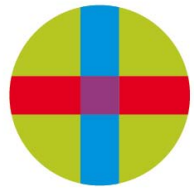
# Elementary Statistics Course

Alfredo Sánchez Alberca (asalber@ceu.es)



CEU
Universidad
San Pablo

# License

**Elementary Statistics Course**

Alfredo Sánchez Alberca (asalber@gmail.com).

# Contenidos

# Introduction to Statistics

# What is Statistics?

## Definition (Statistics)

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

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


Data

# What is Statistics?

## Definition (Statistics)

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

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

# What is Statistics?

## Definition (Statistics)

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

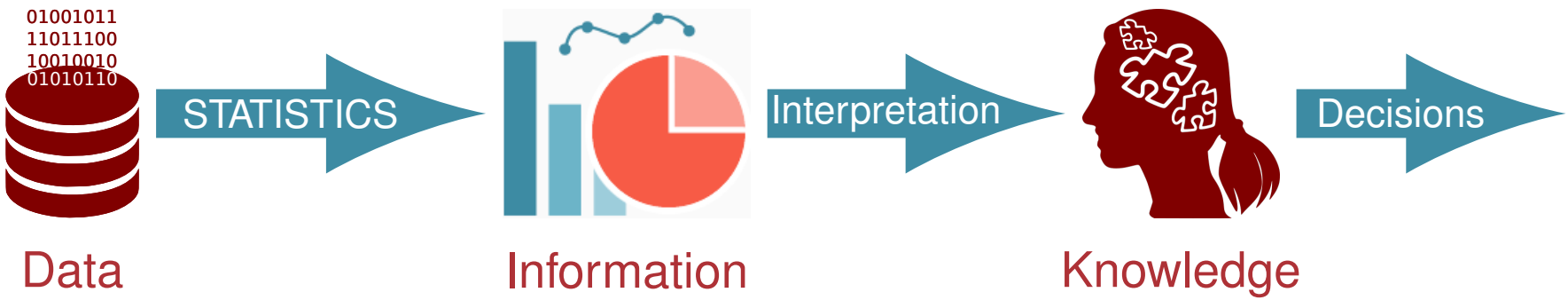
The role of Statistics is to extract information from data in order to gain knowledge for taking decisions.

