. The mean. Mean takes more information from the sample than the others, as it takes into account the magnitude of data.
2. The median. Median takes less information than mean but more than mode, as it takes into account the order of data.
3. The 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 it 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.

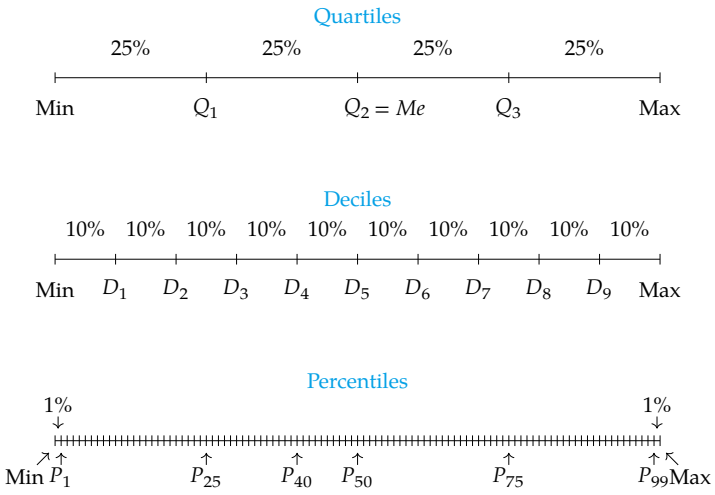
The most used are:

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).

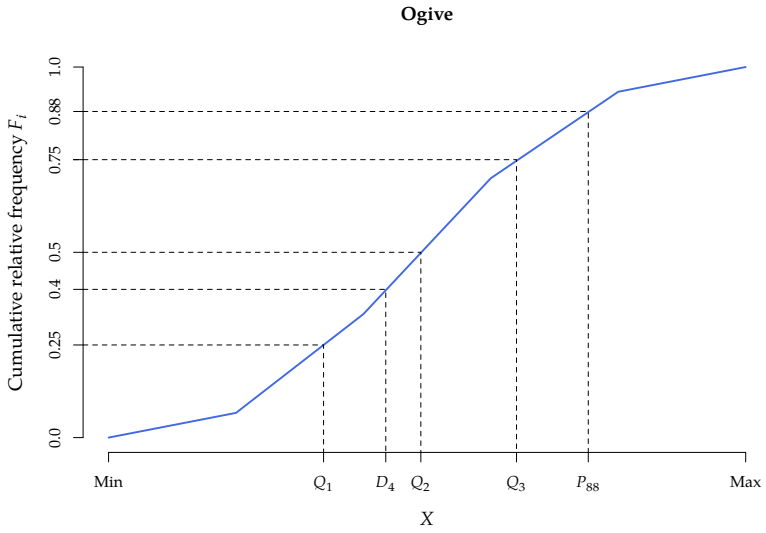
Quantiles



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

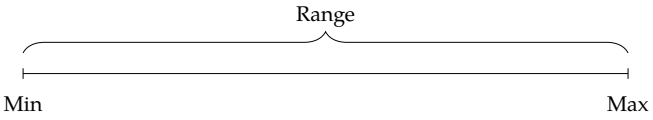
- Variance
- Standard deviation
- Coefficient of variation

Range

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

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

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

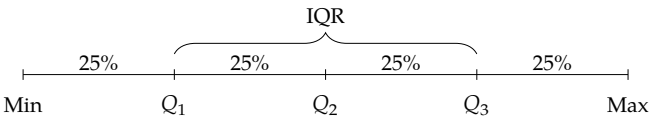


Interquartile range

The following measure avoids 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

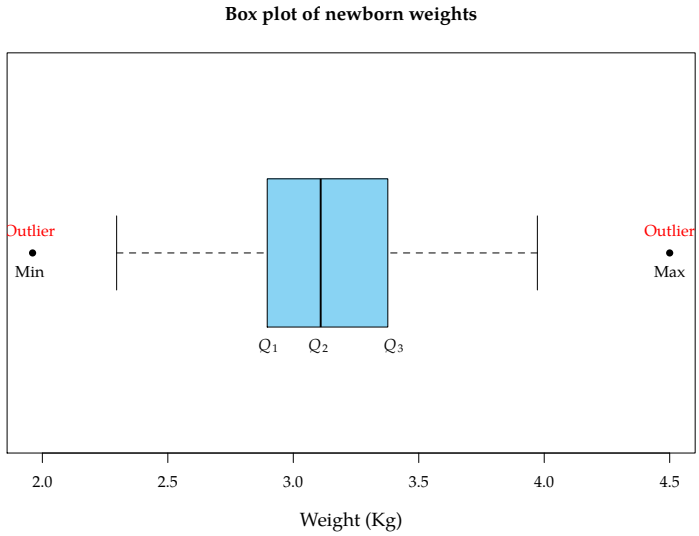
The dispersion of a variable in a sample can be graphically represented with a **box plot**, that represent five descriptive statistics (minimum, quartiles and maximum) known as the *five-numbers*. It consist in a box, drawn from the lower to the upper quartile, that represent the interquartile range,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

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

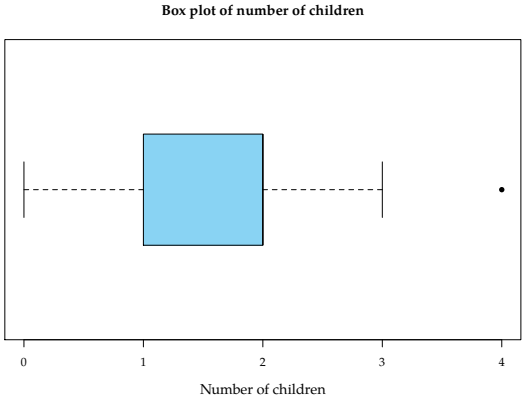
   The fences define the interval where data are considered normal. Any value outside that interval is considered an outlier. For the lower whisker draw a segment from the lower quartile to the lower value in the sample grater than or equal to f_1, and for the upper whisker draw a segment from the upper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

5. Finally, if there are outliers, draw a dot at every outlier.

Box plot construction

Example of number of children

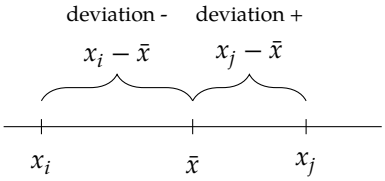
1. Calculate the quartiles: $Q_1 = 1$ children and $Q_2 = Q_3 = 2$ children
2. Draw the box.
3. Calculate the fences $f_1 = 1 - 1.5 * 1 = -0.5$ and $f_2 = 2 + 1.5 * 1 = 3.5$.
4. Draw the whiskers: $w_1 = 0$ children and $w_2 = 3$ children.
5. Draw the outliers: 4 children.



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 is 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 the 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 its interpretation it is 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.

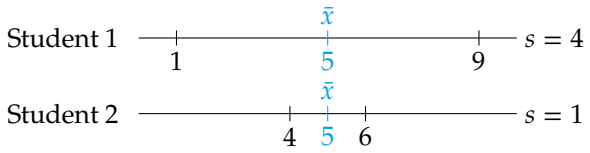
$$s = +\sqrt{s^2}$$

Variance and standard deviation interpretation

Both variance and standard deviation measure 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 they are big, the sample data are far from the mean, and the mean does not represent so well.

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 is 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 standard deviation and the absolute value of the mean of the sample.

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

The coefficient of variation measures the relative dispersion of data around the sample mean. The bigger the coefficient of variation, the bigger the relative dispersion with respect to the mean and the less representative the mean is. Since it has no units, the coefficient of variation it is very helpful to compare dispersion in distributions of different variables.

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 peakedness 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.

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

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

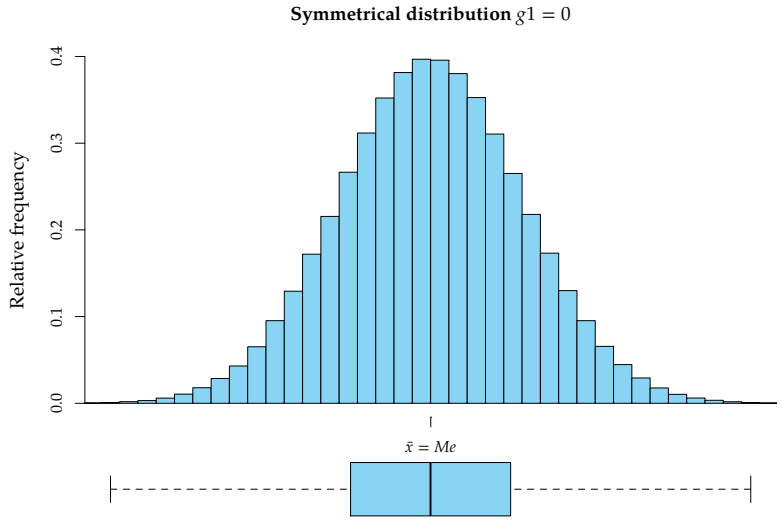
Interpretation of the coefficient of skewness

According to the sign of g_1 we have:

- $g_1 = 0$ 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 left-skewed (it has longer tail to the left).
- $g_1 > 0$ indicates that there are more values below the mean than above it, but the values above are further from it, and the distribution is right-skewed (it has longer tail to the right).

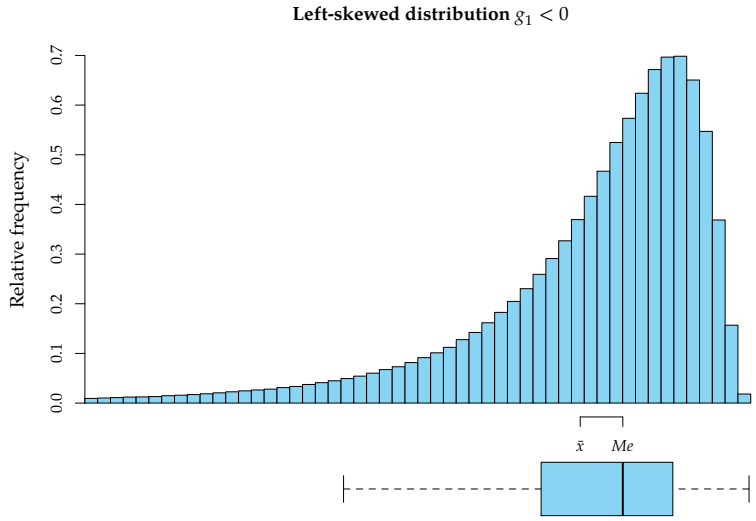
Coefficient of skewness

Example of symmetrical distribution



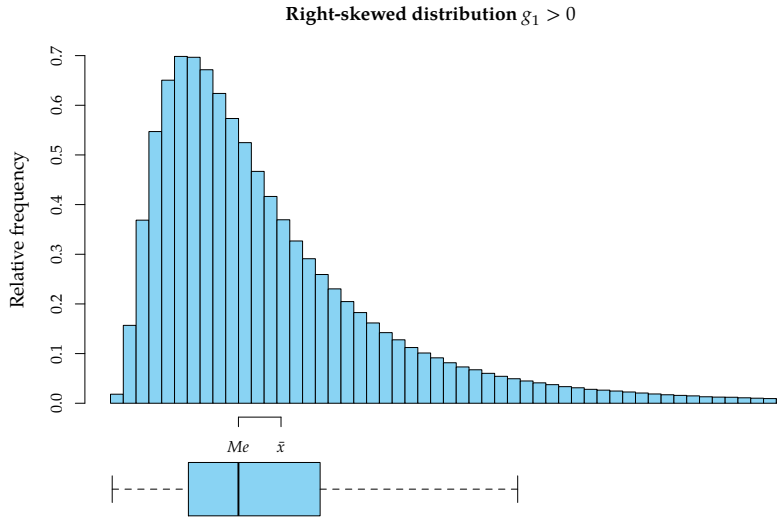
Coefficient of skewness

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 concentration of values around the mean and the length of the tails of the distribution. The normal (Gaussian bell-shaped) distribution is taken as a reference.

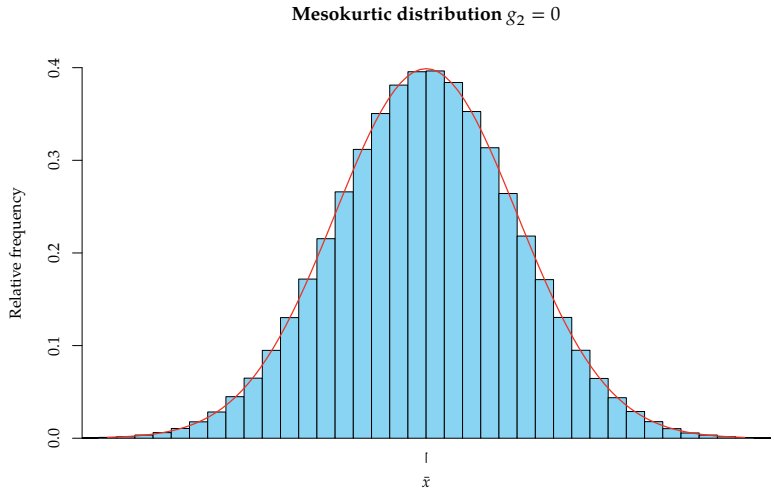
Coefficient of kurtosis

According to the sign fo g_2 we have:

- $g_2 = 0$ indicates that the kurtosis is normal, that is, the concentration of values around the mean is the same than in a Gaussian bell-shaped distribution (*mesokurtic*).
- $g_2 < 0$ indicates that the kurtosis is less than normal, that is, the concentration of values around the mean is less than in a Gaussian bell-shaped distribution (*platykurtic*).
- $g_2 > 0$ indicates that the kurtosis is greater than normal, that is, the concentration of values around the mean is greater than in a Gaussian bell-shaped distribution (*leptokurtic*).

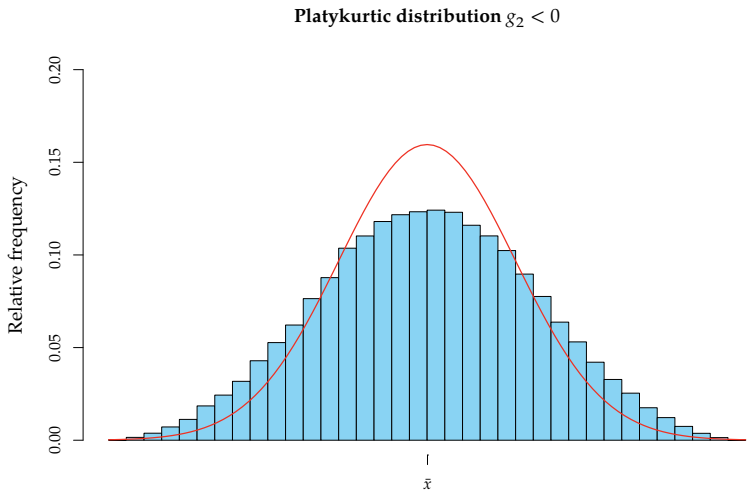
Coefficient of kurtosis

Example of mesokurtic distribution



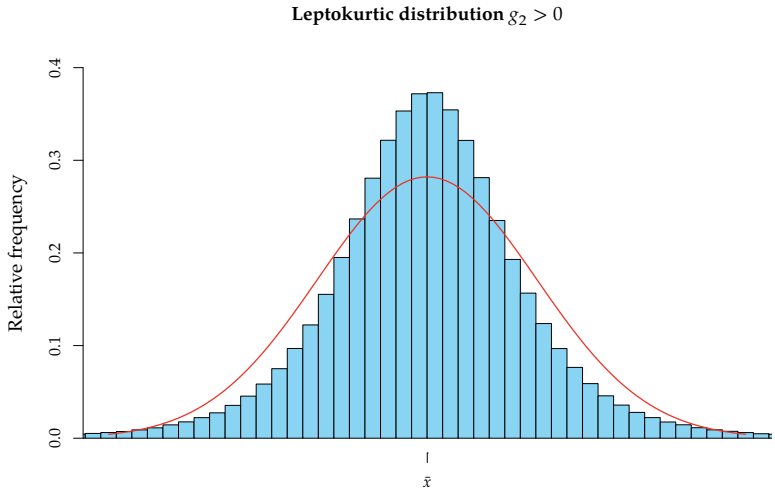
Coefficient of kurtosis

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.

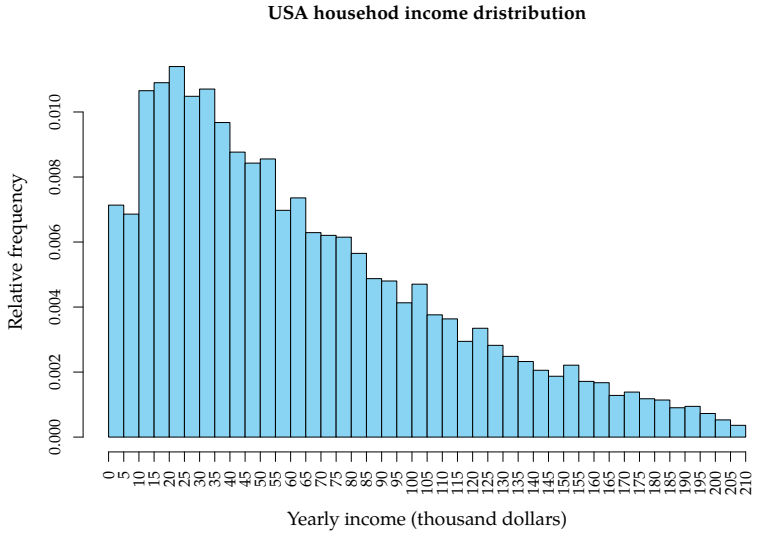
Normal distributions are symmetrical and mesokurtic, and therefore, their coefficients of symmetry and kurtosis are equal to 0. So, a way of checking if a sample comes from a normal population is looking how far are the coefficients of skewness and kurtosis from 0.

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.

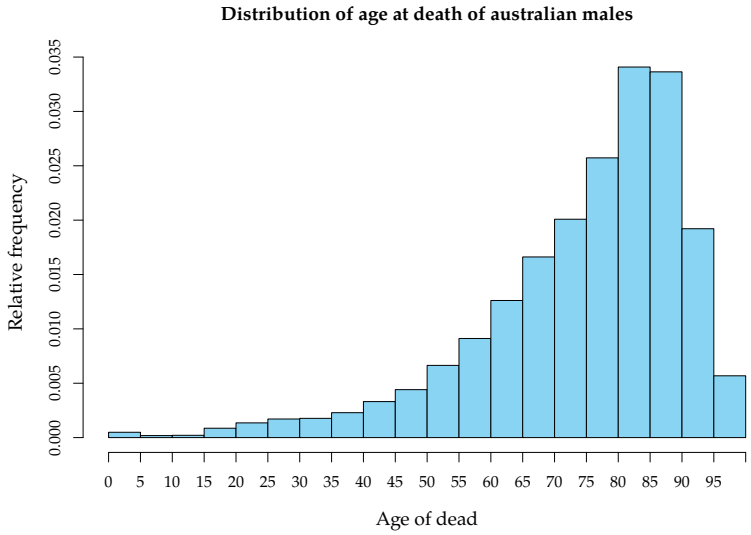
Non-normal right-skewed distribution

Household incomes



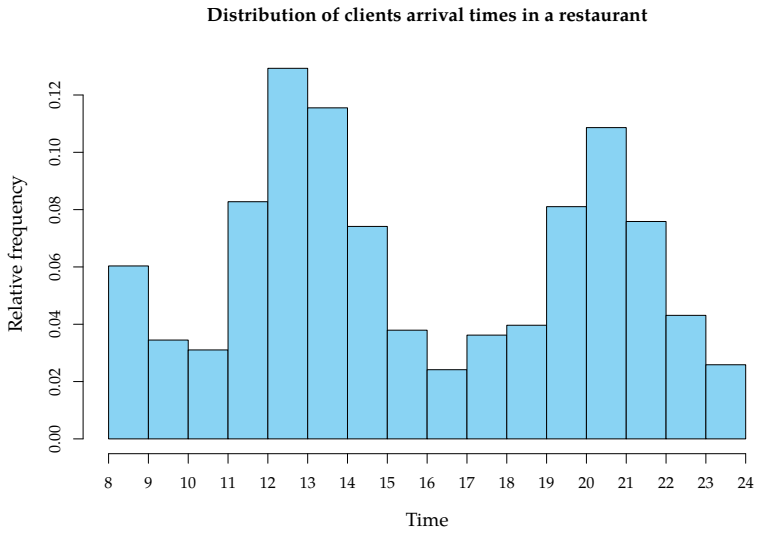
Non-normal left-skewed distribution

Age at death



Non-normal bimodal distribution

Arrival time of clients of a restaurant



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 have the following sample of heights in metres:

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

it is 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 is 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 is obvious that these data are easier to work with than the original ones. In essence, we have applied the following transformation to the data:

$$Y = 100X - 165$$

Linear transformations

One of the most common transformations is the *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 does not 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 results from subtracting the mean from X and dividing it 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 is useful to avoid the dependency of the variable from its measurement units.