# What is Statistics?

## Definition (Statistics)

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

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

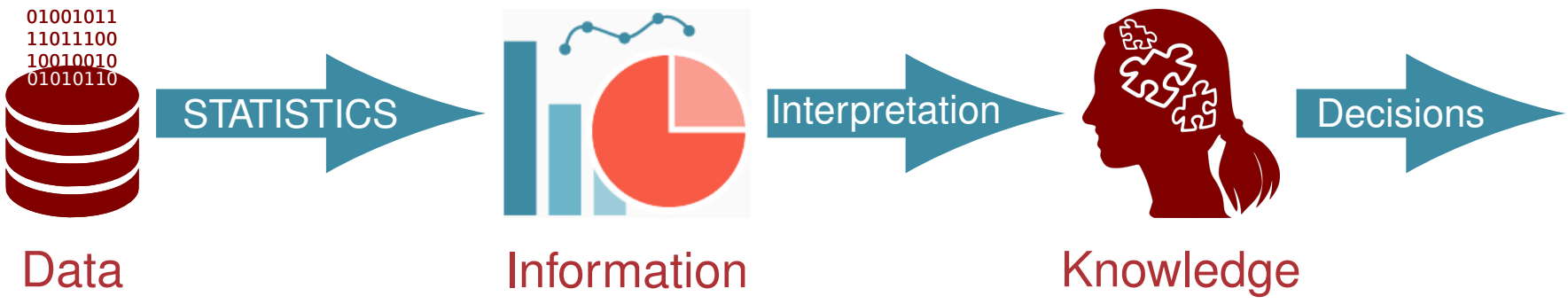


# What is Statistics?

## Definition (Statistics)

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

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



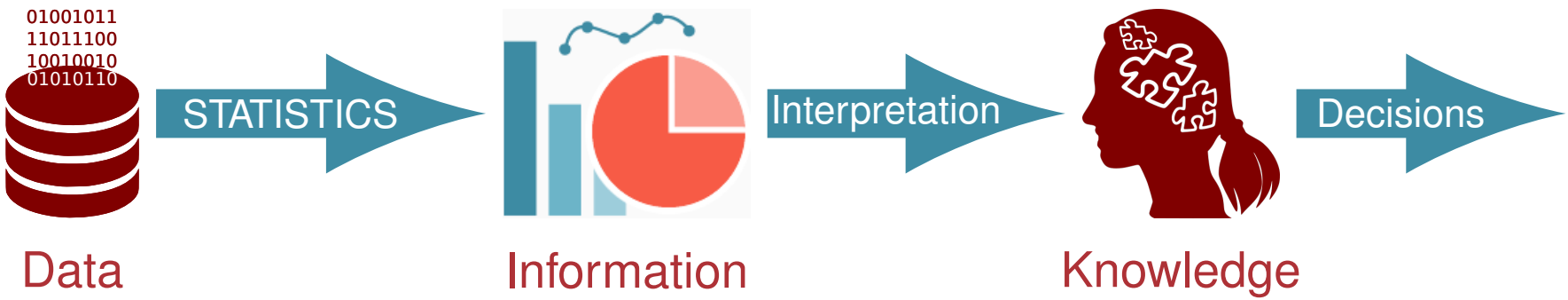

Statistics is essential in any scientific or technical discipline which require data handling, especially with large volumes of data, such as Physics, Chemistry, Medicine, Psychology, Economics or Social Sciences.

# What is Statistics?

## Definition (Statistics)

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

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




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

But, *Why is necessary Statistics?*

# A changing World

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

*Variability is the reason for Statistics!*

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

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

# Statistical population

## Definition (Population)

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