The standardized variable always has 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 grade, 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 it 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 above the mean of X, while the student with an 8 in Y is only 0.95 times the standard deviation above 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

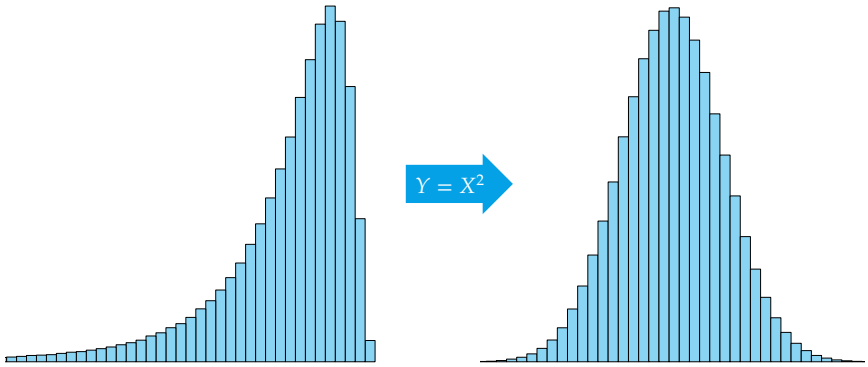
Student:	1	2	3	4	5
X :	-1.5	0	-0.5	1.5	0.5
Y :	-1.26	1.26	0.95	0	-0.95
$\sum$	-2.76	1.26	0.45	1.5	-0.45

the best student is the fourth one.

Non-linear transformations

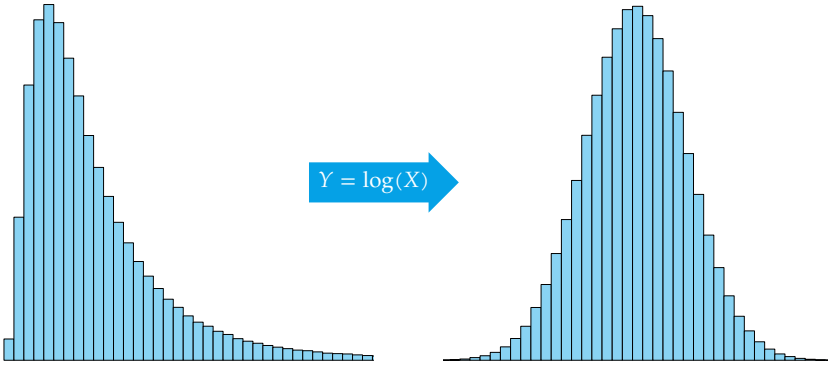
Non-linear transformations are also common to correct non-normality of distributions.

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



Non-linear transformation

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



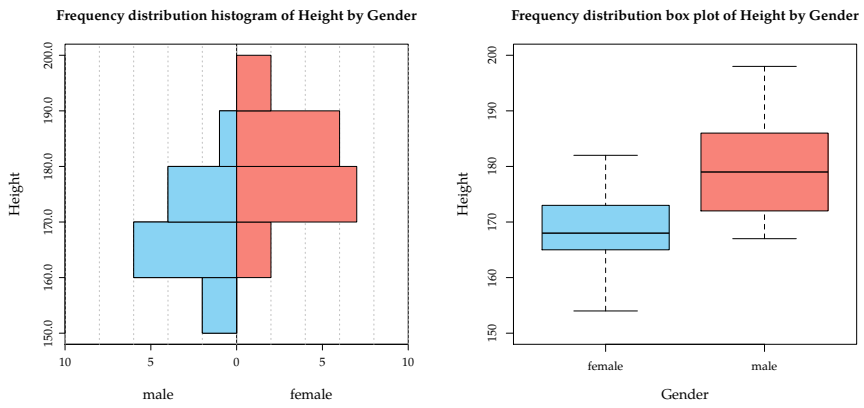
Factors

Sometimes it is interesting to describe the frequency distribution of the main variable for different subsamples corresponding to the categories of another variable, 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



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).

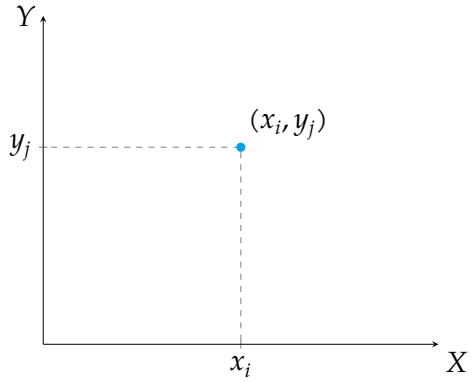
The joint frequency table is

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 collection of points on a XY coordinate system.

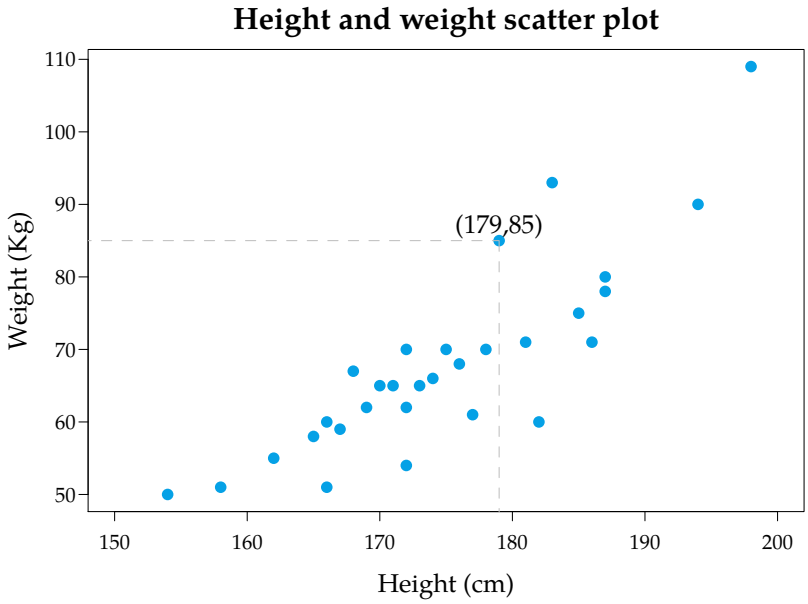
Usually the independent variable is represented in the X axis and the dependent variable in the Y axis. For every data pair (x_i, y_j) in the sample a dot is drawn on the plane with those coordinates.



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

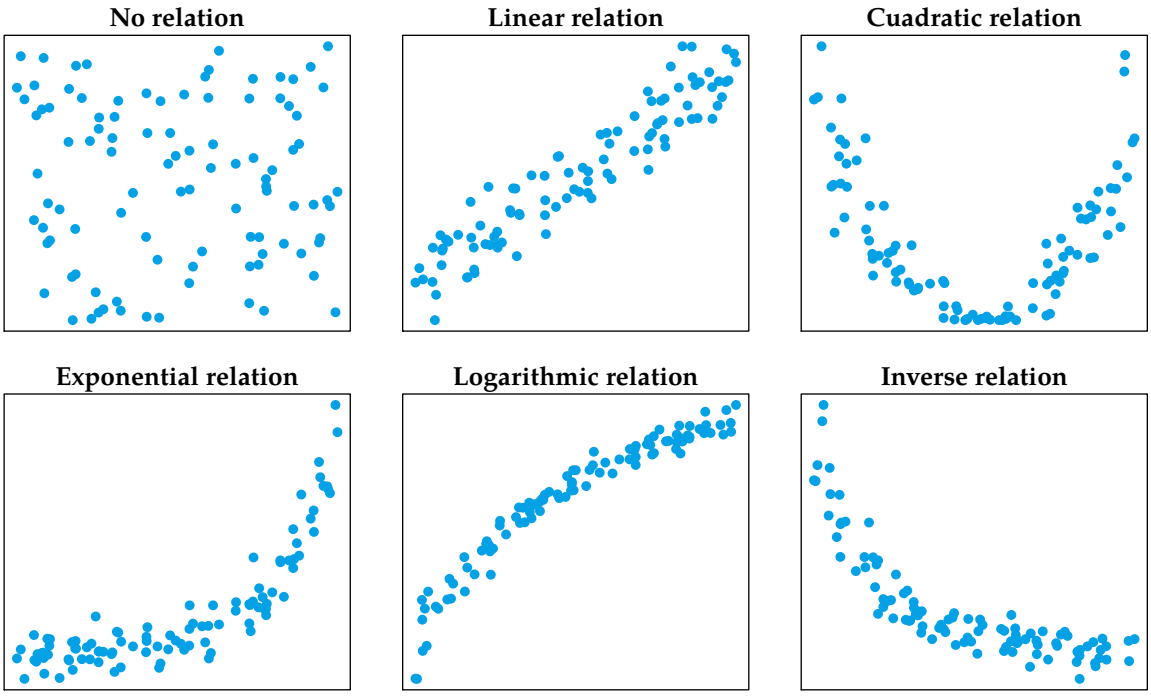
Scatter plot

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.

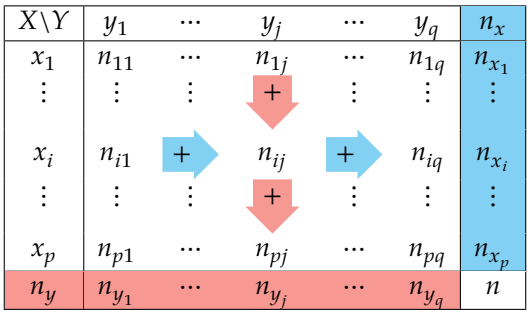


Marginal frequency distributions

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 by adding 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

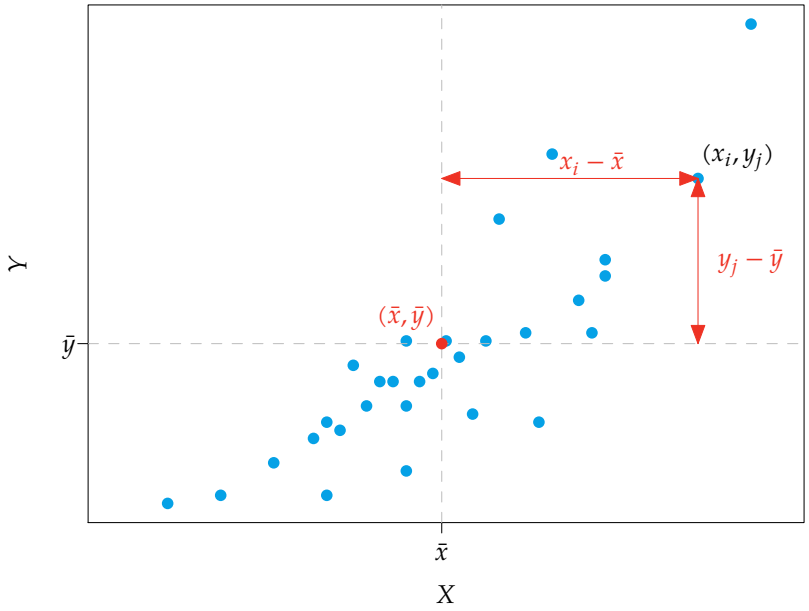
and the corresponding statistics are

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

4.2 Covariance

Deviations from the means

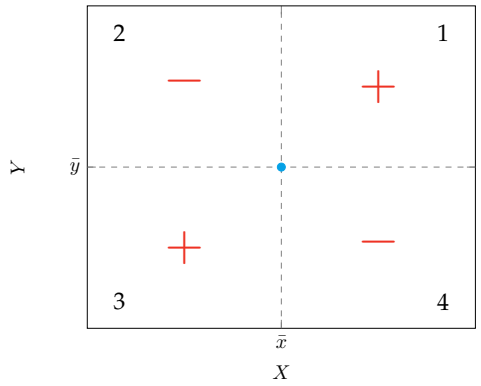
To study the relation between two variables, we have to analyze the joint variation of them.



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:

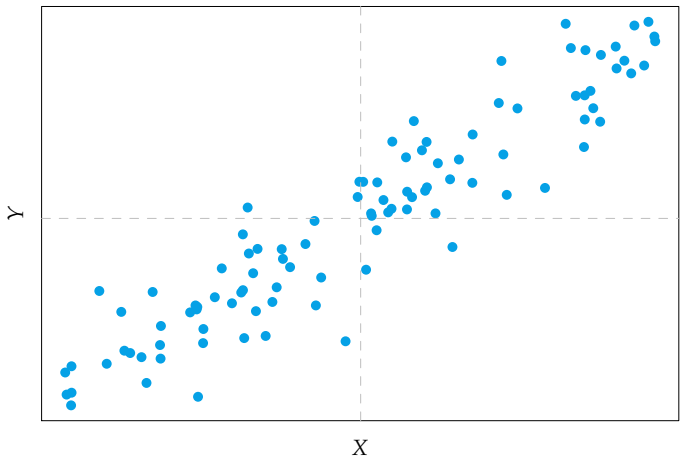
Quadrant	$(x_i - \bar{x})$	$(y_j - \bar{y})$	$(x_i - \bar{x})(y_j - \bar{y})$
1	+	+	+
2	−	+	−
3	−	−	+
4	+	−	−



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.

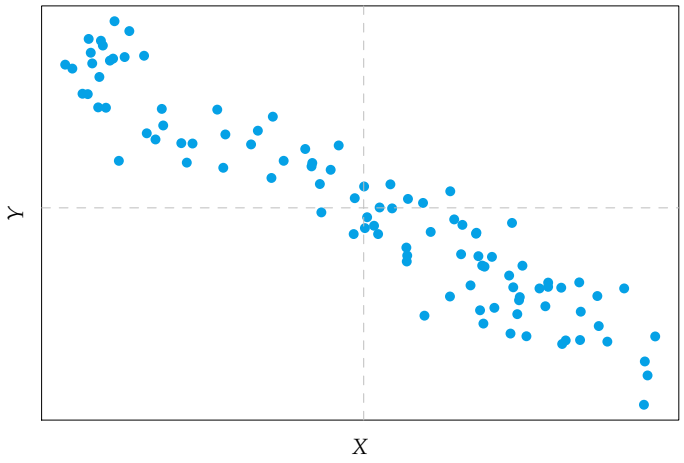
Increasing linear relation



$$\sum (x_i - \bar{x})(y_j - \bar{y}) > 0$$

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



$$\sum (x_i - \bar{x})(y_j - \bar{y}) < 0$$

Covariance

Using the products of deviations from the means 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

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

we get that the covariance is equal to

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

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 called **regression**, and the function is known as **regression function** or **regression model**.

Simple regression models

There are a lot of types of regression models. The most common models 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

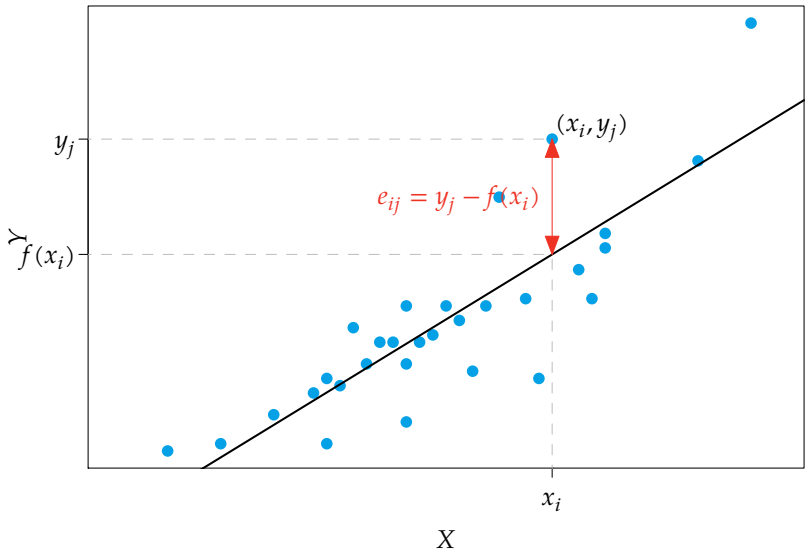
Once chosen the type of regression model, we have to determine which function of that family explains better the relation between the dependent and the independent variables, that is, the function that predicts better the dependent variable.

That function is the function that minimizes 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 determines the function that

minimizes 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 minimizes the predictive errors on Y, therefore it 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

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

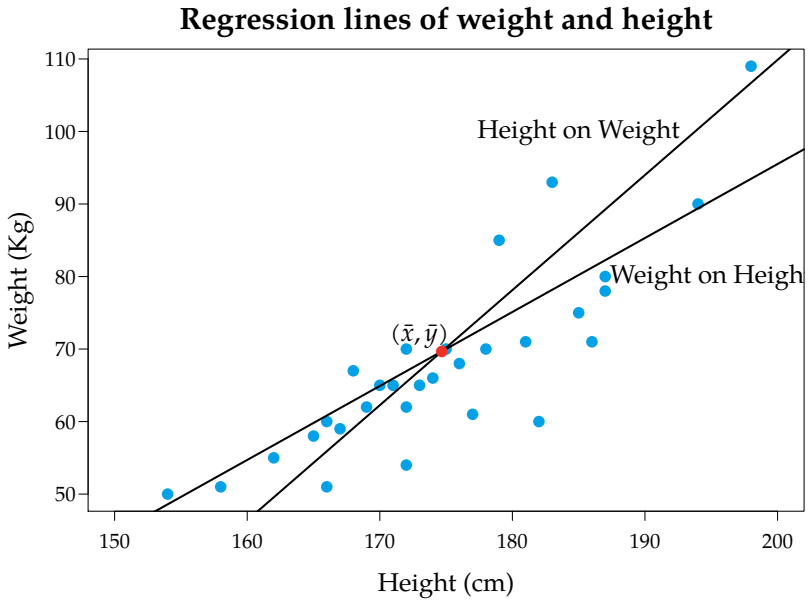
And the regression line of height on weight is

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

Observe that the regression lines are different!

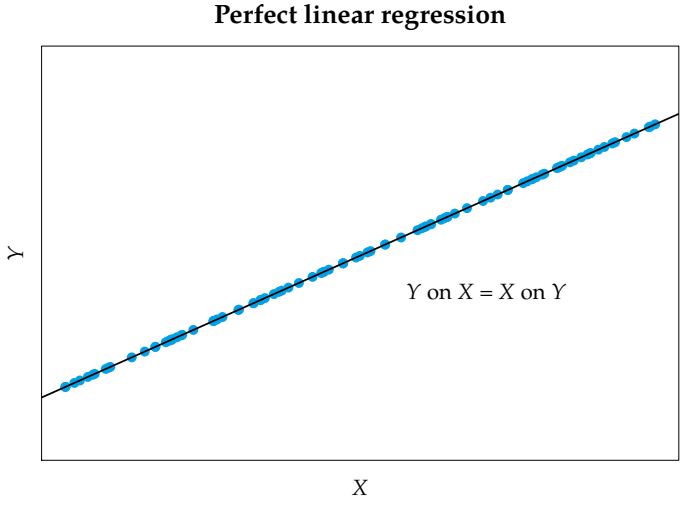
Regression lines

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.

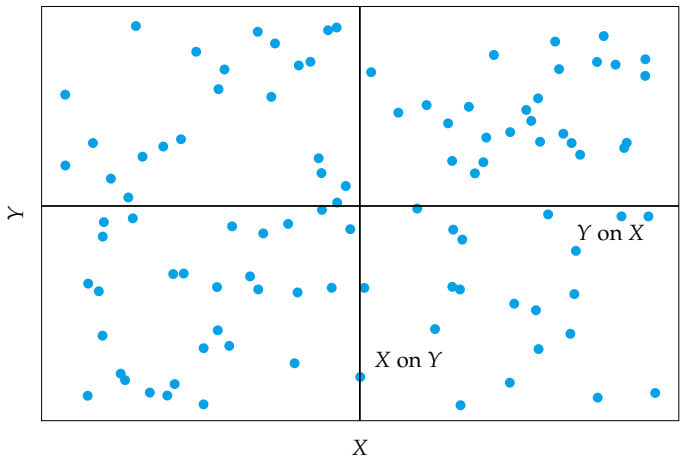


If there is no linear relation between the variables, then both regression lines are constant and equals to the respective means,

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

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 as 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 the independent variable increases.

Regression coefficient

Example of heights and weights

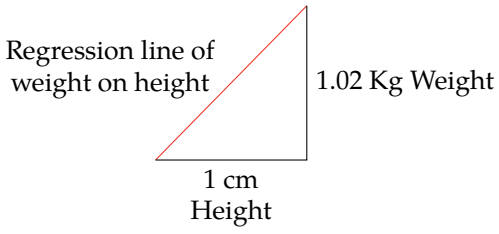
In the sample of heights and weights, the regression line of weight on height was

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

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 the height increases.



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 $y = f(x)$ of a two-dimensional variable (X, Y), its *sample residual variance* is the average of the squared residuals,

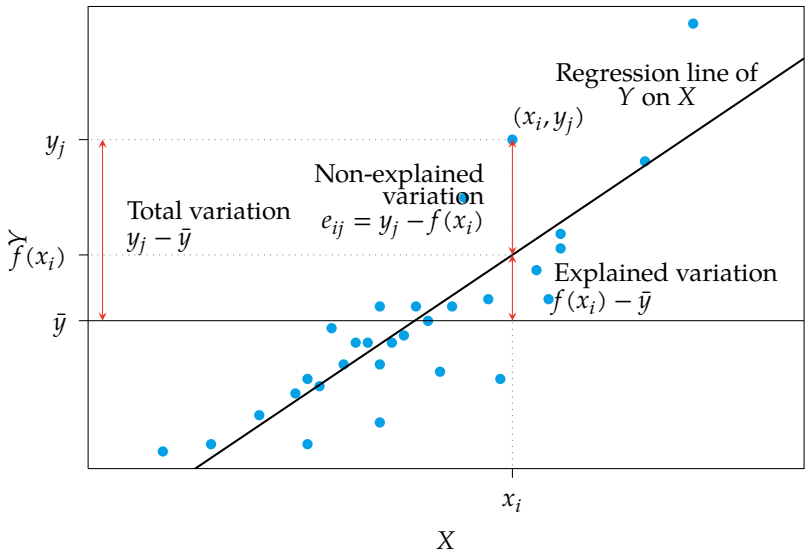
$$s^2_{ry} = \frac{\sum e^2_{ij} n_{ij}}{n} = \frac{\sum (y_j - f(x_i))^2 n_{ij}}{n}.$$

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 equal to 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, we have

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

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

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

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,

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

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

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

The correlation coefficient measures not only 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

The example of heights and weights

In the sample of heights and weights, we had

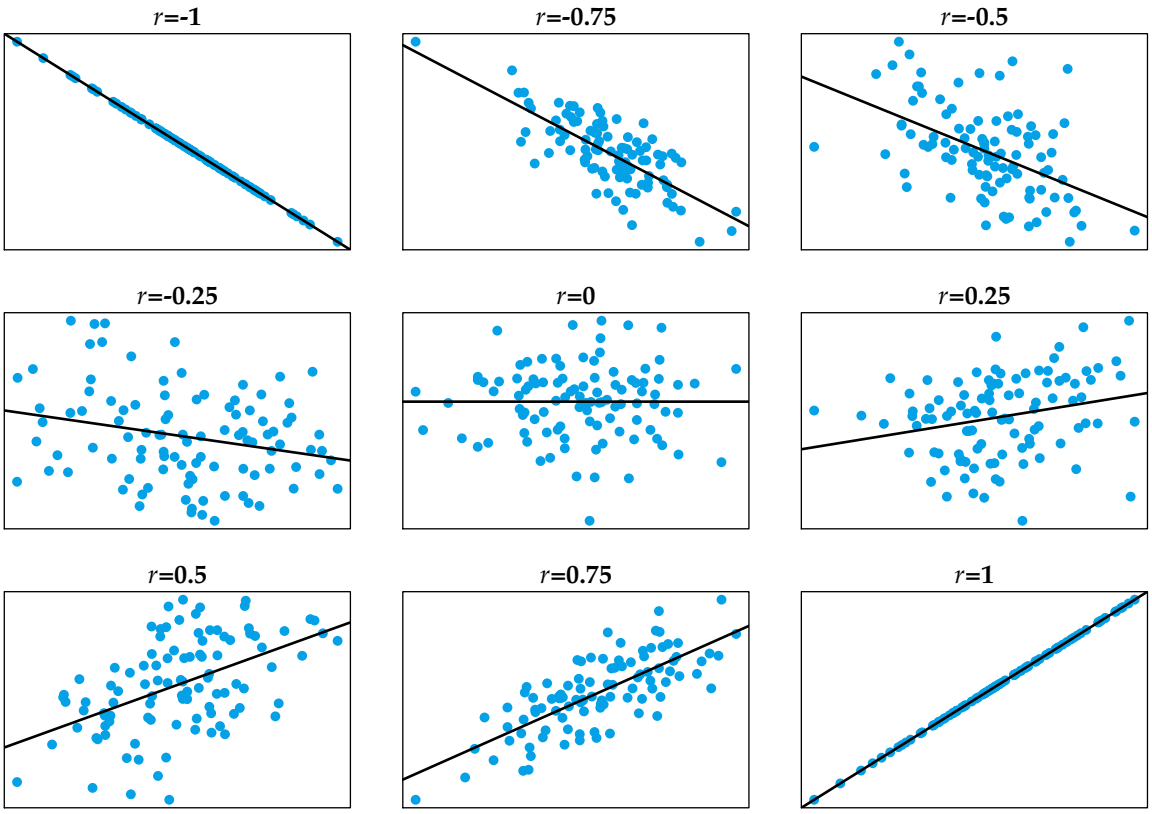
$$\begin{array}{ll} \bar{x} = 174.67 \text{ cm} & s_x^2 = 102.06 \text{ cm}^2 \\ \bar{y} = 69.67 \text{ Kg} & s_y^2 = 164.42 \text{ Kg}^2 \\ s_{xy} = 104.07 \text{ cm} \cdot \text{ Kg} & \end{array}$$

Thus, the correlation coefficient is

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

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 are.
- The variability of the population distribution: The greater the variation, the more difficult to predict and the less reliable the predictions are.
- The sample size: The greater the sample size, the more information we have and the more reliable the predictions are.

In addition, we have to take into account that a regression model is only valid for the range of values observed in 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 the 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 $z = \log y$:
$$z = \log y = \log(e^{a+b/x}) = a + b(1/x) = a + bt.$$

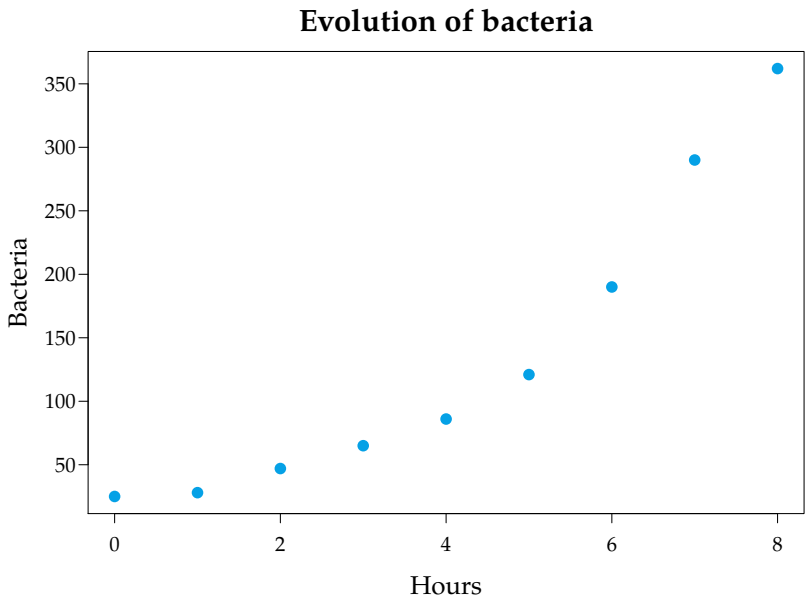
Exponential relation

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.



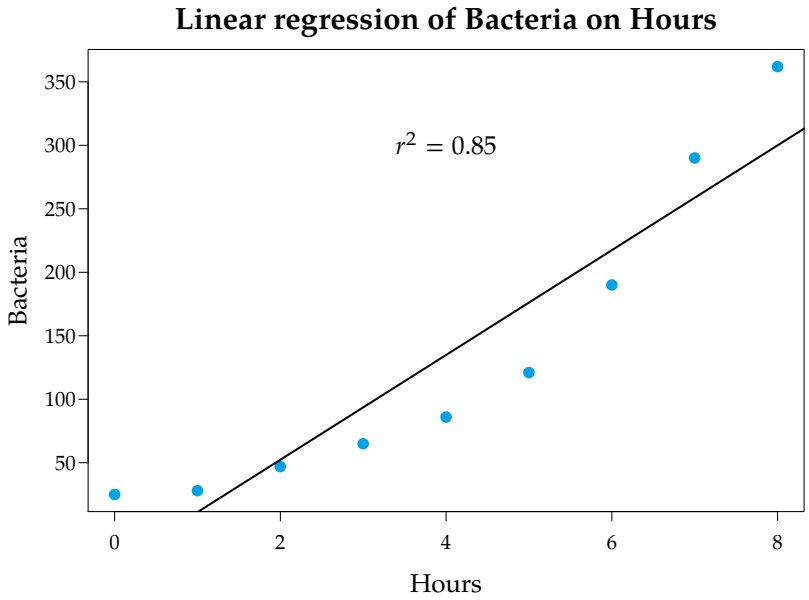
Exponential relation

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?

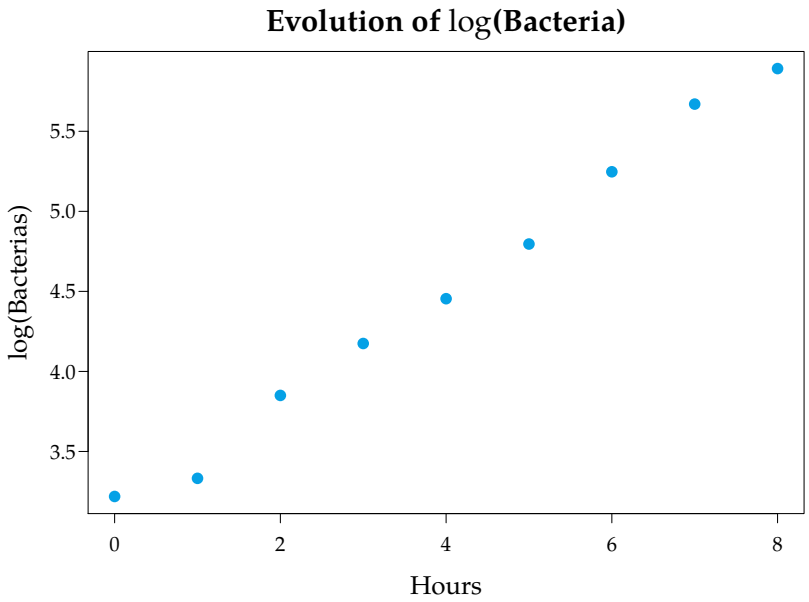
Fitting an exponential regression model

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,

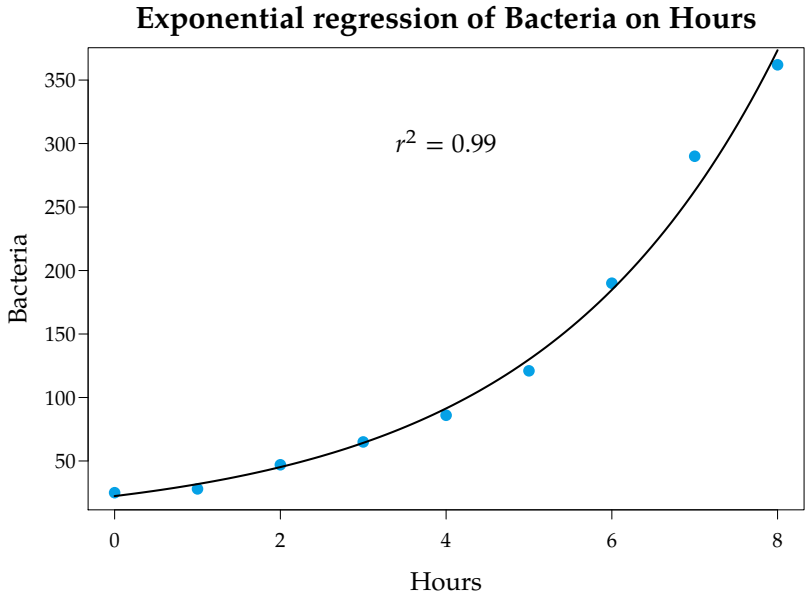
$$\log(\text{Bacteria}) = 3.107 + 0.352\,\text{Horas},$$

and, undoing the change of variable,

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

with $r^2 = 0.99$.

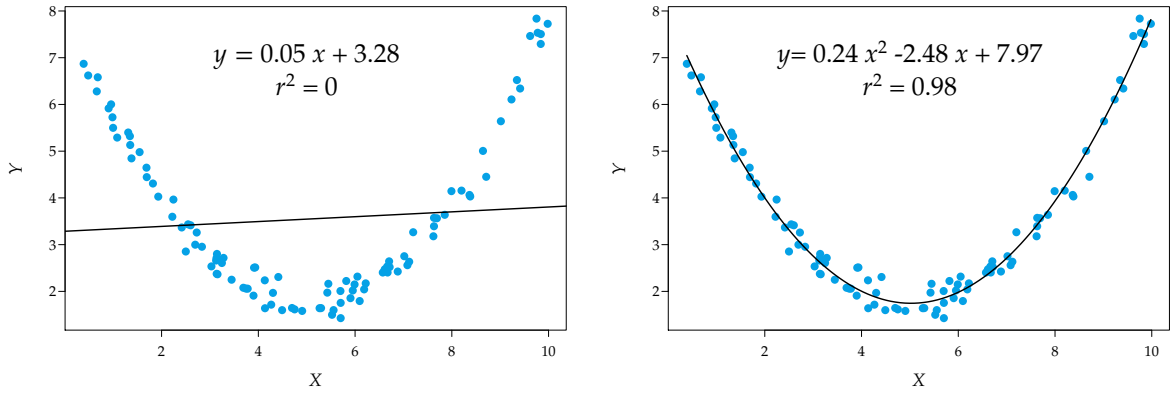
Thus, the exponential model fits much better than the linear model.



4.8 Regression risks

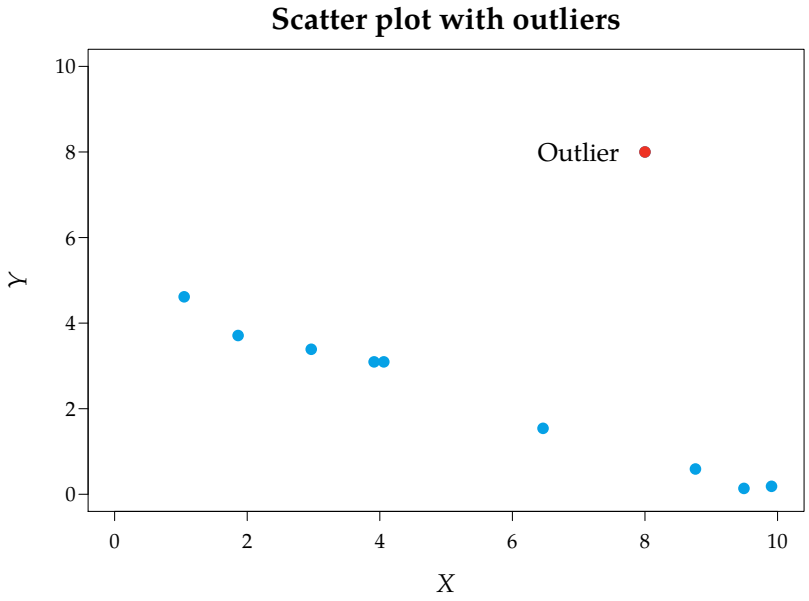
Lack of fit doesn't mean independence

It is 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 does not mean that the variables are independent*, because there could be a different type of relation.



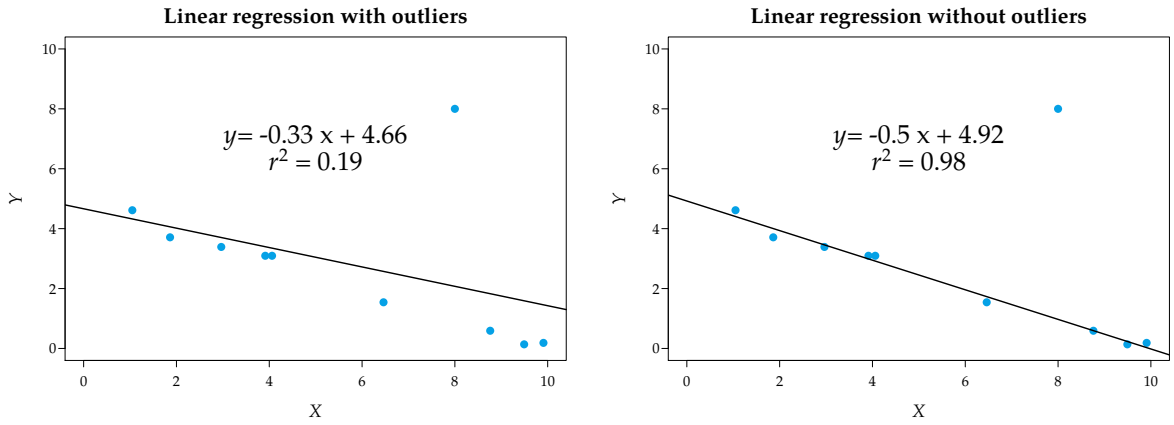
Outliers in regression

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



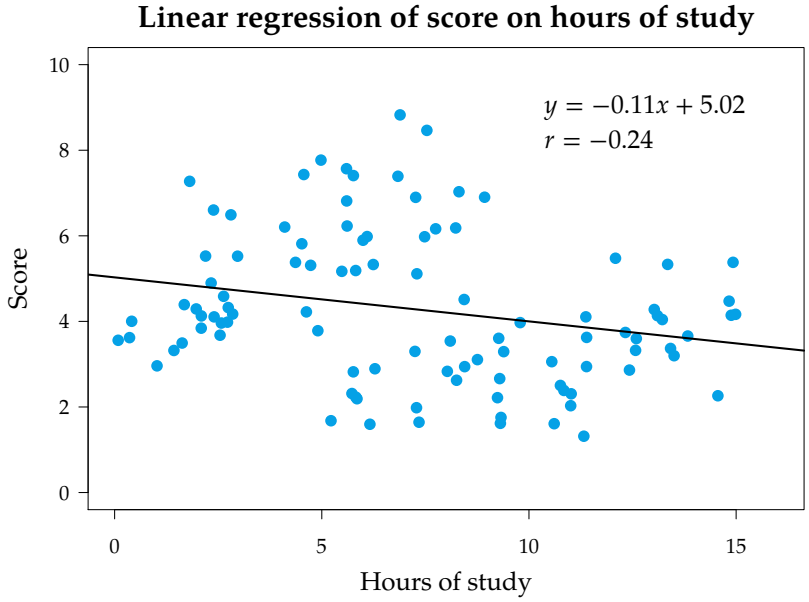
Outliers influence in regression

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



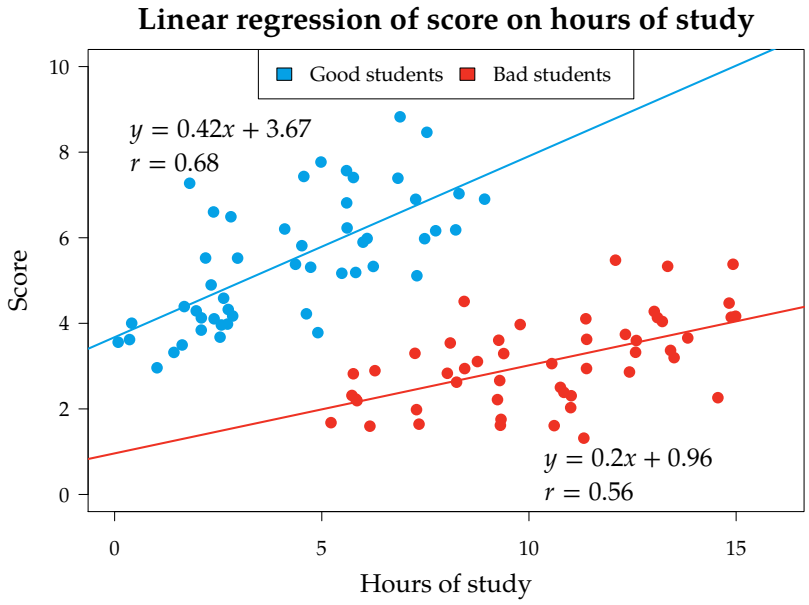
Simpson's paradox

Sometimes a trend can disappears or even reverses when we split the sample into groups according to a qualitative variable that is related to the dependent variable. This is known as the *Simpson's paradox*.



Simpson's paradox

Sometimes a trend can disappears or even reverses when we split the sample into groups according to a qualitative variable that is related to the dependent variable. This is known as the *Simpson's paradox*.


Linear regression of score on hours of study
Good students
Bad students
$y = 0.42x + 3.67$
$r = 0.68$
$y = 0.2x + 0.96$
$r = 0.56$
Score
Hours of study
0
2
4
6
8
10
0
5
10
15

5 Probability

Introduction

Descriptive Statistics provides methods to describe the variables measured in the sample and their relations, but it does not allow to draw any conclusion about the population.

Now it is time to take the leap from the sample to the population and the bridge for that is **probability theory**.

Remember that the sample has a limited information about the population, and in order to draw valid conclusions for the population the sample must be representative of it. For that reason, to guarantee the representativeness of the sample, this must be drawn randomly. This means that the choice of individuals in the sample is by chance.

Probability theory will provide us the tools to control the random in the sampling and to determine the level of reliability of the conclusions drawn from the sample.

5.1 Random experiments and events

Random experiments

The study of a characteristic of the population is conducted through random experiments.

Definition 29 (Random experiment)**.** A *random experiment* is an experiment that meets two conditions:

1. The set of possible outcomes is known.
2. It is impossible to predict the outcome with absolute certainty.

Example. Gambling are typical examples of random experiments. The roll of a dice, for example, is a random experiment because

1. It is known the set of possible outcomes: $\{1, 2, 3, 4, 5, 6\}$.
2. Before rolling the dice, it is impossible to predict with absolute certainty the outcome.