## Definition (Population size)

The number of individuals in a population is known as the *population size* and is represented by $N$.

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

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

Studied population: Individuals truly accessible in the study.

# Drawbacks in the population study

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

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

However, this is not always possible for several reasons:

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

# Statistics Sample

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

## Definition (Sample)

A *sample* is a subset of the population.

## Definition (Sample size)

The number of individuals of the sample is called *sample size* and is represented by $n$.

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

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

# Sample size determination

One of the most interesting questions that arise:

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

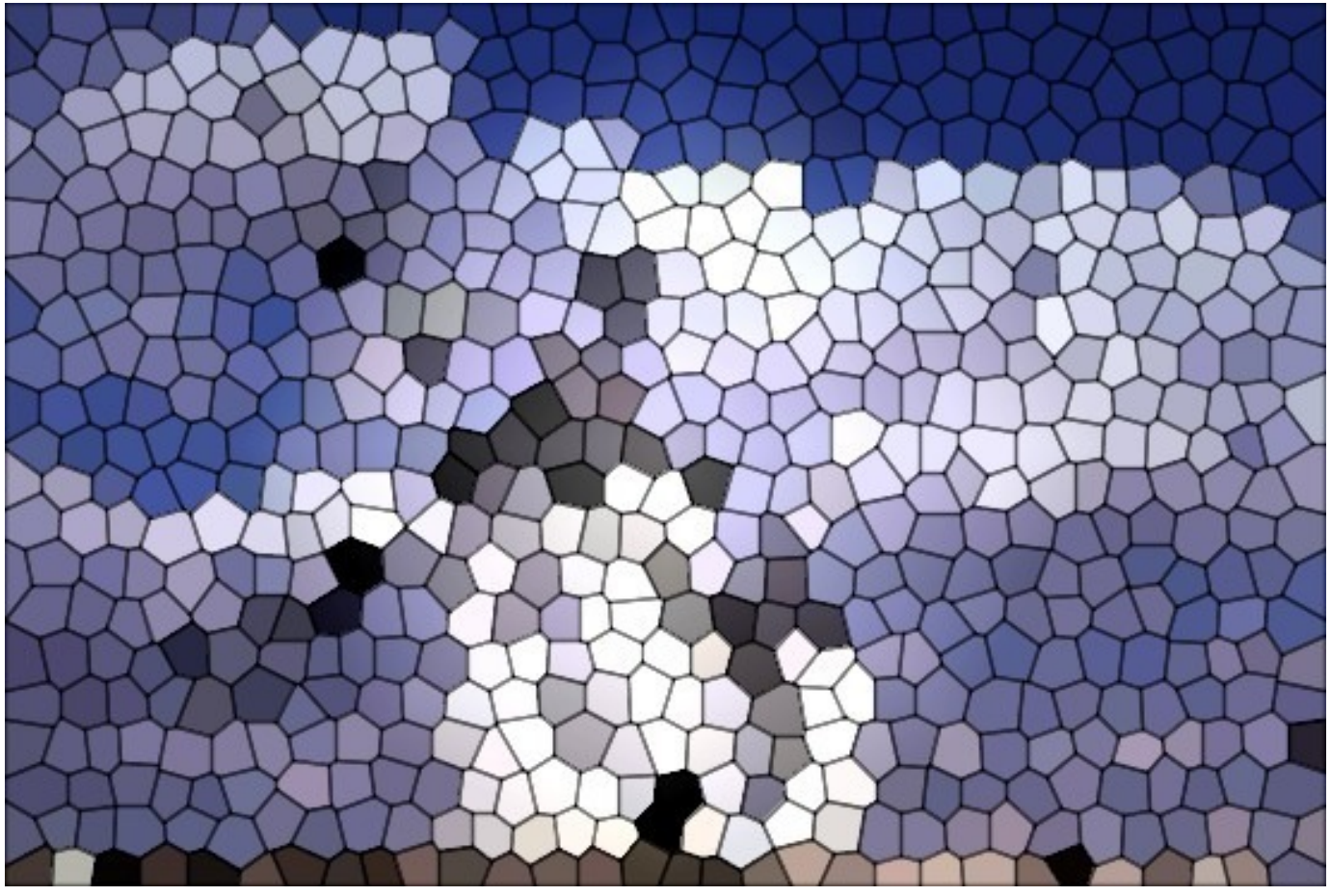
The answer depends of several factors, as the population variability or the desired reliability for extrapolations on the population.