Another non-gambling example is the random choice of an individual of a human population and the determination of its blood type.

Generally, the draw of a sample by a random method is an random experiment.

Sample space

Definition 30 (Sample space)**.** The set Ω of the possible outcomes of a random experiment is known as the *sample space*.

Example Some examples of sample spaces are:

- For the toss of a coin $\Omega = \{heads, tails\}$.
- For the roll of a dice $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- For the blood type of an individual drawn by chance $\Omega = \{\mathrm{A}, \mathrm{B}, \mathrm{AB}, 0\}$.
- For the height of an individual drawn by chance $\Omega = \mathbb{R}^+$.

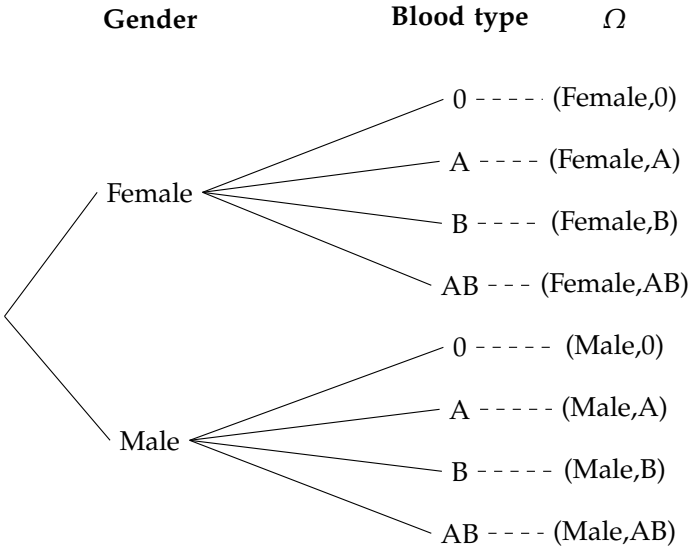
Tree diagrams

In experiments where more than one variable is measured, the determination of the sample space can be difficult. In such a cases, it is advisable to use a **tree diagram** to construct the sample space.

In a tree diagram every variable is represented in a level of the tree and every possible outcome of the variable as a branch.

Tree diagram

Example of gender and blood type



Random events

Definition 31 (Random event)**.** A *random event* is any subset of the sample space Ω of a random experiment.

There are different types of events:

- **Impossible event:** Is the event with no elements $\emptyset$. It has no chance of occurring.
- **Elemental events:** Are events with only one element, that is, a singleton.
- **Composed events:** Are events with two or more elements.
- **Sure event:** Is the event that contains the whole sample space Ω. It always happens.

5.2 Set theory

Event space

Definition 32 (Event space)**.** Given a sample space Ω of a random experiment, the *event space* of Ω is the set of all possible events of Ω, and is noted $\mathcal{P}(\Omega)$.

Example. Given the sample space $\Omega = \{a, b, c\}$, its even space is

$$\mathcal{P}(\Omega) = \big\{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\big\}$$

Event operations

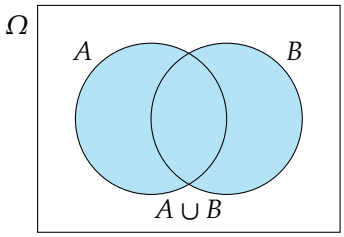
As events are subsets of the sample space, using the set theory we have the following operations on events:

- Union
- Intersection
- Complement
- Difference

Union of events

Definition 33 (Union event)**.** Given two events $A, B \subseteq \Omega$, the *union* of A and B, denoted by $A \cup B$, is the event of all elements that are members of A or B or both.

$$A \cup B = \{x \,|\, x \in A \text{ or } x \in B\}.$$

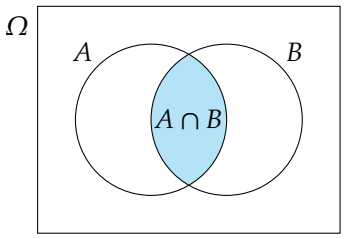


The union event $A \cup B$ happens when *A or B* happen.

Intersection of events

Definition 34 (Intersection event)**.** Given two events $A, B \subseteq \Omega$, the *intersection* of A and B, denoted by $A \cap B$, is the event of all elements that are members of both A and B.

$$A \cap B = \{x \,|\, x \in A \text{ and } x \in B\}.$$



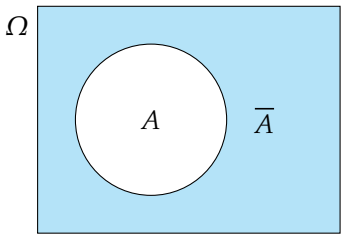
The intersection event $A \cap B$ happens when *A and B* happen.

Two events are **incompatible** if their intersection is empty.

Complement of an event

Definition 35 (Complementary event)**.** Given an event $A \subseteq \Omega$, the *complementary or contrary event* of A, denoted by $\overline{A}$, is the event of all elements of Ω except the elements that are members of A.

$$\overline{A} = \{x \,|\, x \notin A\}.$$

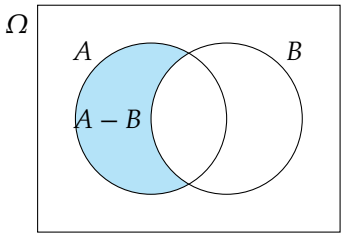


The complementary event $\overline{A}$ happens when A does *not* happen.

Difference of events

Definition 36 (Difference event)**.** Given two events $A, B \subseteq \Omega$, the *difference* of A and B, denoted by $A - B$, is the event of all elements that are members of A but not are members of B.

$$A - B = \{x \,|\, x \in A \text{ and } x \notin B\} = A \cap \overline{B}.$$



The difference event $A - B$ happens when A happens but B does not.

Event operations

Example

Given the sample space of rolling a dice $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the events $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$,

- The union of A and B is $A \cup B = \{1, 2, 3, 4, 6\}$.
- The intersection of A and B is $A \cap B = \{2, 4\}$.
- The complement of A is $\overline{A} = \{1, 3, 5\}$.
- The events A and $\overline{A}$ are incompatible.
- The difference of A and B is $A - B = \{6\}$, and the difference of B and A is $B - A = \{1, 3\}$.

Algebra of events

Given the events $A, B, C \subseteq \Omega$, the following properties are meet:

1. $A \cup A = A, \quad A \cap A = A$ (idempotence).
2. $A \cup B = B \cup A, \quad A \cap B = B \cap A$ (commutative).
3. $(A \cup B) \cup C = A \cup (B \cup C), \quad (A \cap B) \cap C = A \cap (B \cap C)$ (associative).
4. $(A \cup B) \cap C = (A \cap C) \cup (B \cap C), \quad (A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ (distributive).
5. $A \cup \emptyset = A, \quad A \cap \Omega = A$ (neutral element).

6. $A \cup \Omega = \Omega, \quad A \cap \emptyset = \emptyset$ (absorbing element).
7. $A \cup \overline{A} = \Omega, \quad A \cap \overline{A} = \emptyset$ (complementary symmetric element).
8. $\overline{\overline{A}} = A$ (double contrary).
9. $\overline{A \cup B} = \overline{A} \cap \overline{B}, \quad \overline{A \cap B} = \overline{A} \cup \overline{B}$ (Morgan's laws).

5.3 Probability definition

Classical definition of probability

Definition 37 (Probability — Laplace)**.** Given a sample space Ω of a random experiment where all elements of Ω are equally likely, the *probability* of an event $A \subseteq \Omega$ is the quotient between the number of elements of A and the number of elements of Ω

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{number of favorable outcomes}}{\text{number of possible outcomes}}$$

This definition is well known, but it has important restrictions:

- It is required that all the elements of the sample space are equally likely (*equiprobability*).
- It can not be used with infinite sample spaces.

Watch out! These conditions are not meet in many real experiments.

Classical definition of probability

Example

Given the sample space of rolling a dice $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the event $A = \{2, 4, 6\}$, the probability of A is

$$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = 0.5.$$

However, given the sample space of the blood type of a random individual $\Omega = \{O, A, B, AB\}$, it is not possible to use the classical definition to compute the probability of having group A,

$$P(A) \neq \frac{|A|}{|\Omega|} = \frac{1}{4} = 0.25,$$

because the blood types are not equally likely in human populations.

Frequency definition of probability

Theorem 38 (Law of large numbers)**.** *When a random experiment is repeated a large number of times, the relative frequency of an event tends to the probability of the event.*

The following definition of probability uses this theorem.

Definition 39 (Frequency probability)**.** Given a sample space Ω of a replicable random experiment, the *probability* of an event $A \subseteq \Omega$ is the relative frequency of the event A in an infinite number of repetitions of the experiment

$$P(A) = lim_{n \to \infty} \frac{n_A}{n}$$

This definition also have some drawbacks

- It computes an estimation of the real probability.
- The repetition of the experiment must be in identical conditions.

Frequency definition of probability

Example

Given the sample space of tossing a coin $\Omega = \{H, T\}$, if after tossing the coin 100 times we got 54 heads, then the probability of H is

$$P(H) = \frac{n_H}{n} = \frac{54}{100} = 0.54.$$

Given the sample space of the blood type of a random individual $\Omega = \{O, A, B, AB\}$, if after drawing a random sample of 1000 persons we got 412 with blood type A, then the probability of A is

$$P(A) = \frac{n_A}{n} = \frac{412}{1000} = 0.412.$$

Axiomatic definition of probability

Definition 40 (Probability — Kolmogórov)**.** Given a sample space Ω of a random experiment, a *probability* function is a function that maps every event $A \subseteq \Omega$ a real number $P(A)$, known as the probability of A, that meets the following axioms:

1. The probability of any event is nonnegative,
$$P(A) \geq 0.$$
2. The probability of the sure event is 1,
$$P(\Omega) = 1$$
3. The probability of the union of two incompatible events ($A \cap B = \emptyset$) is the sum of their probabilities
$$P(A \cup B) = P(A) + P(B).$$

Properties of the axiomatic probability

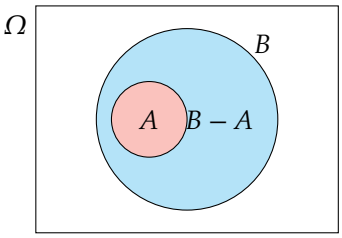
From the previous axioms is possible to deduce some important properties of a probability function.

Given a sample space Ω of a random experiment and the events $A, B \subseteq \Omega$, the following properties are meet:

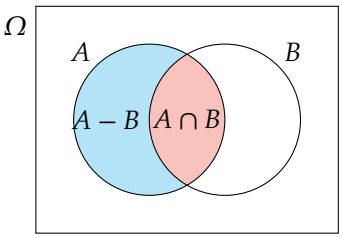
1. $P(\overline{A}) = 1 - P(A)$.
2. $P(\emptyset) = 0$.
3. If $A \subseteq B$ then $P(A) \leq P(B)$.
4. $P(A) \leq 1$. This means that $P(A) \in [0, 1]$.
5. $P(A - B) = P(A) - P(A \cap B)$.
6. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
7. If $A = \{e_1, \dots, e_n\}$, where e_i $i = 1, \dots, n$ are elemental events, then
$$P(A) = \sum_{i=1}^{n} P(e_i).$$

Proof.

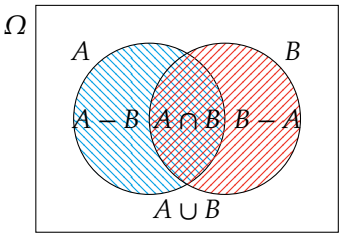
1. $\overline{A} = \Omega \Rightarrow P(A \cup \overline{A}) = P(\Omega) \Rightarrow P(A) + P(\overline{A}) = 1 \Rightarrow P(\overline{A}) = 1 - P(A)$.
2. $\emptyset = \overline{\Omega} \Rightarrow P(\emptyset) = P(\overline{\Omega}) = 1 - P(\Omega) = 1 - 1 = 0$.
3. $B = A \cup (B - A)$. As A and $B - A$ are incompatible, $P(B) = P(A \cup (B - A)) = P(A) + P(B - A) \geq P(A)$.
   If we think of probabilities as areas, it is easy to see graphically,



4. $A \subseteq \Omega \Rightarrow P(A) \leq P(\Omega) = 1$.
5. $A = (A - B) \cup (A \cap B)$. As $A - B$ and $A \cap B$ are incompatible, $P(A) = P(A - B) + P(A \cap B) \Rightarrow P(A - B) = P(A) - P(A \cap B)$.
   If we think of probabilities as areas, it is easy to see graphically,



6. $A \cup B = (A - B) \cup (B - A) \cup (A \cap B)$. As $A - B$, $B - A$ and $A \cap B$ are incompatible, $P(A \cup B) = P(A - B) + P(B - A) + P(A \cap B) = P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) = P(A) + P(B) - P(A \cup B)$.
   If we think again of probabilities as areas, it is easy to see graphically because the area of $A \cap B$ is added twice (one for A and other for B), so it must be subtracted once.



7. $A = \{e_1, \cdots, e_n\} = \{e_1\} \cup \cdots \cup \{e_n\} \Rightarrow P(A) = P(\{e_1\} \cup \cdots \cup \{e_n\}) = P(\{e_1\}) + \cdots P(\{e_n\})$.

Probability interpretation

As set by the previous axioms, the probability of an event A, is a real number $P(A)$ that always ranges from 0 to 1.

In a certain way, this number expresses the plausibility of the event, that is, the chances that the event A occurs in the experiment. Therefore, it also gives a measure of the uncertainty about the event.

- The maximum uncertainty correspond to probability $P(A) = 0.5$ (A and $\overline{A}$ have the same chances of happening).

- The minimum uncertainty correspond to probability $P(A) = 1$ (A will happen with absolute certainty) and $P(A) = 0$ (A won't happen with absolute certainty).

When $P(A)$ is closer to 0 than to 1, the chances of not happening A are greater than the chances of happening A. On the contrary, when $P(A)$ is closer to 1 than to 0, the chances of happening A are greater than the chances of not happening A.

5.4 Conditional probability

Conditional experiments

Occasionally, we can get some information about the experiment before its realization. Usually that information is given as an event B of the same sample space that we know that is true before we conduct the experiment.

In such a case, we will say that B is a *conditioning* event and the probability of another event A is known as a **conditional probability** and expressed

$$P(A|B).$$

This must be read as *probability of A given B* or *probability of A under the condition B*.

Conditional experiments

Example

Usually, conditioning events change the sample space and therefore the probabilities of events.

Assume that we have a sample of 100 women and 100 men with the following frequencies

	Non-smokers	Smokers
Females	80	20
Males	60	40

Then, using the frequency definition of probability, the probability of being smoker from the whole sample is

$$P(\text{Smoker}) = \frac{60}{200} = 0.3.$$

However, if we know that the person is a woman, then the sample is reduced to the first row, and the probability of being smoker is

$$P(\text{Smoker}|\text{Female}) = \frac{20}{100} = 0.2.$$

Conditional probability

Definition 41 (Conditional probability)**.** Given a sample space Ω of a random experiment, and two events $A, B \subseteq \Omega$, the probability of A *conditional* on B occurring is

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

as long as, $P(B) \neq 0$.

This definition allows to calculate conditional probabilities without changing the original sample space.

Example. In the previous example

$$P(\text{Smoker}|\text{Female}) = \frac{P(\text{Smoker} \cap \text{Female})}{P(\text{Female})} = \frac{20/200}{100/200} = \frac{20}{100} = 0.2.$$

Probability of the intersection event

From the definition of conditional probability it is possible to derive the formula for the probability of the intersection of two events.

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B).$$

Example. In a population there are a 30% of smokers and we know that there are a 40% of smokers with breast cancer. The probability of a random person being smoker and having breast cancer is

$$P(\text{Smoker} \cap \text{Cancer}) = P(\text{Smoker})P(\text{Cancer}|\text{Smoker}) = 0.3 \times 0.4 = 0.12.$$

Independence of events

Sometimes, the conditioning event does not change the original probability of the main event.

Definition 42 (Independent events)**.** Given a sample space Ω of a random experiment, two events $A, B \subseteq \Omega$ are *independents* if the probability of A does not change when conditioning on B, and vice-versa, that is,

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B),$$

if $P(A) \neq 0$ and $P(B) \neq 0$.

This means that the occurrence of one event does not give relevant information to change the uncertainty of the other.

When two events are independent, the probability of the intersection of them is equal to the product of their probabilities,

$$P(A \cap B) = P(A)P(B).$$

Independence of events

Example of tossing coins

The sample space of tossing twice a coin is $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$ and all the elements are equiprobable if the coin is fair. Thus, applying the classical definition of probability we have

$$P((H,H)) = \frac{1}{4} = 0.25.$$

If we name $H_1 = \{(H,H),(H,T)\}$, that is, having heads in the first toss, and $H_2 = \{(H,H),(T,H)\}$, that is, having heads in the second toss, we can get the same result assuming that these events are independent,

$$P(H,H) = P(H_1 \cap H_2) = P(H_1)P(H_2) = \frac{2}{4}\frac{2}{4} = \frac{1}{4} = 0.25.$$

5.5 Probability space

Probability space

Definition 43 (Probability space)**.** A *probability space* of a random experiment is a triplet $(\Omega, \mathcal{F}, P)$ where

- Ω is the sample space of the experiment.

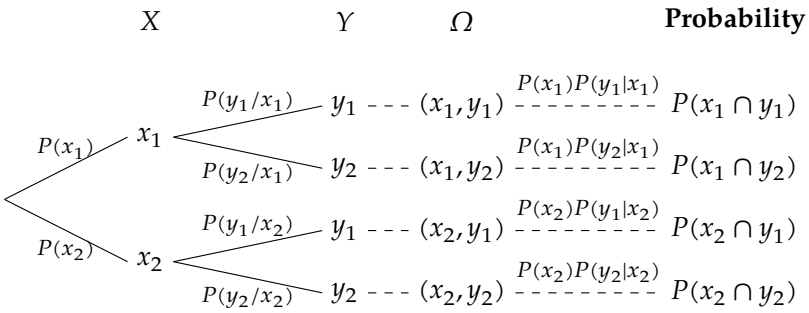
- $\mathcal{F}$ is a set of events of the experiment.
- P is a probability function.

If we know the probabilities of all the elements of Ω, then we can calculate the probability of every event in $\mathcal{F}$ and we can construct easily the probability space.

Probability space construction

In order to determine the probability of every elemental event we can use a tree diagram, using the following rules:

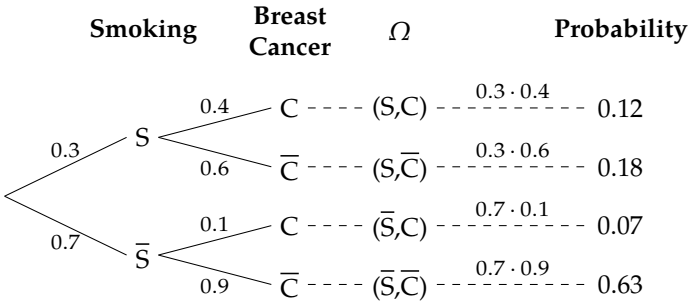
1. For every node of the tree, label the incoming edge with the probability of the variable in that level having the value of the node, conditioned by events corresponding to its ancestor nodes in the tree.
2. The probability of every elemental event in the leaves is the product of the probabilities on edges that go form the root to the leave.



Probability tree with dependent variables

Example of smoking and cancer

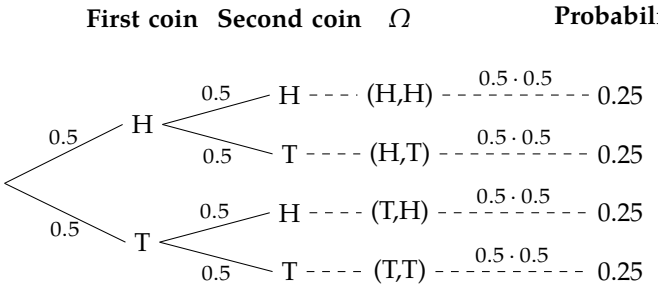
In a population there are a 30% of smokers and we know that there are a 40% of smokers with breast cancer, while only 10% of non-smokers have breast cancer. The probability tree of the probability space of the random experiment consisting of picking a random person and measuring the variables smoking and breast cancer is shown below.



Probability tree with independent variables

Example of tossing coins

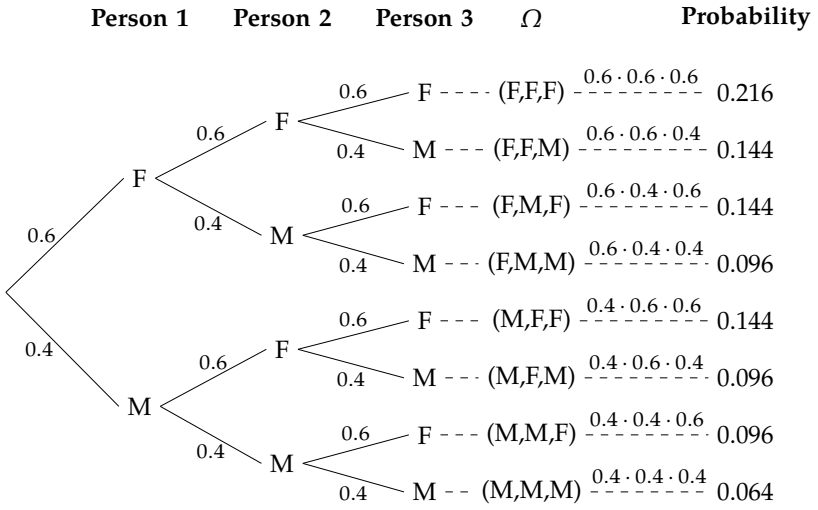
The probability tree of the random experiment of tossing two coins is shown below.