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

# Sample size determination

## Small sample of pixels of a picture

What picture is it?

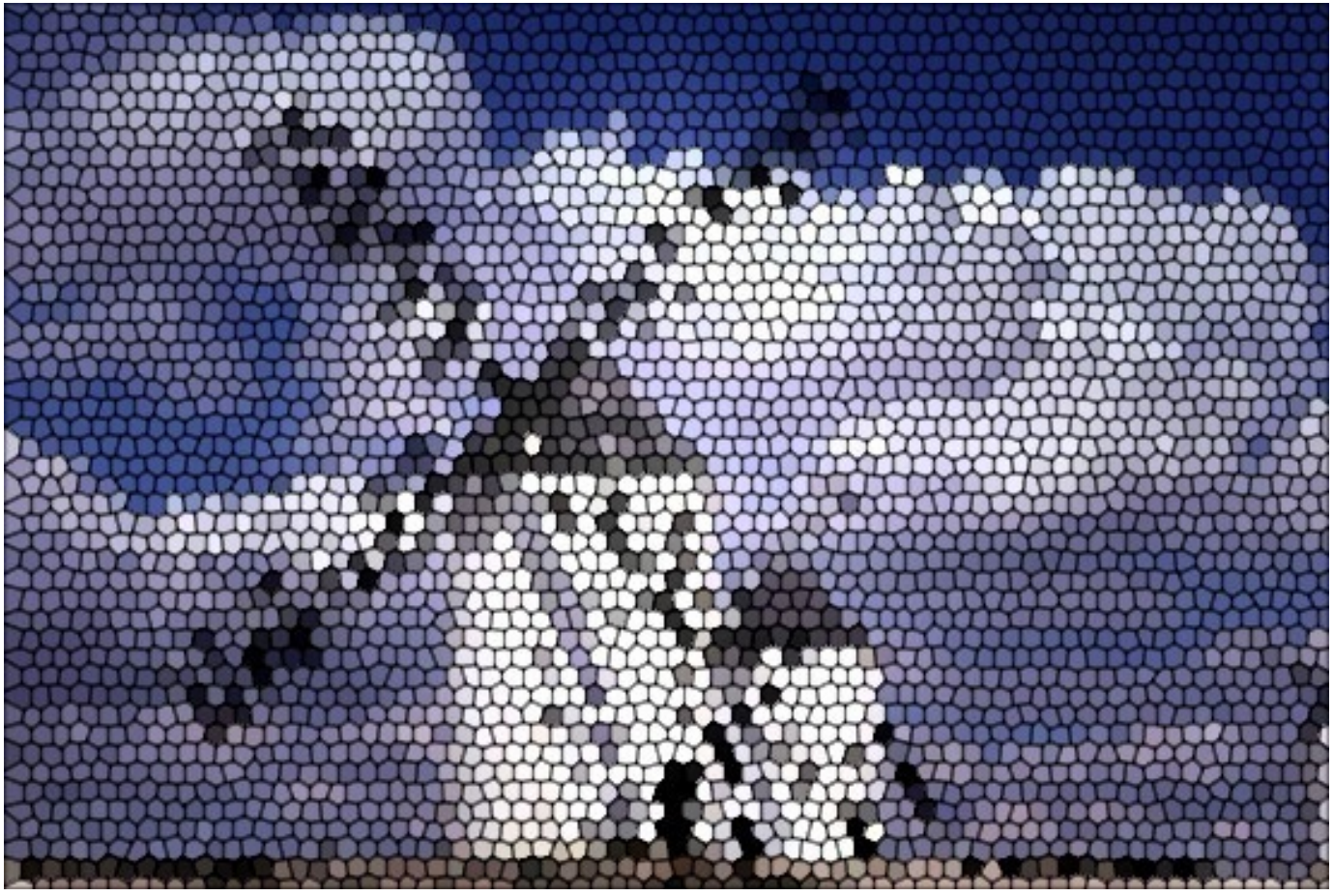


*With a small sample size it's difficult to find the image content out!*

# Sample size determination

## Large sample of pixels of a picture

What picture is it?