Probability tree with independent variables

Example of a sample of size 3

In a population there are 40% of males and 60% of females, the probability tree of drawing a random sample of three persons is shown below.

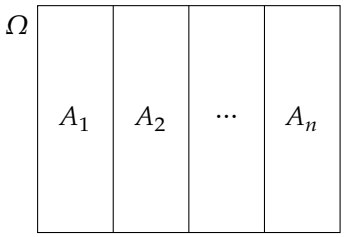


5.6 Total probability theorem

Partition of the sample space

Definition 44 (Partition of the sample space)**.** A collection of events $A_1, A_2, \ldots, A_n$ of the same sample space Ω is a *partition* of the sample space if it satisfies the following conditions

1. The union of the events is the sample space, that is, $A_1 \cup \cdots \cup A_n = \Omega$.
2. All the events are mutually incompatible, that is, $A_i \cap A_j = \emptyset \ \forall i \neq j$.



Usually it is easy to get a partition of the sample space splitting a population according to some categorical variable, like for example gender, blood type, etc.

Total probability theorem

If we have a partition of a sample space, we can use it to calculate the probabilities of other events in the same sample space.

Theorem 45 (Total probability)**.** *Given a partition $A_1, \dots, A_n$ of a sample space Ω, the probability of any other event B of the same sample space can be calculated with the formula*

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(A_i)P(B|A_i).$$

Total probability theorem

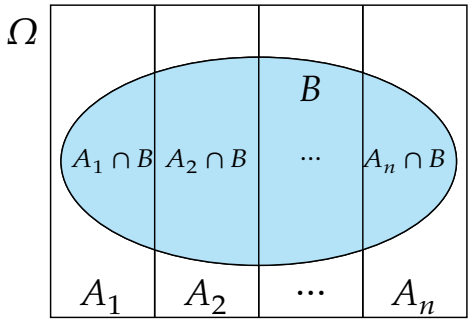
Proof

The proof of the theorem is quite simple. As $A_1, \dots, A_n$ is a partition of Ω, we have

$$B = B \cap \Omega = B \cap (A_1 \cup \cdots \cup A_n) = (B \cap A_1) \cup \cdots \cup (B \cap A_n).$$

And all the events of this union are mutually incompatible as $A_1, \dots, A_n$ are, thus

$$P(B) = P((B \cap A_1) \cup \cdots \cup (B \cap A_n)) = P(B \cap A_1) + \cdots + P(B \cap A_n) =$$
$$= P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n) = \sum_{i=1}^{n} P(A_i)P(B|A_i).$$



Total probability theorem

Example of diagnosis

A symptom S can be caused by a disease D, but it can also be present in persons without the disease. In a population, the rate of people with the disease is 0.2. We know also that 90% of persons with the disease have the symptom, while only 40% of persons without the disease have it.

What is the probability that a random person of the population has the symptom?

To answer the question we can apply the total probability theorem using the partition $\{D, \overline{D}\}$:

$$P(S) = P(D)P(S|D) + P(\overline{D})P(S|\overline{D}) = 0.2 \cdot 0.9 + 0.8 \cdot 0.4 = 0.5.$$

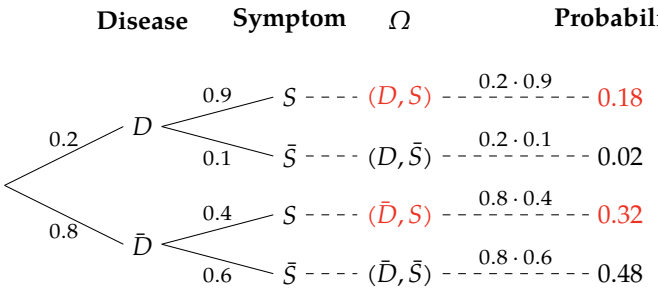
That is, half of the population has the symptom.

Indeed, it is a weighted mean of probabilities!

Total probability theorem

Example of diagnosis with a tree diagram

The answer to the previous question is even clearer with the tree diagram of the probability space.



$$P(S) = P(D \cap S) + P(\overline{D} \cap S) = P(D)P(S|D) + P(\overline{D})P(S|\overline{D})$$
$$= 0.2 \cdot 0.9 + 0.8 \cdot 0.4 = 0.18 + 0.32 = 0.5.$$

5.7 Bayes theorem

Bayes theorem

A partition of a sample space $A_1, \cdots, A_n$ may also be interpreted as a set of feasible hypothesis for a fact B.

In such cases it may be helpful to calculate the posterior probability $P(A_i|B)$ of every hypothesis.

Theorem 46 (Bayes)**.** *Given a partition $A_1, \ldots, A_n$ of a sample space Ω and another event B of the same sample space, the conditional probability of every even A_i $i = 1, \ldots, n$ on B can be calculated with the following formula*

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}.$$

Bayes theorem

Example of diagnosis

In the previous example, a more interesting question is about the diagnosis for a person with the symptom.

In this case we can interpret D and $\overline{D}$ as the two feasible hypothesis for the symptom S. The prior probabilities for them are $P(D) = 0.2$ and $P(\overline{D}) = 0.8$. That means that if we do not have information about the symptom, the diagnosis would be that the person does not have the disease.

However, if after examining the person we observe the symptom, that information changes the uncertainty about the hypothesis, and we need calculate the posterior probabilities to diagnose, that is,

$$P(D|S) \text{ and } P(\overline{D}|S)$$

Bayes theorem

Example of diagnosis

To calculate the posterior probabilities we can use the Bayes theorem.

$$P(D|S) = \frac{P(D)P(S|D)}{P(D)P(S|D) + P(\overline{D})P(S|\overline{D})} = \frac{0.2 \cdot 0.9}{0.2 \cdot 0.9 + 0.8 \cdot 0.4} = \frac{0.18}{0.5} = 0.36,$$
$$P(\overline{D}|S) = \frac{P(\overline{D})P(S|\overline{D})}{P(D)P(S|D) + P(\overline{D})P(S|\overline{D})} = \frac{0.8 \cdot 0.4}{0.2 \cdot 0.9 + 0.8 \cdot 0.4} = \frac{0.32}{0.5} = 0.64.$$

As we can see the probability of having the disease has increased. Nevertheless, the probability of not having the disease is still greater than the probability of having it, and for that reason, the diagnosis is not having the disease.

In this case it is said the the symptom S is *not decisive* in order to diagnose the disease.

5.8 Epidemiology

Epidemiology

One of the branches of Medicine that makes an intensive use of probability is **Epidemilogy**, that study the distribution and causes of diseases in populations identifying risk factors for disease and targets for preventive healthcare.

In Epidemiology we are interested in how often appears a *medical event* D (typically a disease like flu, a risk factor like smoking or a protection factor like a vaccine) that is measured as a nominal variable with two categories (occurrence or not of the event).

There are different measures related to the frequency of a medical event. The most important are:

- Prevalence
- Incidence
- Relative risk
- Odds ratio

Prevalence

Definition 47 (Prevalence)**.** The *prevalence* of a medical event D is the proportion of a particular population that is affected by a medical event.

$$\text{Prevalence}(D) = \frac{\text{Num people affected by } D}{\text{Population size}}$$

Often, the prevalence is estimated from a sample as the relative frequency of people affected by the event in the sample. It is also common to express that frequency as a percentage.

Example. To estimate the prevalence of flu a sample of 1000 persons has been studied and 150 of them had flu. Thus, the prevalence of flu is approximately 150/1000=0.15, that is, a 15%.

Incidence

Incidence measures the likelihood of occurrence of a medical event in a population within a given period of time. Incidence can be measured as a cumulative proportion or as a rate.

Definition 48 (Cumulative incidence)**.** The *cumulative incidence* of a medical event D is the proportion of people that acquired the event in a period of time, that is, the number of new cases with the event in the period of time divided by the size of the population at risk.

$$R(D) = \frac{\text{Num of new cases with } D}{\text{Population at risk size}}$$

Example. A population initially contains 1000 persons without flu and after two years of observation 160 of them got the flu. The incidence proportion of flu is 160 cases per 1000 persons per two years, i.e. 16% per two years.

Incidence rate or Absolute risk

Definition 49 (Incidence rate)**.** The *incidence rate* or *absolute risk* of a medical event D is the number of new cases with the event divided by the size of the population at risk and by the number of units of time in a given period.

$$R(D) = \frac{\text{Num of new cases with } D}{\text{Population at risk size} \times \text{Num of time units}}$$

Example. A population initially contains 1000 persons without flu and after two years of observation 160 of them got the flu. If we consider the year as the unit of time, the incidence rate of flu is 160 cases per 1000 persons divided by two years, i.e. 80 cases per 1000 persons-year or 8% persons per year.

Prevalence vs Incidence

Prevalence must not be confused with incidence. Prevalence indicates how widespread the medical event is, and is more a measure of the burden of the event on society with no regard to time at risk or when subjects may have been exposed to a possible risk factor, whereas incidence conveys information about the risk of being affected by the event.

Prevalence can be measured in cross-sectional studies at a particular time, while in order to measure incidence we need a longitudinal study observing the individuals during a period of time.

Incidence is usually more useful than prevalence in understanding the event etiology: for example, if the incidence of a disease in a population increases, then there is a risk factor that promotes it.

When the incidence is approximately constant for the duration of the event, prevalence is approximately the product of event incidence and average event duration, so

$$\text{prevalence} = \text{incidence} \times \text{duration}$$

Comparing risks

In order to determine if a factor or characteristic is associated with the medical event we need to compare the risk of the medical event in two populations, one exposed to the factor and the other not exposed. The group of people exposed to the factor is known as the *treatment group* or the *experimental group* and the group of people unexposed as the *control group*.

Usually the cases observed for each group are represented in a 2×2 table like the one below.

	Event D	No event $\overline{D}$
Treatment group (exposed)	a	b
Control group (unexposed)	c	d

Attributable risk or Risk difference *RD*

Definition 50 (Attributable risk)**.** The *attributable risk* or *risk difference* of a medical event D for people exposed to a factor is the difference between the absolute risks of the treatment group and the control group.

$$AR(D) = R_T(D) - R_C(D) = \frac{a}{a+b} - \frac{c}{c+d}.$$

The attributable risk is the risk of an event that is specifically due to the factor of interest.

Observe that the attributable risk can be positive, when the risk of the treatment group is greater than the risk of the control group, and negative, on the contrary.

Attributable risk *RR*

Example of a vaccine

To determine the effectiveness of a vaccine against the flu, a sample of 1000 person without flu was selected at the beginning of the year. Half of them were vaccinated (treatment group) and the other received a placebo (control group). The table below summarize the results at the end of the year.

	Flu D	No flu $\overline{D}$
Treatment group (vaccinated)	20	480
Control group (Unvaccinated)	80	420

The attributable risk of getting the flu for people vaccinated is

$$AR(D) = \frac{20}{20+480} - \frac{80}{80+420} = -0.12.$$

This means that the risk of getting flu in vaccinated people is a 12% less than in unvaccinated.

Relative risk *RR*

Definition 51 (Relative risk)**.** The *relative risk* of a medical event D for people exposed to a factor is the quotient between the proportions of people that acquired the event in a period of time in the treatment and control groups. That is, the quotient between the incidences of the treatment and the control groups.

$$RR(D) = \frac{\text{Risk in treatment group}}{\text{Risk in control group}} = \frac{R_T(D)}{R_C(D)} = \frac{a/(a+b)}{c/(c+d)}$$

Relative risk compares the risk of a medical event between the treatment and the control groups.

- $RR = 1 \Rightarrow$ There is no association between the event and the exposure to the factor.
- $RR < 1 \Rightarrow$ Exposure to the factor decreases the risk of the event.
- $RR > 1 \Rightarrow$ Exposure to the factor increases the risk of the event.

The further from 1, the stronger the association.

Relative risk *RR*

Example of a vaccine

To determine the effectiveness of a vaccine against the flu, a sample of 1000 person without flu was selected at the beginning of the year. Half of them were vaccinated (treatment group) and the other received a placebo (control group). The table below summarize the results at the end of the year.

	Flu D	No flu $\overline{D}$
Treatment group (vaccinated)	20	480
Control group (Unvaccinated)	80	420

The relative risk of getting the flu for people vaccinated is

$$RR(D) = \frac{20/(20+480)}{80/(80+420)} = 0.25.$$

This means that vaccinated people were only one-fourth as likely to develop flu as were unvaccinated people, i.e. the vaccine reduce the risk of flu by 75%.

Odds

An alternative way of measuring the risk of a medical event is the *odds*.

Definition 52 (Odds)**.** The *odds* of a medical event D in a population is the quotient between the people that acquired the event and people that not in a period of time.

$$ODDS(D) = \frac{\text{Num new cases with } D}{\text{Num cases without } D} = \frac{P(D)}{P(\overline{D})}$$

Unlike incidence, that is a proportion less than or equal to 1, the odds can be greater than 1. However, it is possible to convert an odds into a probability with the formula

$$P(D) = \frac{ODDS(D)}{ODDS(D) + 1}$$

Example A population initially contains 1000 persons without flu and after a year 160 of them got the flu. The odds of flu is 160/840.

Observe that the incidence is 160/1000.

Odds ratio *OR*

Definition 53 (Odds ratio)**.** The *odds ratio* of a medical event D for people exposed to a factor is the quotient between the odds of the event of the treatment and the control groups.

$$OR(D) = \frac{\text{Odds in treatment group}}{\text{Odds in control group}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Odds ratio compares the odds of a medical event between the treatment and the control groups. The interpretation is similar to the relative risk.

- $OR = 1 \Rightarrow$ There is no association between the event and the exposure to the factor.
- $OR < 1 \Rightarrow$ Exposure to the factor decreases the risk of the event.
- $OR > 1 \Rightarrow$ Exposure to the factor increases the risk of the event.

The further from 1, the stronger the association.

Odds ratio *OR*

Example of a vaccine

To determine the effectiveness of a vaccine against the flu, a sample of 1000 person without flu was selected at the beginning of the year. Half of them were vaccinated (treatment group) and the other received a placebo (control group). The table below summarize the results at the end of the year.

	Flu D	No flu $\overline{D}$
Treatment group (vaccinated)	20	480
Control group (Unvaccinated)	80	420

The odds ratio of getting the flu for people vaccinated is

$$OR(D) = \frac{20/480}{80/420} = 0.21875.$$

This means that the odds of getting the flu versus not getting the flu in vaccinated individuals is almost one fifth of that in unvaccinated, i.e. approximately for every 22 persons vaccinated with flu there will be 100 persons unvaccinated with flu.

Relative risk vs Odds ratio

Relative risk and odds ratio are two measures of association but their interpretation is slightly different. While the relative risk expresses a comparison of risks between the treatment and control groups, the odds ratio expresses a comparison of odds, that is not the same than the risk. Thus, an odds ratio of 2 *does not* mean that the treatment group has the double of risk of acquire the medical event.

The interpretation of the odds ratio is trickier because is counterfactual, and give us how many times is more frequent the event in the treatment group in comparison with the control group, assuming that in the control group the event is as frequent as the non-event.

The advantage of the odds ratio is that it does not depend on the prevalence or the incidence of the event, and must be used necessarily when the number of people with the medical event is selected arbitrarily in both groups, like in the case-control studies.

Relative risk vs Odds ratio

Example of lung cancer and smoking

In order to determine the association between lung cancer and smoking two samples were selected (the second one with the double of non-cancer individuals) getting the following results:

Sample 1

	Cancer	No cancer
Smokers	60	80
Non-smokers	40	320

$$RR(D) = \frac{60/(60+80)}{40/(40+320)} = 3.86.$$
$$OR(D) = \frac{60/80}{40/320} = 6.$$

Sample 2

	Cancer	No cancer
Smokers	60	160
Non-smokers	40	640

$$RR(D) = \frac{60/(60+160)}{40/(40+640)} = 4.64.$$
$$OR(D) = \frac{60/160}{40/640} = 6.$$

Thus, when we change the incidence or the prevalence of the event (lung cancer) the relative risk changes, while the odds ratio not.

Relative risk vs Odds ratio

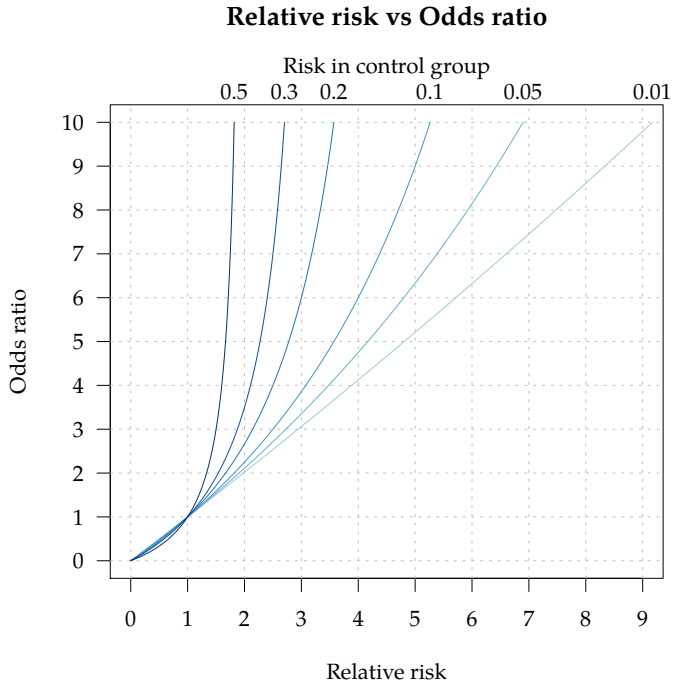
The relation between the relative risk and the odds ratio is given by the following formula

$$RR = \frac{OR}{1 - R_C + R_C * OR} = OR * \frac{1 - R_T}{1 - R_C},$$

where R_C and R_T are the prevalence or the incidence in control and treatment groups respectively.

The odds ratio always overestimate the relative risk when it is greater than 1 and underestimate it when it is less than 1. However, with rare medical events (with very small prevalence or incidence) the relative risk and the odds ratio are almost the same.

Relative risk vs Odds ratio



5.9 Diagnostic tests

Diagnostic tests

In Epidemiology it is common to use diagnostic test to diagnose diseases.

In general, diagnostic tests are not fully reliable and have some risk of misdiagnosis as it is represented in the table below.

	Presence of disease D	Absence of disease $\overline{D}$
Test outcome positive +	True Positive TP	False Positive FP
Test outcome negative −	False Negative FN	True Negative TN

Sensitivity and specificity of a diagnostic test

The performance of a diagnostic test depends on the following two probabilities.

Definition 54 (Sensitivity)**.** The *sensitivity* of a diagnostic test is the proportion of positive outcomes in persons with the disease

$$P(+|D) = \frac{TP}{TP + FN}$$

Definition 55 (Specificity)**.** The *specificity* of a diagnostic test is the proportion of negative outcomes in persons without the disease

$$P(-|\overline{D}) = \frac{TN}{TN + FP}$$

Sensitivity and specificity interpretation

Usually, there is a trade-off between sensitivity and specificity.

A test with high sensitivity will detect the disease in most sick persons, but it will produce also more false positives than a less sensitive test. This way, a positive outcome in a test with high sensitivity is not useful for confirming the disease, but a negative outcome is useful for ruling out the disease, since it rarely give negative outcomes in sick people.

On the other hand, a test with a high specificity will rule out the disease in most healthy persons, but it will produce also more false negatives than a less specific test. Thus, a negative outcome in a test with high specificity is not useful for ruling out the disease, but a positive is useful to confirm the disease, since it rarely give positive outcomes in healthy people.

Sensitivity and specificity interpretation

Deciding on a test with greater sensitivity or a test with greater specificity depends on the type of disease and the goal of the test. In general, we will use a sensitive test when:

- The disease is serious and it is important to detect it.
- The disease is curable.
- The false positives do not provoke serious traumas.

And we will use a specific test when:

- The disease is important but difficult or impossible to cure.
- The false positives provoke serious traumas.
- The treatment of false positives can have dangerous consequences.

Predictive values of a diagnostic test

But the most important aspect of a diagnostic test is its predictive power, that is measured with the following two posterior probabilities.