*With a large sample is easier to find the image content out!*

# 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 their content out!*

# Types of reasoning

Population

From general to particular

Deduction

Sample

# Types of reasoning

Population

From general to particular

Deduction

Induction

From particular to general

Sample

# Types of reasoning

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

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

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

# Sampling

## Definition (Sampling)

The process of selecting the elements included in a sample is known as *sampling*.


Population Sample

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

*Our goal is to get a representative sample!*

# Types of sampling

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

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

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

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

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

# Simple random sampling

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

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

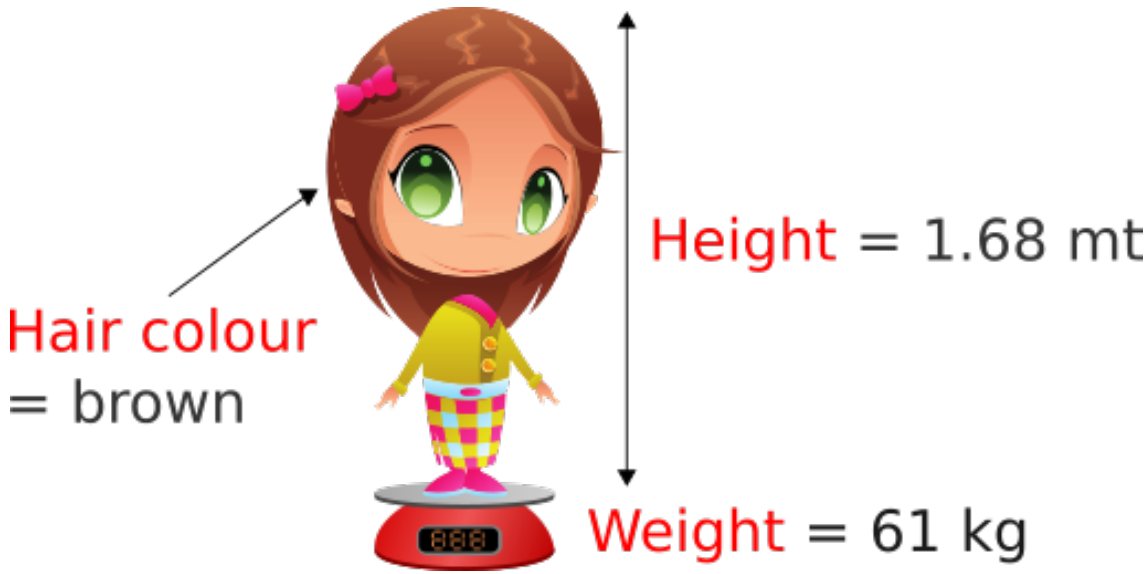
The only way of doing a random sampling is to assign a unique identity number to each population individual (conducting a *census*) and performing a random drawing.