Definition 56 (Positive predictive value PPV)**.** The *positive predictive value* of a diagnostic test is the proportion of persons with the disease to persons with a positive outcome

$$P(D|+) = \frac{TP}{TP + FP}$$

Definition 57 (Negative predictive value NPV)**.** The *negative predictive value* of a diagnostic test is the proportion of persons without the disease to persons with a negative outcome

$$P(\overline{D}|-) = \frac{TN}{TN + FN}$$

Predictive values interpretation

Positive and negative predictive values allow to confirm or to rule out the disease, respectively, if they reach at least a threshold of 0.5.

$$\begin{array}{rcl} PPV > 0.5 & \Rightarrow & \text{Disease diagnostic} \\ NPV > 0.5 & \Rightarrow & \text{Not disease diagnostic} \end{array}$$

However, these probabilities depends on prevalence of the disease $P(D)$. They can be calculated from the sensitivity and the specificity of the diagnostic test using the Bayes theorem.

$$PPV = P(D|+) = \frac{P(D)P(+|D)}{P(D)P(+|D) + P(\overline{D})P(+|\overline{D})}$$
$$NPV = P(\overline{D}|-) = \frac{P(\overline{D})P(-|\overline{D})}{P(D)P(-|D) + P(\overline{D})P(-|\overline{D})}$$

Thus, with frequent diseases, the positive predictive value increases, and with rare diseases, the negative predictive value increases.

Diagnostic tests

Example

A diagnostic test for the flu has been tried in a random sample of 1000 persons. The results are summarized in the table below.

	Presence of flu D	Absence of flu $\overline{D}$
Test outcome +	95	90
Test outcome −	5	810

According to this sample, the prevalence of the flu can be estimated as

$$P(D) = \frac{95+5}{1000} = 0.1.$$

The sensitivity of this diagnostic test is

$$P(+|D) = \frac{95}{95+5} = 0.95.$$

And the specificity is

$$P(-|\overline{D}) = \frac{810}{90+810} = 0.9.$$

Diagnostic tests

Example cont.

The predictive positive value of the diagnostic test is

$$PPV = P(D|+) = \frac{95}{95+90} = 0.5135.$$

As this value is over 0.5, this means that we will diagnose the flu if the outcome of the test is positive. However, the confidence in the diagnostic will be low, as this value is pretty close to 0.5.

On the other hand, the predictive negative value is

$$NPV = P(\overline{D}|-) = \frac{810}{5+810} = 0.9939.$$

As this value is almost 1, that means that is almost sure that a person does not have the flu if he or she gets a negative outcome in the test.

Thus, this test is a powerful test to rule out the flu, but not so powerful to confirm it.

Likelihood ratios of a diagnostic test

The following measures are usually derived from sensitivity and specificity.

Definition 58 (Positive likelihood ratio $LR+$)**.** The *positive likelihood ratio* of a diagnostic test is the ratio between the probability of positive outcomes in persons with the disease and healthy persons respectively,

$$LR+ = \frac{P(+|D)}{P(+|\overline{D})} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$$

Definition 59 (Negative likelihood ratio $LR-$)**.** The *negative likelihood ratio* of a diagnostic test is the ratio between the probability of negative outcomes in persons with the disease and healthy persons respectively,

$$LR- = \frac{P(-|D)}{P(-|\overline{D})} = \frac{1 - \text{Sensitivity}}{\text{Specificity}}$$

Likelihood ratios interpretation

Positive likelihood ratio can be interpreted as the number of times that a positive outcome is more probable in people with the disease than in people without it.

On the other hand, negative likelihood ratio can be interpreted as the number of times that a negative outcome is more probable in people without the disease than in people with it.

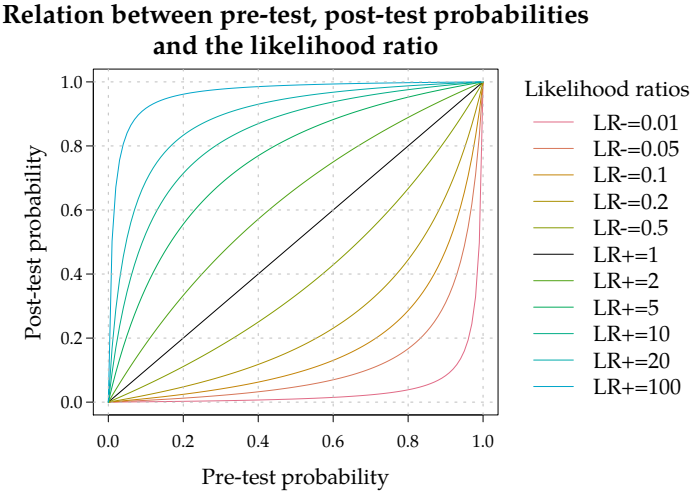
Post-test probabilities can be calculated from pre-test probabilities through likelihood ratios.

$$P(D|+) = \frac{P(D)P(+|D)}{P(D)P(+|D) + P(\overline{D})P(+|\overline{D})} = \frac{P(D)LR+}{1 - P(D) + P(D)LR+}$$

Thus,

- A likelihood ratio greater than 1 increases the probability of disease.
- A likelihood ratio less than 1 decreases the probability of disease.
- A likelihood ratio 1 does not change the pre-test probability.

Likelihood ratios interpretation



6 Discrete random variables

Random variables

The process of drawing a sample randomly is a random experiment and any variable measured in the sample is a *random variable* because the values taken by the variable in the individuals of the sample are a matter of chance.

Definition 60 (Random variable)**.** A *random variable* X is a function that maps every element of the sample space of a random experiment to a real number.

$$X : \Omega \to \mathbb{R}$$

The set of values that the variable can assume is called the *range* and is represented by $\text{Ran}(X)$.

In essence, a random variable is a variable whose values come from a random experiment, and every value has a probability of occurrence.

Example. The variable X that measures the outcome of rolling a dice is a random variable and its range is

$$\text{Ran}(X) = \{1, 2, 3, 4, 5, 6\}$$

Types of random variables

There are two types of random variables:

Discrete They take isolated values, and their range is numerable. Example. Number of children of a family, number of smoked cigarettes, number of subjects passed, etc.

Continuous They can take any value in a real interval, and their range is non-numerable. Example. Weight, height, age, cholesterol level, etc.

The way of modelling each type of variable is different. In this chapter we are going to study how to model discrete random variables.

6.1 Probability distribution of a discrete random variable

Probability distribution of a discrete random variable

As values of a discrete random variable are linked to the elementary events of a random experiment, every value has a probability.

Definition 61 (Probability function)**.** The *probability function* of a discrete random variable X is the function $f(x)$ that maps every value x_i of the variable to its probability,

$$f(x_i) = P(X = x_i).$$

We can also accumulate probabilities the same way that we accumulated sample frequencies.

Definition 62 (Distribution function)**.** The *distribution function* of a discrete random variable X is the function $F(x)$ that maps every value x_i of the variable to the probability of having a value less than or equal to x_i,

$$F(x_i) = P(X \leq x_i) = f(x_1) + \cdots + f(x_i).$$

Probability distribution of a discrete random variable

The range of a discrete random variable and its probability function is known as **Probability Distribution** of the variable, and it is usually presented in a table

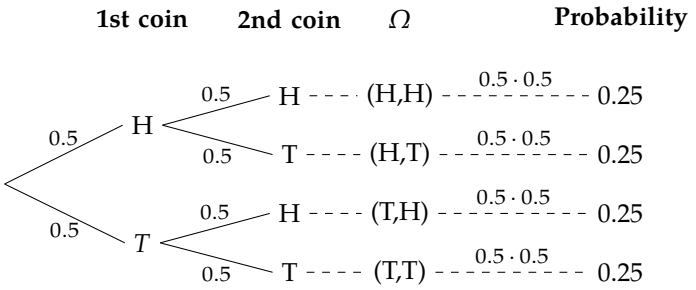
X	x_1	x_2	$\cdots$	x_n	$\sum$
$f(x)$	$f(x_1)$	$f(x_2)$	$\cdots$	$f(x_n)$	1
$F(x)$	$F(x_1)$	$F(x_2)$	$\cdots$	$F(x_n) = 1$	

The same way that the sample frequency table shows the distribution of values of a variable in the sample, the probability distribution of a discrete random variable shows the distribution of values in the whole population.

Probability distribution of a discrete random variable

Example of tossing two coins

Let X be the discrete random variable that measures the number of heads after tossing two coins. The probability tree of the probability space of the random experiment is shown below.



According to this, the probability distribution of X is

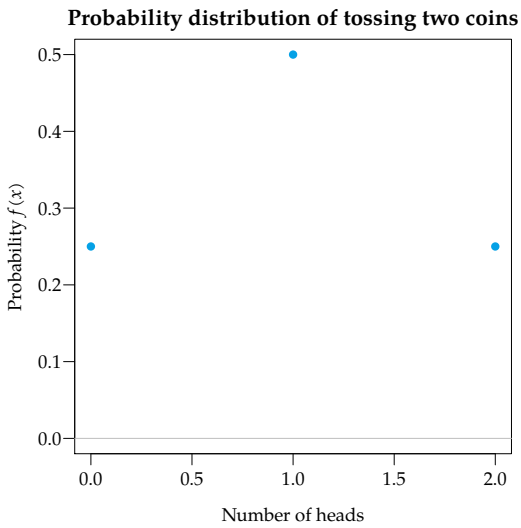
X	0	1	2
$f(x)$	0.25	0.5	0.25
$F(x)$	0.25	0.75	1

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.25 & \text{if } 0 \leq x < 1 \\ 0.75 & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

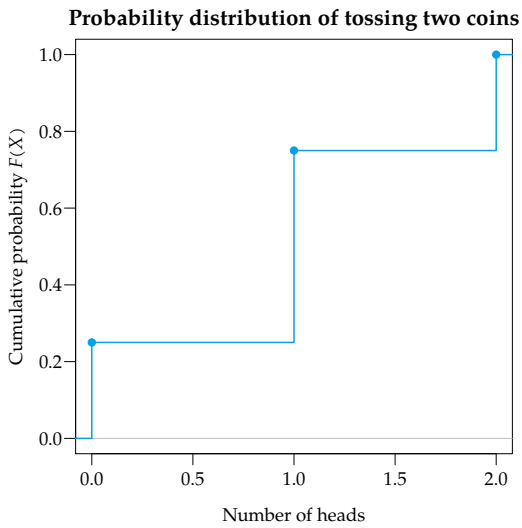
Probability distribution charts

Example of tossing two coins

Probability function



Distribution function



Population statistics

The same way we use sample statistics to describe the sample frequency distribution of a variable, we use population statistics to describe the probability distribution of a random variable in the whole population.

The population statistics definition is analogous to the sample statistics definition, but using probabilities instead of relative frequencies.

The most important are [1]:

- Mean:

$$\mu = E(X) = \sum_{i=1}^{n} x_i f(x_i)$$

- Variance:

$$\sigma^2 = Var(X) = \sum_{i=1}^{n} x_i^2 f(x_i) - \mu^2$$

- Standard deviation:

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

Population statistics

Example of tossing two coins

In the random experiment of tossing two coins the probability distribution is

X	0	1	2
$f(x)$	0.25	0.5	0.25
$F(x)$	0.25	0.75	1

[1] To distinguish population statistics from sample statistics we use Greek letters

The main population statistics are

$$\mu = \sum_{i=1}^{n} x_i f(x_i) = 0 \cdot 0.25 + 1 \cdot 0.5 + 2 \cdot 0.25 = 1 \text{ heads},$$
$$\sigma^2 = \sum_{i=1}^{n} x_i^2 f(x_i) - \mu^2 = (0^0 \cdot 0.25 + 1^2 \cdot 0.5 + 2^2 \cdot 0.25) - 1^2 = 0.5 \text{ heads}^2,$$
$$\sigma = +\sqrt{0.5} = 0.71 \text{ heads}.$$

Discrete probability distribution models

According to the type of experiment where the random variable is measured, there are different probability distributions models. The most important are

- Discrete uniform
- Binomial
- Poisson

6.2 Discrete uniform distribution

Discrete uniform probability distribution model $U(a,b)$

When all the values of a random variable X have equal probability, the probability distribution of X is *uniform*.

Definition 63 (Discrete uniform distribution $U(a,b)$)**.** A discrete random variable X follows a *discrete uniform distribution model* with parameters a and b, noted $X \sim U(a,b)$, if its range is $\text{Ran}(X) = \{a, a+1, \dots, b\}$ and its probability function is

$$f(x) = \frac{1}{b-a+1}.$$

Observe that a and b are the minimum and the maximum of the range respectively.

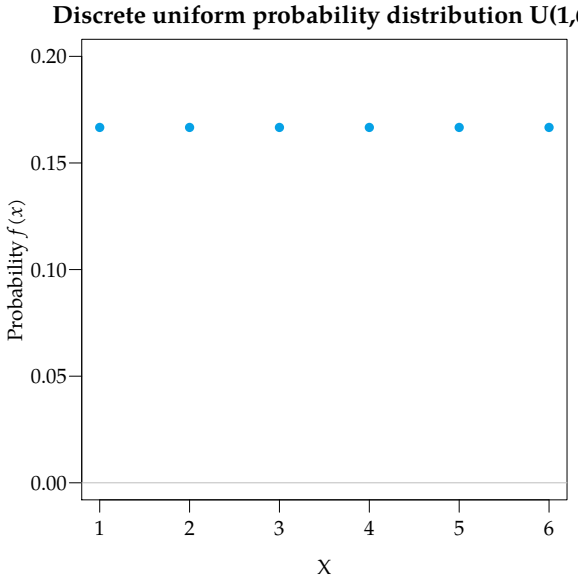
The mean and the variance are

$$\mu = \sum_{i=0}^{b-a} \frac{a+i}{b-a+1} = \frac{a+b}{2} \qquad \sigma^2 = \sum_{i=0}^{b-a} \frac{(a+i-\mu)^2}{b-a+1} = \frac{(b-a+1)^2-1}{12}$$

Discrete uniform probability distribution model $U(a,b)$

Example of rolling a dice

The variable that measures the outcome of rolling a dice follows a discrete uniform distribution model $U(1,6)$.



6.3 Binomial distribution

Binomial distribution

Usually the binomial distribution corresponds to a variable measured in a random experiment with the following features:

- The experiment consist in a sequence of n repetitions of the same trial.
- Each trial is repeated in identical conditions and produces two possible outcomes known as *Success* or *Failure*.
- The trials are independent.
- The probability of Success is the same in all the trials and is $P(\text{Success}) = p$.

Under these conditions, the discrete random variable X that measures the number of successes in the n trials follows a *binomial distribution model* with parameters n and p.

Binomial distribution model $B(n,p)$

Definition 64 (Binomial distribution $(B(n,p))$**.** A discrete random variable X follows a *binomial distribution model* with parameters n and p, noted $X \sim B(n,p)$, if its range is $\text{Ran}(X) = \{0,1,\dots,n\}$ and its probability function is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

Observe that n is known as the number of repetitions of a trial and p is known as the probability of Success in every repetition.

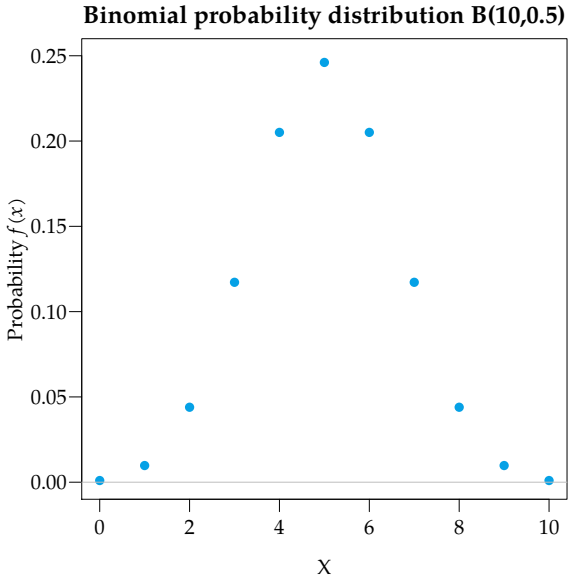
The mean and the variance are

$$\mu = n \cdot p \qquad \sigma^2 = n \cdot p \cdot (1-p).$$

Binomial distribution model $B(n,p)$

Example of tossing 10 coins

The variable that measures the number of heads after tossing 10 coins follows a binomial distribution model $B(10, 0.5)$.



Binomial distribution model $B(n,p)$

Example of tossing 10 coins

If $X \sim B(10, 0.5)$ is the random variable that measures the number of heads after tossing 10 coins, then

- The probability of getting 4 heads is

$$f(4) = \binom{10}{4} 0.5^4 (1-0.5)^{10-4} = \frac{10!}{4!6!} 0.5^4 0.5^6 = 210 \cdot 0.5^{10} = 0.2051.$$

- The probability of getting 2 or less heads is

$$\begin{aligned} F(2) &= f(0) + f(1) + f(2) = \\ &= \binom{10}{0} 0.5^0 (1-0.5)^{10-0} + \binom{10}{1} 0.5^1 (1-0.5)^{10-1} + \binom{10}{2} 0.5^2 (1-0.5)^{10-2} = \\ &= 0.0547. \end{aligned}$$

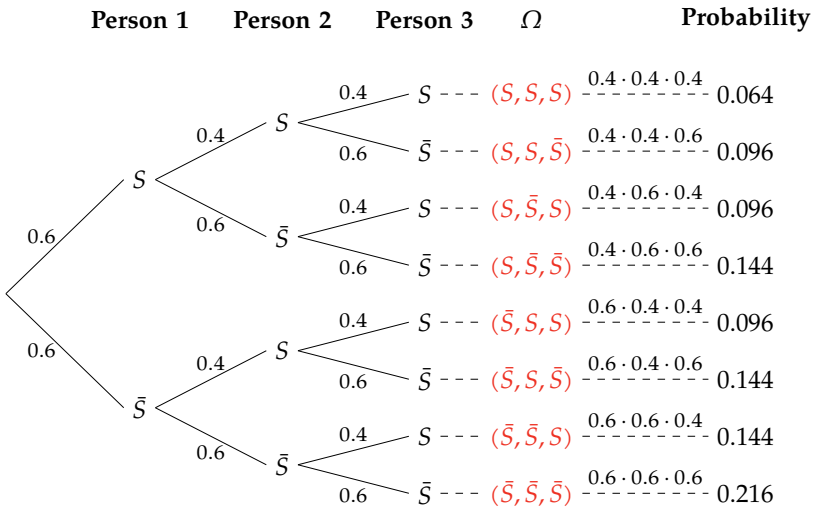
- And the expected number of heads is

$$\mu = 10 \cdot 0.5 = 5 \text{ heads.}$$

Binomial distribution function $B(n,p)$

Example of random sampling with replacement

In a population there are a 40% of smokers. The variable X that measures the number of smokers in a random sample with replacement of 3 persons follows a binomial distribution model $X \sim B(3, 0.4)$.



$$f(0) = \binom{3}{0}0.4^0(1-0.4)^{3-0} = 0.6^3, \qquad f(1) = \binom{3}{1}0.4^1(1-0.4)^{3-1} = 3\cdot 0.4\cdot 0.6^2,$$
$$f(2) = \binom{3}{2}0.4^2(1-0.4)^{3-2} = 3\cdot 0.4^2\cdot 0.6, \qquad f(3) = \binom{3}{3}0.4^3(1-0.4)^{3-3} = 0.4^3.$$

6.4 Poisson distribution

Poisson distribution

Usually the Poisson distribution correspond to a variable measured in a random experiment with the following features:

- The experiment consists of observing the number of events occurring in a fixed interval of time or space. For instance, number of births in a month, number of emails in one hour, number of red blood cells in a volume of blood, etc.
- The events occur independently.
- The experiment produces the same average rate of events λ for every interval unit.

Under these conditions, the discrete random variable X that measures the number of events in an interval unit follows a *Poisson distribution model* with parameter λ.

Poisson distribution model $P(\lambda)$

Definition 65 (Poisson distribution $P(\lambda)$)**.** A discrete random variable X follows a *Poisson distribution model* with parameter λ, noted $X \sim P(\lambda)$, if its range is $\text{Ran}(X) = \{0, 1, \dots, \infty\}$ and its probability function is

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

Observe that λ is the average rate of event for an interval unit, and it will change if the interval changes.

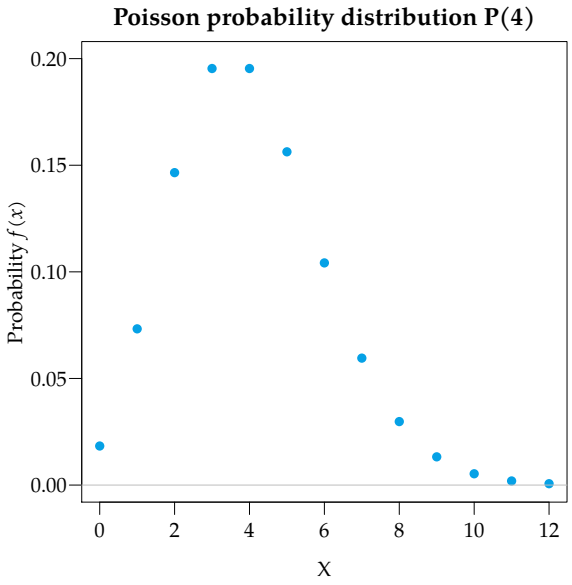
The mean and the variance are

$$\mu = \lambda \qquad \sigma^2 = \lambda.$$

Poisson distribution $P(\lambda)$

Example of number of births in a city

In a city there are an average of 4 births every day. The random variable X that measures the number of births in a day in the city follows a Poisson distribution model $X \sim P(4)$.