# Statistical variables and data

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

## Definition (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.



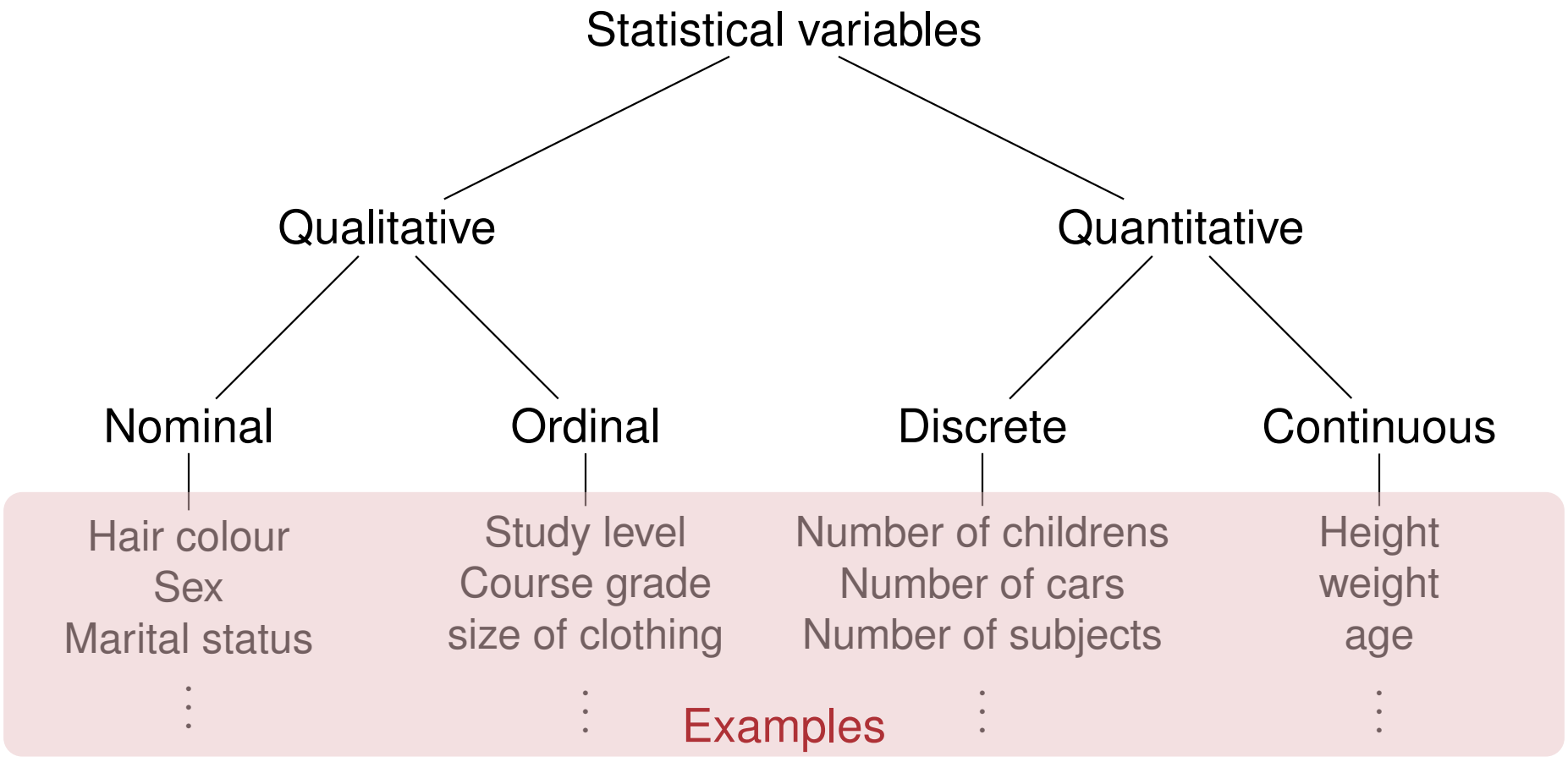
# Types of statistical variables

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 the categories.
    Example: The eyes or hair colour.
  - Ordinal: There is a natural order between the categories.
    Example: The education level.
- Quantitative variables: They measure numeric quantities. They can be:
  - Discrete: Their values are isolated numbers (usually integers).
    Example: The number of children or cars in a family.
  - Continuous: They can take any value in a real interval.
    Example: The height, weight or age of a person.

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

# Types of statistical variables



# Types of statistical variables

## Choosing the appropriate variable

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

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

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

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

# Types of statistical variables

According to their role in the study:

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

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

# Types of statistical studies

According to their role in the study:

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

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

# The data table

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

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

**Example**

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

# Phases of a statistical study

Usually a statistical study goes through the following phases:

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

# The statistical cycle

Population

# The statistical cycle

Sample

Sampling

Population

# The statistical cycle

Sample

Descriptive

Summary measures

$\bar{x}$ $s^2$

$p$ $g_1$

Sampling

Population

# The statistical cycle

Sample

Summary measures

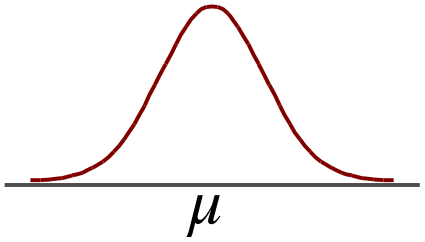
Descriptive

$\bar{x}$ $s^2$

$p$ $g_1$

Sampling

Inferential




Population

Model

# The statistical cycle

Sample

Descriptive

Summary measures

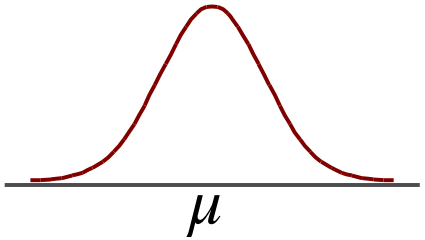
$\bar{x}$ $s^2$

$p$ $g_1$

Inferential

Sampling

Prediction




Population

Model

# Frequency distribution: Tabulation and charts

## 2. Frequency distributions: Tabulation and charts

2.1 Frequency distribution

2.2 Frequency distribution graphs

2.3 Estadísticos muestrales

2.4 Central tendency statistics

# Descriptive Statistics

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

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

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

It has no inferential power $\Rightarrow$ *Do not generalize to the population!*

# Sample classification

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

There are two ways of classifying data:

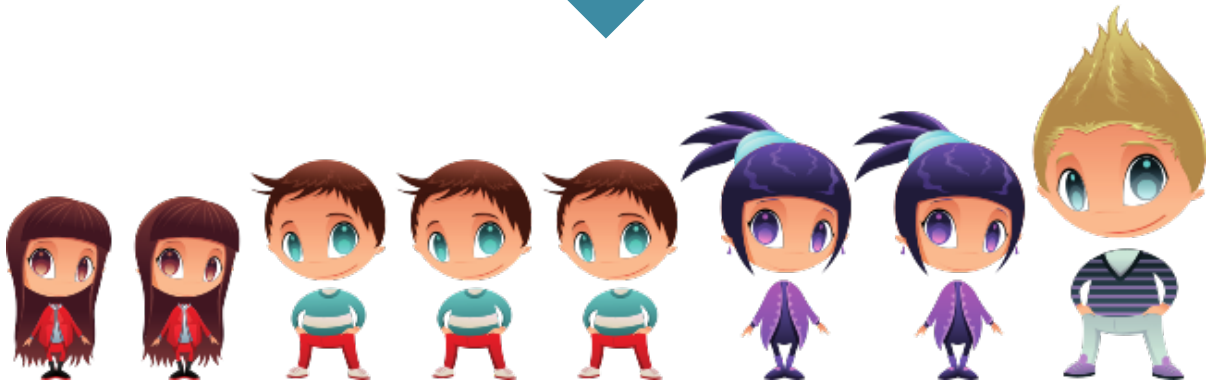
**Non-grouping** Sort values from lowest to highest value (if there is an order). Used with qualitative variables and discrete variables with few distinct values.

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

# Sample classification

$X$ =Height

Classify



# Frequency count

$X$ =Height

Frequency count

# Sample frequencies

## Definition (Sample frequencies)

Given a sample of $n$ values of a variable $X$, for every value $x_i$ of the variable is defined

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

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

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

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

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

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

# Frequency table

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

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

# Frequency table

## Example of quantitative variable and non-grouped data

The number of children in 25 families are:

1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2,
0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2

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

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

# Frequency table

## Example of quantitative variable and grouped data

The heights (in cm) of 30 students are:

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

The frequency table for the height in this sample is

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

# Classes construction

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

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

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

# Frequency table

## Example with qualitative variable

The blood type of 30 people are:

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

The frequency table of the blood type is

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

*Why there are not cumulative frequencies?*

# Frequency distribution graphs

Usually the frequency distribution is also displayed graphically.

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

- Bar chart
- Histogram
- Line chart or ogive.
- Pie chart

# Bar chart

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

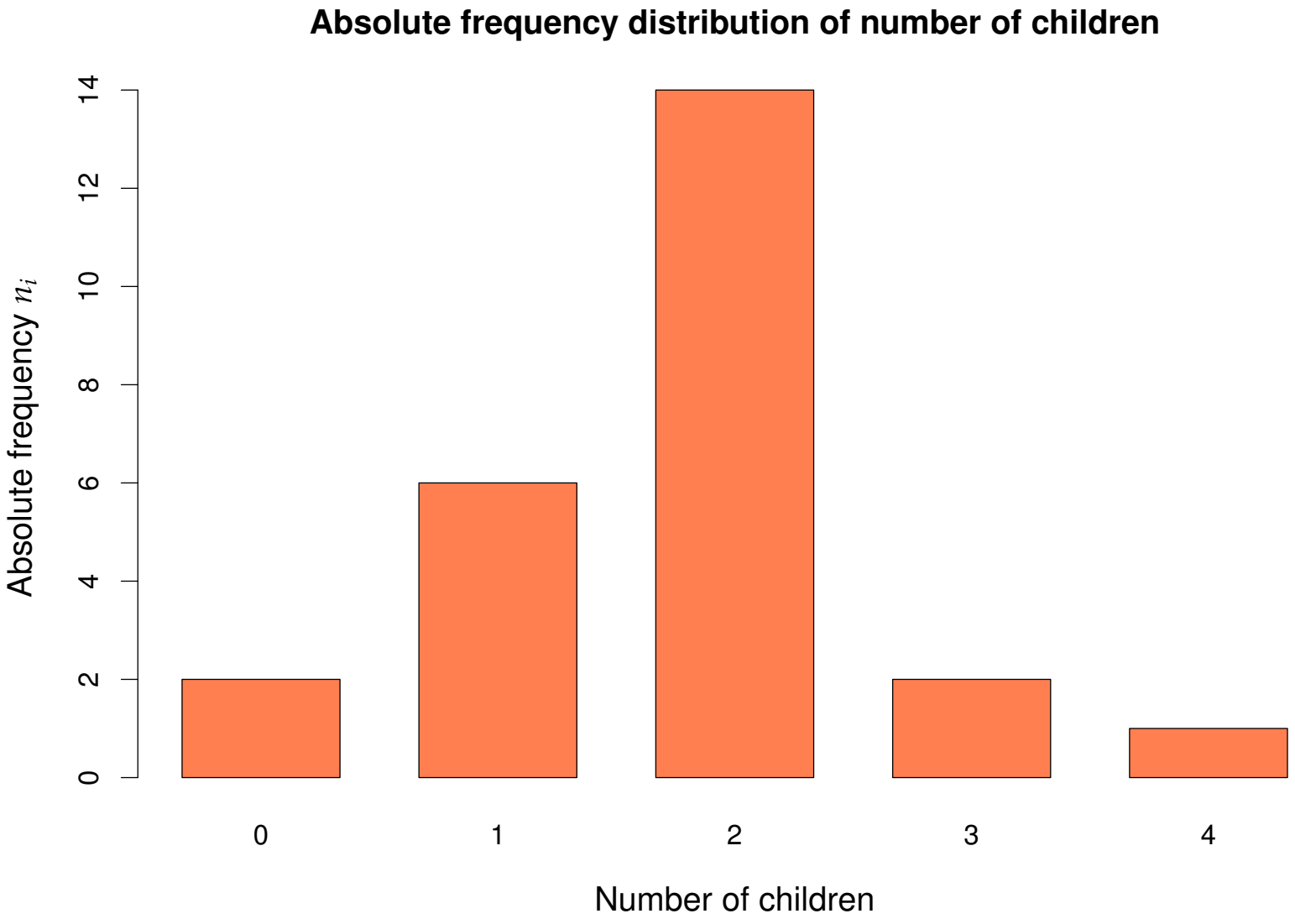
Usually the values or categories of the variable are represented on the $x$-axis, and the frequencies on the $y$-axis. For each value or category of the variable, a bar is draw to the height of its frequency. The width of the bar is not important but bars should be clearly separated among them.