Poisson distribution model $P(\lambda)$

Example of number of births in a city

If $X \sim P(4)$ is the random variable that measures the number of births in the city, then

- The probability that there are 5 births in a day is

$$f(5) = e^{-4}\frac{4^5}{5!} = 0.1563.$$

- The probability that there are less than 2 births in a day is

$$F(1) = f(0) + f(1) = e^{-4}\frac{4^0}{0!} + e^{-4}\frac{4^1}{1!} = 5e^{-4} = 0.0916.$$

- The probability that there are more than 1 birth a day is

$$P(X > 1) = 1 - P(X \leq 1) = 1 - F(1) = 1 - 0.0916 = 0.9084.$$

Approximation of Binomial by Poisson distribution

Law of rare events

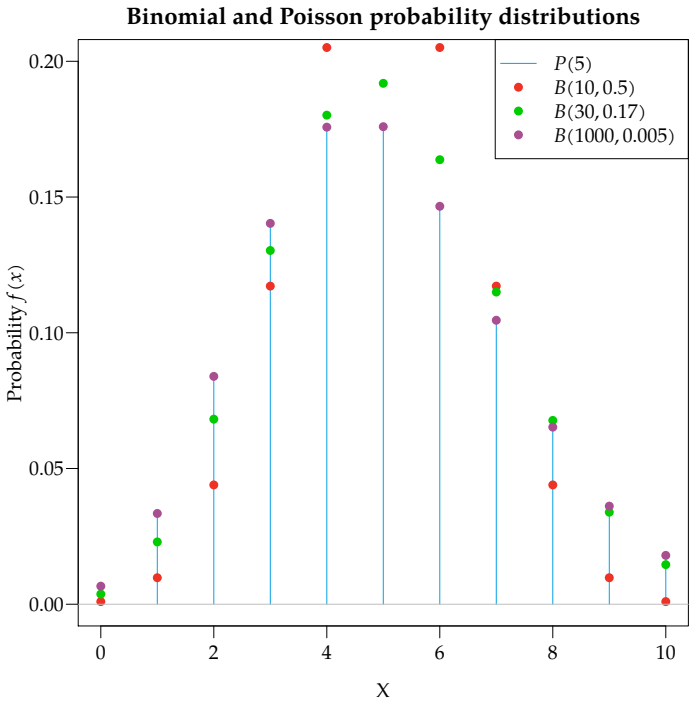
The Poisson distribution can be obtained from the Binomial distribution when the number of trials repetition tends to infinite and the probability of Success tends to zero.

Theorem 66 (Law or rare events)**.** *The Binomial distribution $X \sim B(n,p)$ tends to the Poisson distribution $P(\lambda)$, with $\lambda = n \cdot p$, when n tends to infinite and p tends to zero, that is,*

$$\lim_{n\to\infty, p\to 0} \binom{n}{x} p^x (1-p)^{n-x} = e^{-\lambda}\frac{\lambda^x}{x!}.$$

In practice, this approximation can be used for $n \geq 30$ and $p \leq 0.1$.

Approximation of Binomial by Poisson distribution



Approximation of Binomial by Poisson distribution

Example

A vaccine produces an adverse reaction in 4% of the cases. If a sample of 50 persons are vaccinated, what is the probability of having more than 2 persons with an adverse reaction?

The variable that measures the number of persons with an adverse reaction in the sample follows a Binomial distribution model $X \sim B(50, 0.04)$, but as $n = 50 > 30$ and $p = 0.04 < 0.1$, we can apply the law of rare events and use the Poisson distribution model $P(50 \cdot 0.04) = P(2)$ to do the calculations.

$$\begin{aligned} P(X > 2) &= 1 - P(X \leq 2) = 1 - f(0) - f(1) - f(2) = \\ &= 1 - e^{-2}\frac{2^0}{0!} - e^{-2}\frac{2^1}{1!} - e^{-2}\frac{2^2}{2!} = \\ &= 1 - 5e^{-2} = 0.3233. \end{aligned}$$

7 Continuous random variables

Continuous random variables

Continuous random variables, unlike discrete random variables, can take any value in a real interval. Thus the range of a continuous random variables is infinite and uncountable.

Such a density of values makes impossible to compute the probability for each one of them, and therefore, it's not possible to define a probabilistic model trough a probability function like with discrete random variables.

Besides, usually the measurement of continuous random variable is limited by the precision of the measuring instrument. For instance, when somebody says that is 1.68 meters tall, his or her true height is no exactly 1.68 meters, because the precision of the measuring instrument is only cm (two decimal places). This means that the true height of that person is between 1.675 y 1.685 meters.

Hence, for continuous variables, it makes no sense to calculate the probability of an isolated value, and we will calculate probabilities for intervals.

7.1 Probability distribution of a continuous random variable

Probability density function

To model the probability distribution of a continuous random variable we use a probability density function.

Definition 67 (Probability density function)**.** The *probability density function* of a continuous random variable X is a function $f(x)$ that meets the following conditions:

- It is non-negative: $f(x) \geq 0 \ \forall x \in \mathbb{R}$,
- The area bounded by the curve of the density function and the x-axis is equal to 1, that is,

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

- The probability that X assumes a value between a and b is equal to the area under the density function bounded by a and b, that is,

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)\, dx$$

The probability density function measures the relative likelihood of every value, but watch out!, $f(x)$ is not the probability of x,because $P(X = x) = 0$ for every x value by definition.

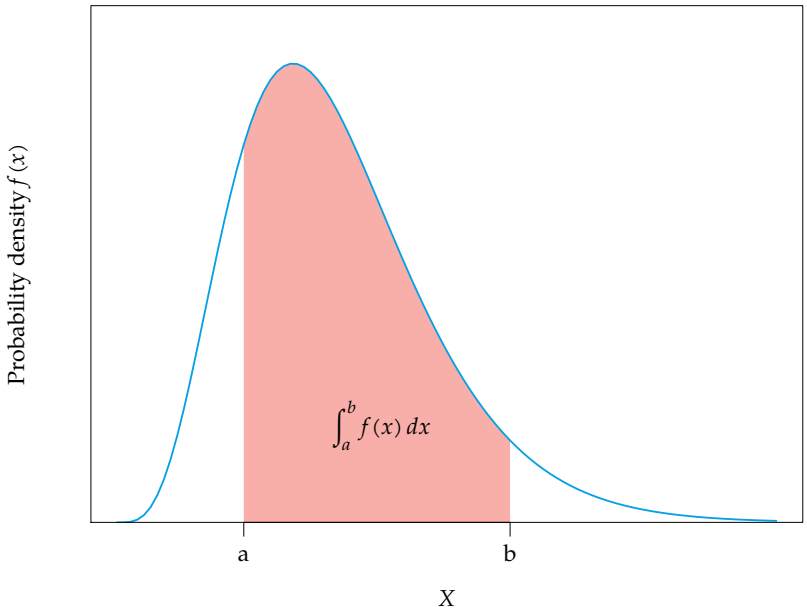
Distribution function

The same way that for discrete random variables, for continuous random variables it makes sense to calculate cumulative probabilities.

Definition 68 (Distribution function)**.** The *distribution function* of a continuous random variable X is a function $F(x)$ that maps every value a to the probability that X takes on a value less than or equal to a, that is,

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)\, dx.$$

Probabilities as areas



$$P(a \leq X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a)$$

Probabilities as areas

Example

Given the following function

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ e^{-x} & \text{si } x \geq 0, \end{cases}$$

let's check that is a density function.

As this function is clearly non-negative, we have to check that total area bounded by the curve and the x-axis is 1.

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\infty} f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{\infty} e^{-x}\,dx = \\ = [-e^{-x}]_0^{\infty} = -e^{-\infty} + e^0 = 1.$$

Now, let's calculate the probability of X having a value between 0 and 2.

$$P(0 \leq X \leq 2) = \int_0^2 f(x)\,dx = \int_0^2 e^{-x}\,dx = [-e^{-x}]_0^2 = -e^{-2} + e^0 = 0.8646.$$

Population statistics

The calculation of the population statistics is similar to the case of discrete variables, but using the density function instead of the probability function, and extending the discrete sum to the integral.

The most important are:

- Mean:

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$$

- Variance:
$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2$$
- Standard deviation:
$$\sigma = +\sqrt{\sigma^2}$$

Population statistics

Example

Let X be a variable with the following probability density function

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ e^{-x} & \text{si } x \geq 0 \end{cases}$$

The mean is

$$\mu = \int_{-\infty}^{\infty} xf(x)\, dx = \int_{-\infty}^{0} xf(x)\, dx + \int_{0}^{\infty} xf(x)\, dx = \int_{-\infty}^{0} 0\, dx + \int_{0}^{\infty} xe^{-x}\, dx =$$
$$= [-e^{-x}(1+x)]_0^{\infty} = 1.$$

and the variance is

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2 = \int_{-\infty}^{0} x^2 f(x)\, dx + \int_{0}^{\infty} x^2 f(x)\, dx - \mu^2 =$$
$$= \int_{-\infty}^{0} 0\, dx + \int_{0}^{\infty} x^2 e^{-x}\, dx - \mu^2 = \left[-e^{-x}(x^2+2x+2)\right]_0^{\infty} - 1^2 = 2e^0 - 1 = 1.$$

Continuous probability distribution models

According to the type of experiment where the random variable is measured, there are different probability distributions models. The most common are:

- Continuous uniform.
- Normal.
- Student's T.
- Chi-square.
- Fisher-Snedecor's F.

7.2 Continuous uniform distribution

Continuous uniform probability distribution model $U(a,b)$

When all the values of a random variable X have equal probability, the probability distribution of X is uniform.

Definition 69 (Continuous uniform distribution $U(a,b)$)**.** A continuous random variable X follows a probability distribution model *uniform* of parameters a and b, noted $X \sim U(a,b)$, if its range is $\text{Ran}(X) = [a,b]$ and its density function is

$$f(x) = \frac{1}{b-a} \quad \forall x \in [a,b]$$

Observe that a and b are the minimum and the maximum of the range respectively, and that the density function is constant.

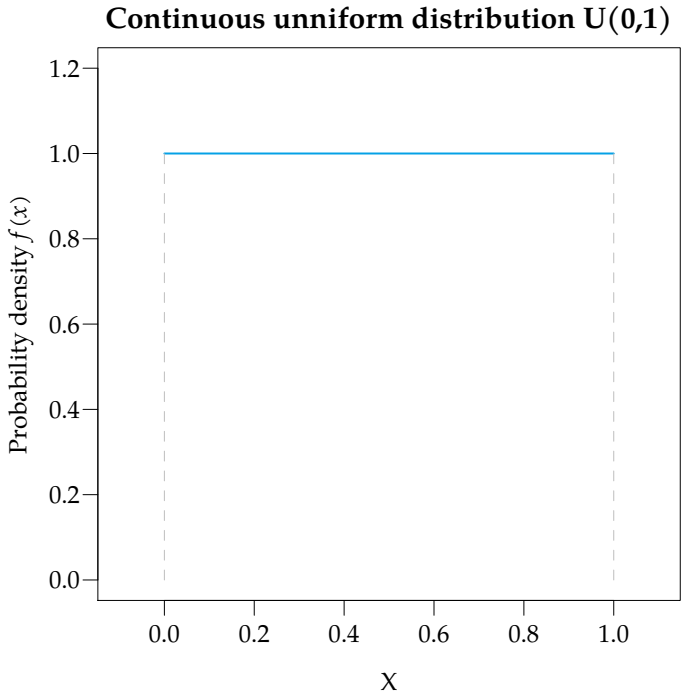
The mean and the variance are

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12}.$$

Continuous uniform probability density function

Example

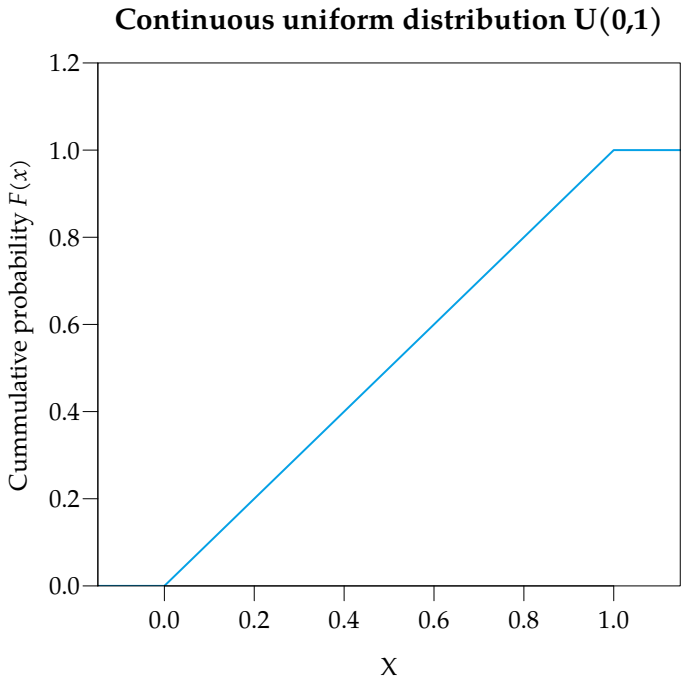
The generation of a random number between 0 and 1 is follows a continuous uniform distribution $U(0,1)$.



Continuous uniform distribution function

Example

As the density function is constant, the distribution function has a linear growth.



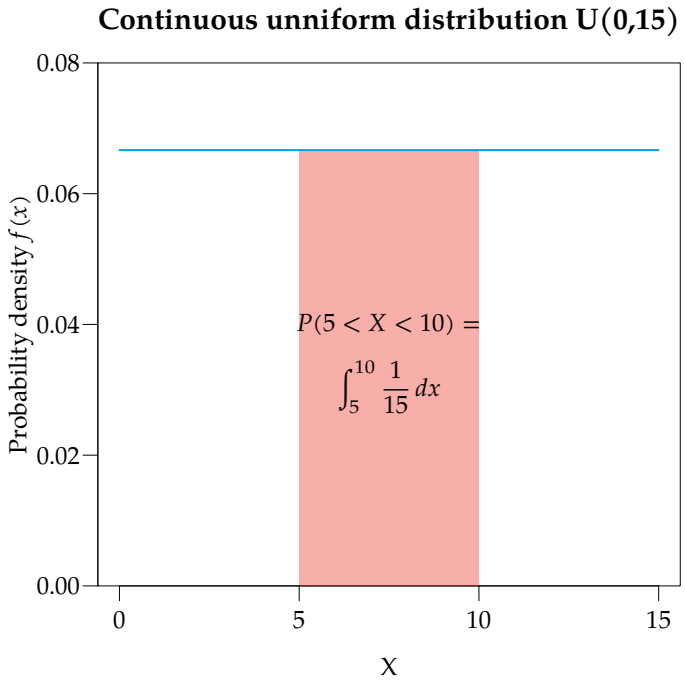
Continuous uniform probability calculation

Example of waiting time for a bus

A bus has a frequency of 15 minutes. Assuming that a person can arrive to the bus station in any time, *what is the probability of waiting for the bus between 5 and 10 minutes?* In this case, the variable X that measures the waiting time follows a continuous uniform distribution $U(0, 15)$ as any waiting time between 0 and 15 is equally likely.

Then, the probability of waiting between 5 and 10 minutes is

$$P(5 \leq X \leq 10) = \int_5^{10} \frac{1}{15}\, dx = \left[\frac{x}{15}\right]_5^{10} = \\ = \frac{10}{15} - \frac{5}{15} = \frac{1}{3}.$$



And the expected waiting (the mean) time is $\mu = \frac{0+15}{2} = 7.5$ minutes.

7.3 Normal distribution

Normal probability distribution model $N(\mu, \sigma)$

The normal distribution model is, without a doubt, the most important continuous distribution model as it is the most common in Nature.

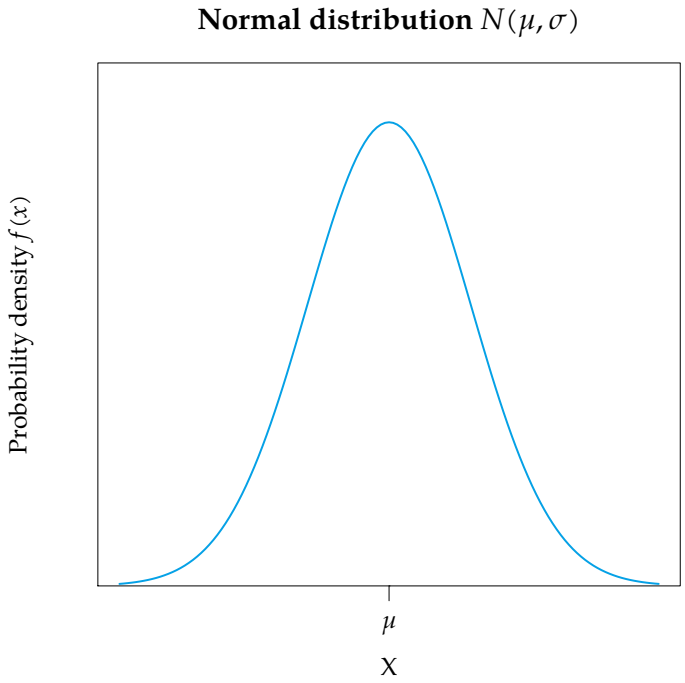
Definition 70 (Normal distribution $N(\mu, \sigma)$)**.** A continuous random variable X follows a probability distribution model *normal* of parameters μ and σ, noted $X \sim N(\mu, \sigma)$, if its range is $\text{Ran}(X) = (-\infty, \infty)$ and its density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The two parameters μ and σ are the mean and the standard deviation of the population respectively.

Normal probability density function

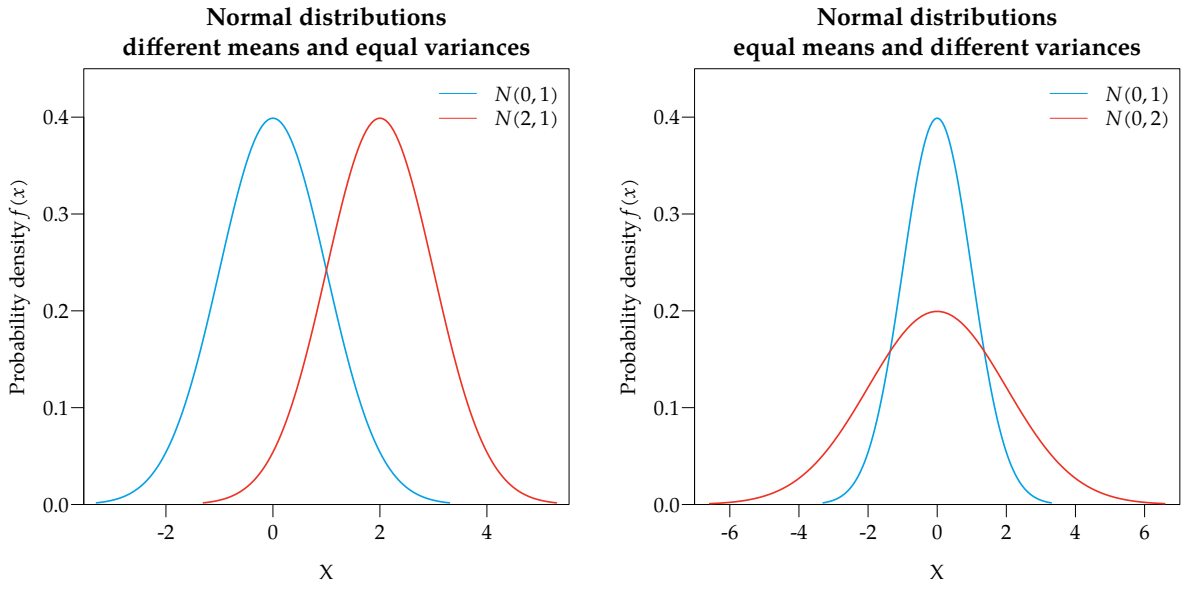
The plot of the probability density function of a normal distribution $N(\mu, \sigma)$ is bell shaped and it is known as a *Gauss bell*.



Normal probability density function

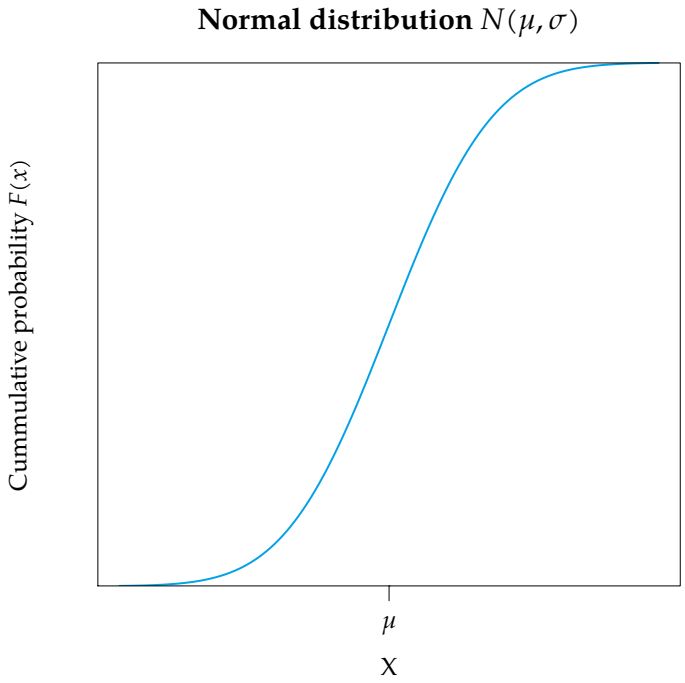
The bell shape depends on the mean μ and the standard deviation σ,

- The mean μ sets the center of the bell.
- The standard deviation sets σ the width of the bell.



Normal distribution function

The plot of the distribution function of a normal distribution is S shaped.



Normal distribution properties

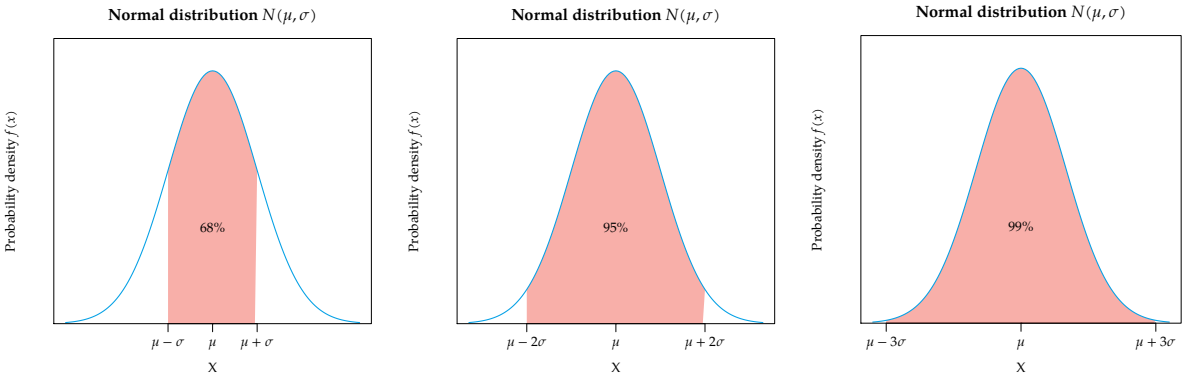
- It is symmetric with respect to the mean, and therefore, the coefficient of skewness is zero, $g_1 = 0$.
- It is mesokurtic, as the density function is bell shaped, and so, the coefficient of kurtosis is zero, $g_2 = 0$.
- The mean, median and mode are the same

$$\mu = Me = Mo.$$

- It asymptotically approaches 0 when x tends to $\pm\infty$.

Normal distribution properties

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68, P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95, P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.99.$$



Normal distribution properties

Example of cholesterol

It is known that the cholesterol level in females of age between 40 and 50 follows a normal distribution with mean 210 mg/dl and standard deviation 20 mg/dl.

According to the Gauss bell properties, this means that

- The 68% of females have a cholesterol level between 210 ± 20 mg/dl, i.e., between 190 and 230 mg/dl.
- The 95% of females have a cholesterol level between $210 \pm 2 \cdot 20$ mg/dl, i.e., between 170 and 250 mg/dl.
- The 99% of females have a cholesterol level between $210 \pm 3 \cdot 20$ mg/dl, i.e., between 150 and 270 mg/dl.

Normal distribution properties

Example blood analysis

In blood analysis it is common to use the interval $\mu \pm 2\sigma$ to detect possible pathologies. In the case of cholesterol, this interval is $[170 \text{ mg/dl}, 250 \text{ mg/dl}]$.

Thus, when a women between 40 and 50 years of age has a cholesterol level out of this interval, it's common to think about some pathology. However this person could be healthy, although the likelihood of that happening is only 5%.

BioReference LABORATORIES

Basic Lipid Evaluation				Goals	Moderate Risk	High Risk
Total Cholesterol	199	mg/dL		<200	200-239	>239
Triglycerides	209	mg/dL		<150	150-199	>199
HDL-C, Direct	29	mg/dL	(Male)	>40	40-20	<20
HDL-C as % of Total Cholesterol	15	%		>25	25-9	<9
Cholesterol / HDL-C Ratio	6.9		(Male)	<4.2	7.4-11.5	>11.5
LDL/HDL-C Ratio	4.79			<3.55	3.55-8.00	>8.00
Non-HDL Cholesterol	170	mg/dL		<130	130-189	>189
VLDL Cholesterol	42	mg/dL		<30	30-40	>40
LDL-C Direct	139	mg/dL		<100	100-160	>160
Lipid Fractions	Performed by LipoScience, Raleigh, NC			Goals	Moderate Risk	High Risk
NMR LipoProfile® Test Results						
LDL-P (Total Number of LDL Particles)	2164	nmol/l		<1000	1300-1600	>1600
Small LDL-P (Number of Small LDL Particles)	1855	nmol/L		<117	437-617	>839
LDL Particle Size	19.6	nm		LARGE - PATTERN A 23.0-20.6		SMALL - PATTERN B 20.5-18.0
HDL-P (Total HDL Particle Count)	23.6	umol/L		>40	40-23	<23
Large HDL-P (Large HDL Subclass Particles)	< 0.7	umol/L		>9.0	9.0-4.0	<4.0
Additional Risk Factors				Goals	Moderate Risk	High Risk
LP [a] -C (Lipoprotein a)	39.7	mg/dL		<20	20-30	>30
hs-CRP	0.7	mg/L		<1.0	1.0-3.0	>3.0
Lp-PLA2 (Lipoprotein associated Phospholipase)	201	ng/dL		<200	200-235	>235
Apolipoprotein A-1	116	mg/dL	(Male)	Normal Range: 104-202		
Apolipoprotein B	126	mg/dL	(Male)	Normal Range: 66-133		
Apo B/A-1 Ratio	1.09		(Male)	Normal Range: 0.45-1.25		

The central limit theorem

This behavior is common in many physical and biological variables in Nature.

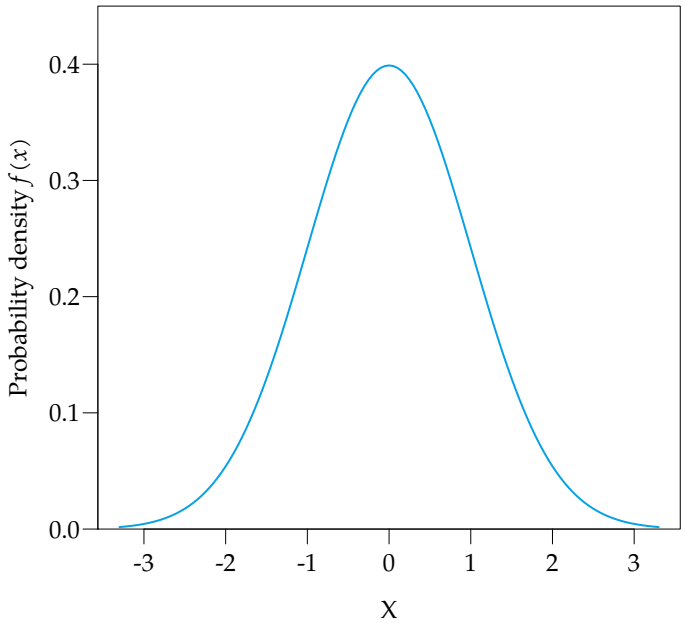
If you think about the distribution of the height, for instance, you can check that most people in the population have a height around the mean, but as the heights move away from the mean, both below and above the mean, there are few and few people with such a heights.

The explanation for this behavior is the **Central Limit Theorem**, that we will see in the next chapter; it states that a continuous random variable whose values depends on a huge number of independent factors adding their effects, always follows a normal distribution.

The standard normal distribution $N(0,1)$

The most important normal distribution has mean zero, $\mu = 0$, and standard deviation one, $\sigma = 1$. It is known as **Standard normal distribution** and usually represented as $Z \sim N(0,1)$.

Standard Normal distribution $N(0,1)$



Calculation of probabilities with the standard normal distribution

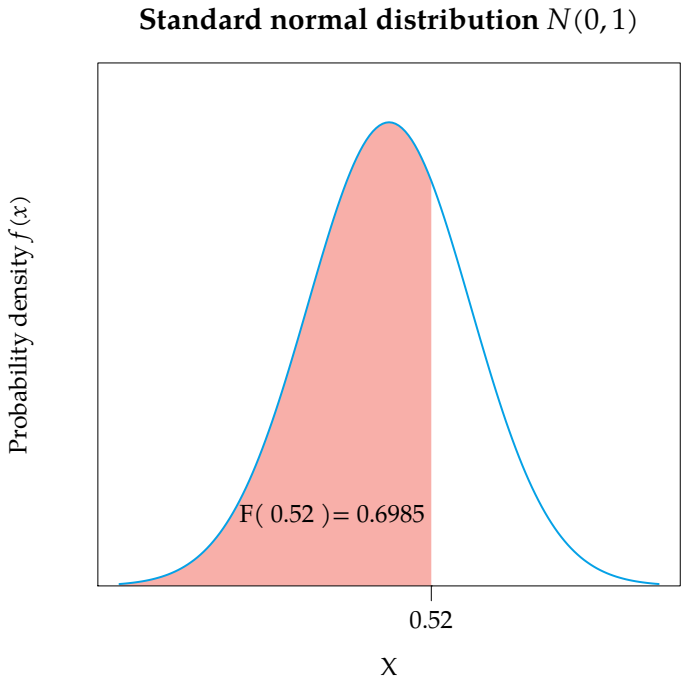
Managing the distribution function table

To avoid integrating the normal density function to compute probabilities it's common to use the distribution function, that is given in a tabular format like the one below.

For instance, to calculate $P(Z \leq 0.52)$

z	**0.00**	**0.01**	**0.02**	⋯
0.0	0.5000	0.5040	0.5080	⋯
0.1	0.5398	0.5438	0.5478	⋯
0.2	0.5793	0.5832	0.5871	⋯
0.3	0.6179	0.6217	0.6255	⋯
0.4	0.6554	0.6591	0.6628	⋯
0.5	0.6915	0.6950	**0.6985**	⋯
⋮	⋮	⋮	⋮	⋱

$0.52 \rightarrow$ row 0.5 + column 0.02

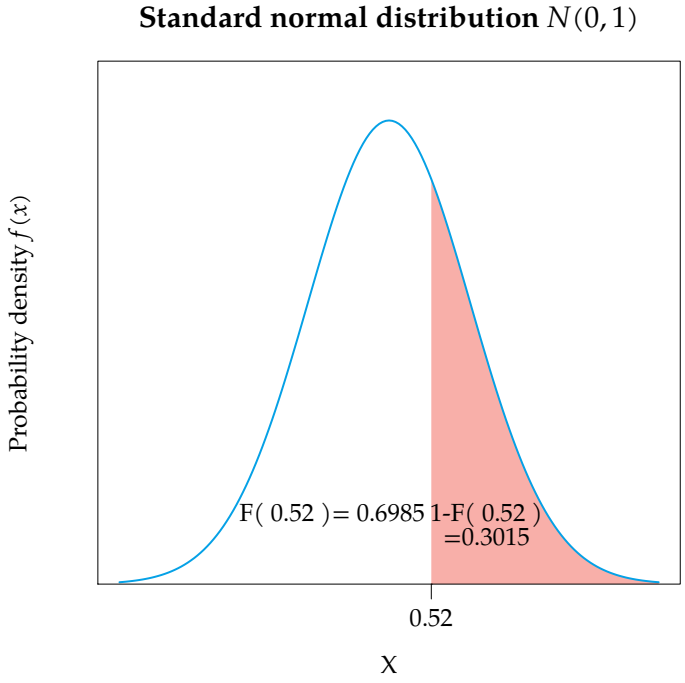


Calculation of probabilities with the standard normal distribution

Right tail probabilities

To compute cumulative probabilities to the right of a value, we can apply the rule for the complement event. For instance,

$$P(Z > 0.52) = 1 - P(Z \leq 0.52) = 1 - F(0.52) = 1 - 0.6985 = 0.3015.$$



Standardization

We have seen how to use the table of the standard normal distribution function to compute probabilities, but, *what to do when the normal distribution is not the standard one?*

In that case we can use standardization to transform any normal distribution in the standard normal distribution.

Theorem 71 (Standardization)**.** *If X is a continuous random variables that follow a Normal probability distribution model with mean μ and standard deviation σ, $X \sim N(\mu, \sigma)$, then the variable that result of subtracting μ to X and dividing by σ, follows a Standard Normal probability distribution,*

$$X \sim N(\mu, \sigma) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Thus, to compute probabilities with a non-standard Normal distribution first we have to standardize the variable before using the table of the standard Normal distribution function.

Standardization

Example

Assume that the grade of an exam X follows a Normal probability distribution model $N(\mu = 6, \sigma = 1.5)$. What percentage of students didn't pass the exam?

As X follows a non-standard Normal distribution model, we have to apply standardization first, $Z = \dfrac{X - \mu}{\sigma} = \dfrac{X - 6}{1.5}$,

$$P(X < 5) = P\left(\frac{X - 6}{1.5} < \frac{5 - 6}{1.5}\right) = P(Z < -0.67).$$

Then we can use the table of the standard Normal distribution function,

$$P(Z < -0.67) = F(-0.67) = 0.2514.$$

Therefore, 25.14% of students didn't pass the exam.

7.4 Chi-square distribution

Chi-square probability distribution model $\chi^2(n)$

Definition 72 (Chi-square distribution $\chi^2(n)$)**.** Given n independent random variables $Z_1, \dots, Z_n$, all of them following a standard normal probability distribution, then the variable

$$\chi^2(n) = Z_1^2 + \cdots + Z_n^2,$$

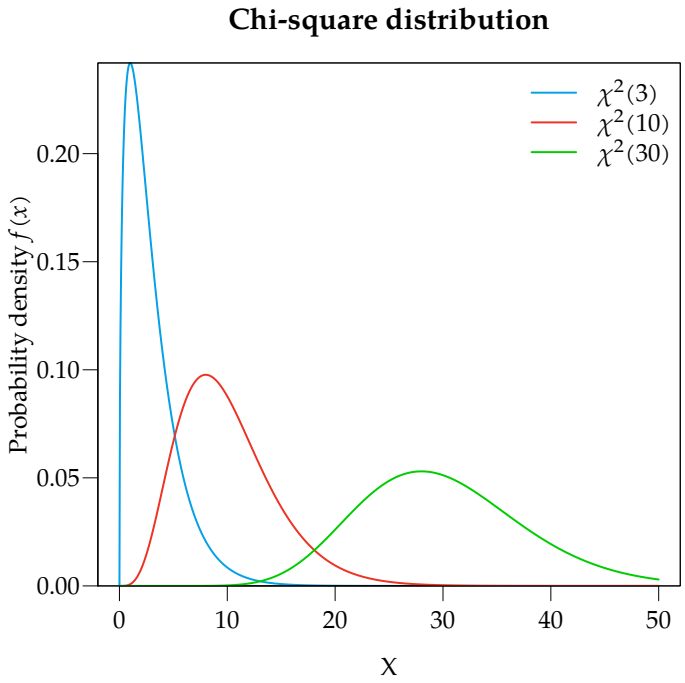
follows a *chi-square probability distribution with n degrees of freedom*.

Its range is $\mathbb{R}^+$ and its mean and variance are

$$\mu = n, \qquad \sigma^2 = 2n.$$

As we will see in the next chapter, the chi-square distribution plays an important role in the estimation of the population variance and in the study of relations between qualitative variables.

Chi-square probability density function



Chi-square distribution properties

- The range is non-negative.
- If $X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$, then
$$X + Y \sim \chi^2(n+m).$$
- It asymptotically approaches to a normal distribution as the degrees of freedom increase.

As we will see in the next chapter, the chi-square distribution plays an important role in the estimation of the population variance and in the study of relations between qualitative variables.

7.5 Student's t distribution

Student's t probability distribution $T(n)$

Definition 73 (Student's t distribution $T(n)$)**.** Given a variable Z following a standard normal distribution model, $Z \sim N(0,1)$, and a variable X following a chi-square distribution model with n degrees of freedom, $X \sim \chi^2(n)$, independent, the variable

$$T = \frac{Z}{\sqrt{X/n}},$$

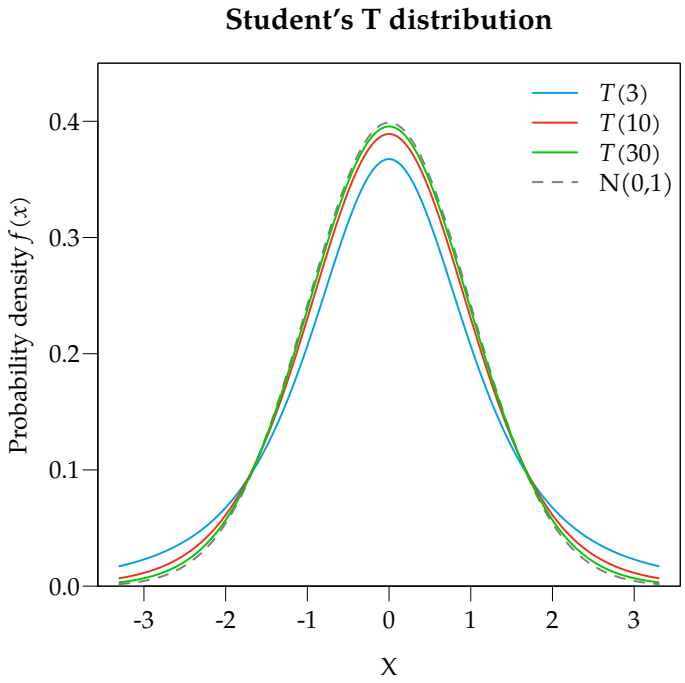
follows a *Student's t probability distribution model with n degrees of freedom*.

Its range is $\mathbb{R}$ and its mean and variance are

$$\mu = 0, \qquad \sigma^2 = \frac{n}{n-2} \text{ if } n > 2.$$

As we will see in the next chapter, the Student's t distribution plays an important role in the estimation of the population mean.

Student's t probability density function



Student's t distribution properties

- The mean, the median and the mode are the same, $\mu = Me = Mo$.
- It is symmetric, $g_1 = 0$.
- It asymptotically approaches to the standard normal distribution as the degrees of freedom increase. In practice for $n \geq 30$ both distributions are approximately the same.

$$T(n) \overset{n \to \infty}{\approx} N(0,1).$$

As we will see in the next chapter, the Student's t distribution plays an important role in the estimation of the population mean.

7.6 Fisher-Snedecor's F distribution

Fisher-Snedecor's F probability distribution $F(m,n)$

Definition 74 (Fisher-Snedecor's F distribution $F(m,n)$)**.** Given two independent variables X and Y both following a chi-square probability distribution model with m an n degrees of freedom respectively, $X \sim \chi^2(m)$ and $Y \sim \chi^2(n)$, then the variable

$$F = \frac{X/m}{Y/n},$$

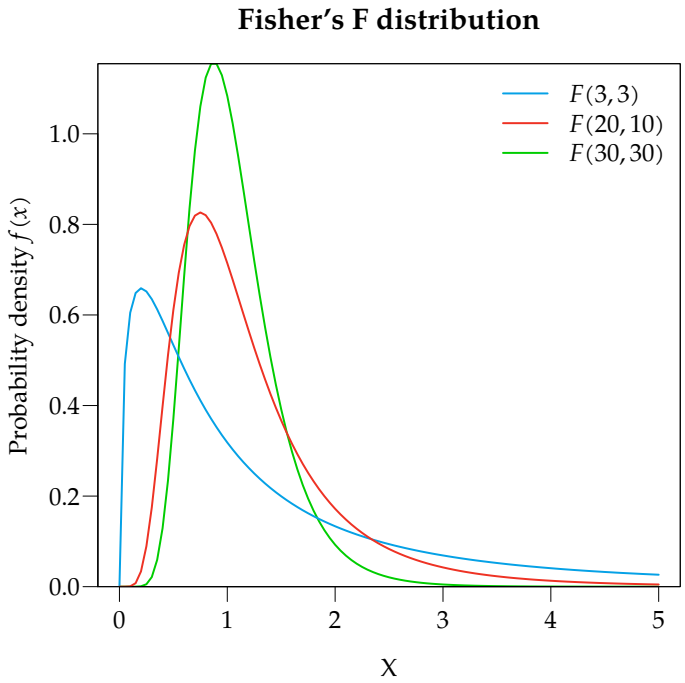
follows a *Fisher-Snedecor's F probability distribution model with m and n degrees of freedom*.

Its range is $\mathbb{R}^+$ and its mean and variance are

$$\mu = \frac{n}{n-2}, \qquad \sigma^2 = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \text{ if } n > 4.$$

As we will see in the next chapter, the Fisher-Snedecor's F distribution plays an important role in the comparison of population variances and in the analysis of variance test (ANOVA).

Fisher-Snedecor's F probability density function



Fisher-Snedecor's F distribution properties

- The range is non-negative.
- It satisfies

$$F(m,n) = \frac{1}{F(n,m)}.$$

Thus, if we name $f(m,n)_p$ the value that satisfies $P(F(m,n) \leq f(m,n)_p) = p$, then

$$f(m,n)_p = \frac{1}{f(n,m)_{1-p}}$$

which is helpful in order to compute probabilities from the table of the distribution function.