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

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

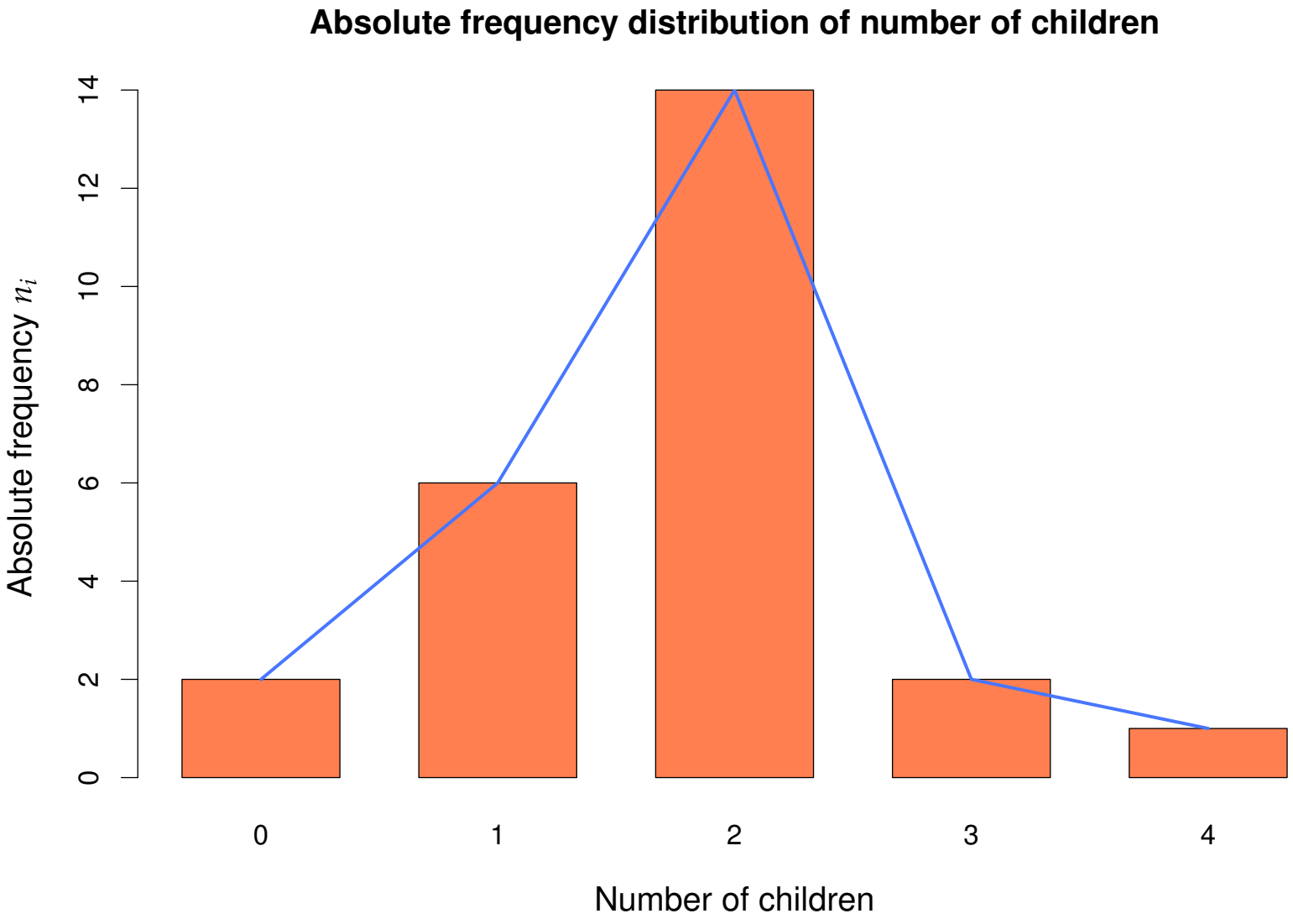
# Absolute frequency bar chart

## Non-grouped data



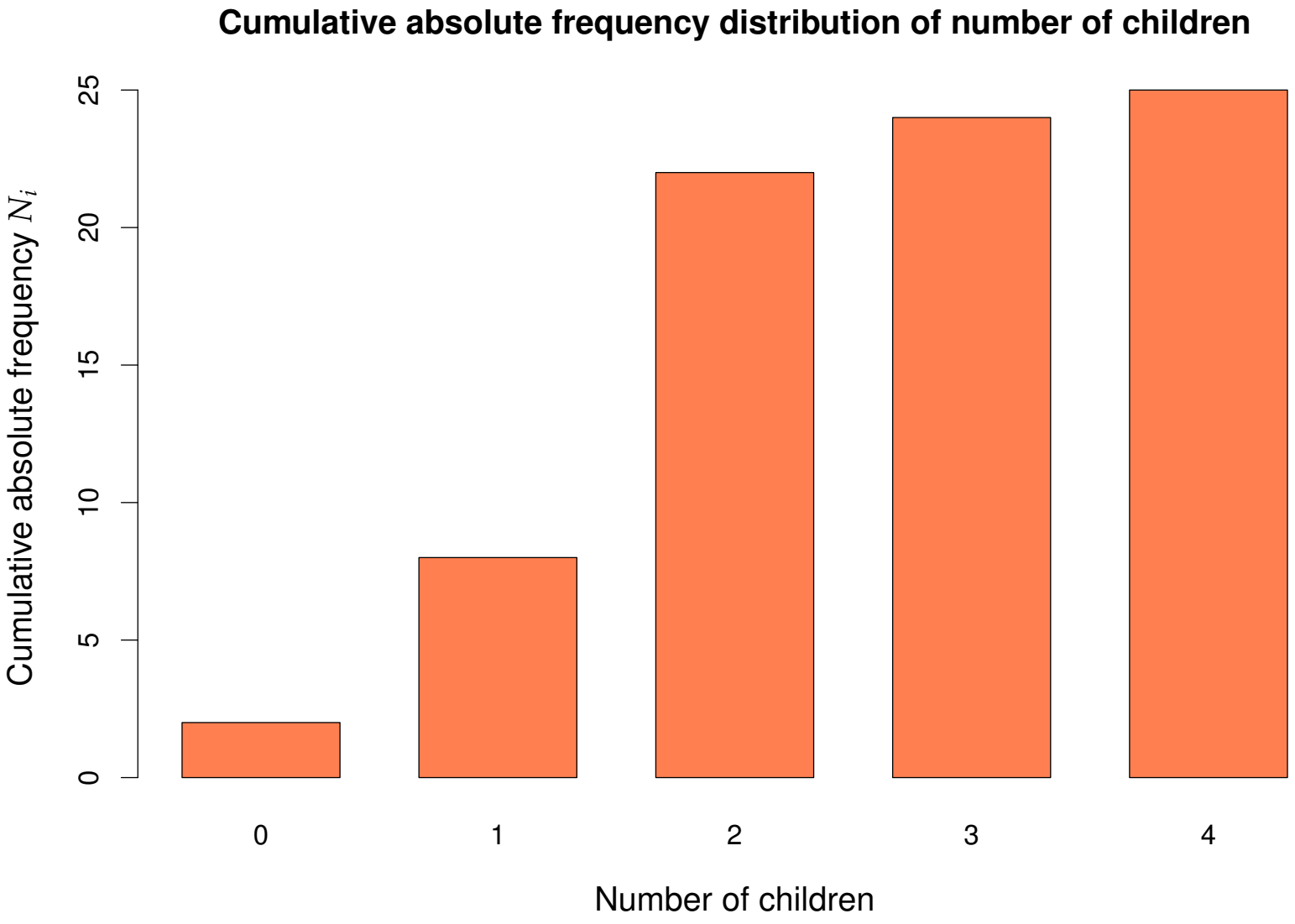
# Absolute frequency line chart or polygon

Non-grouped data



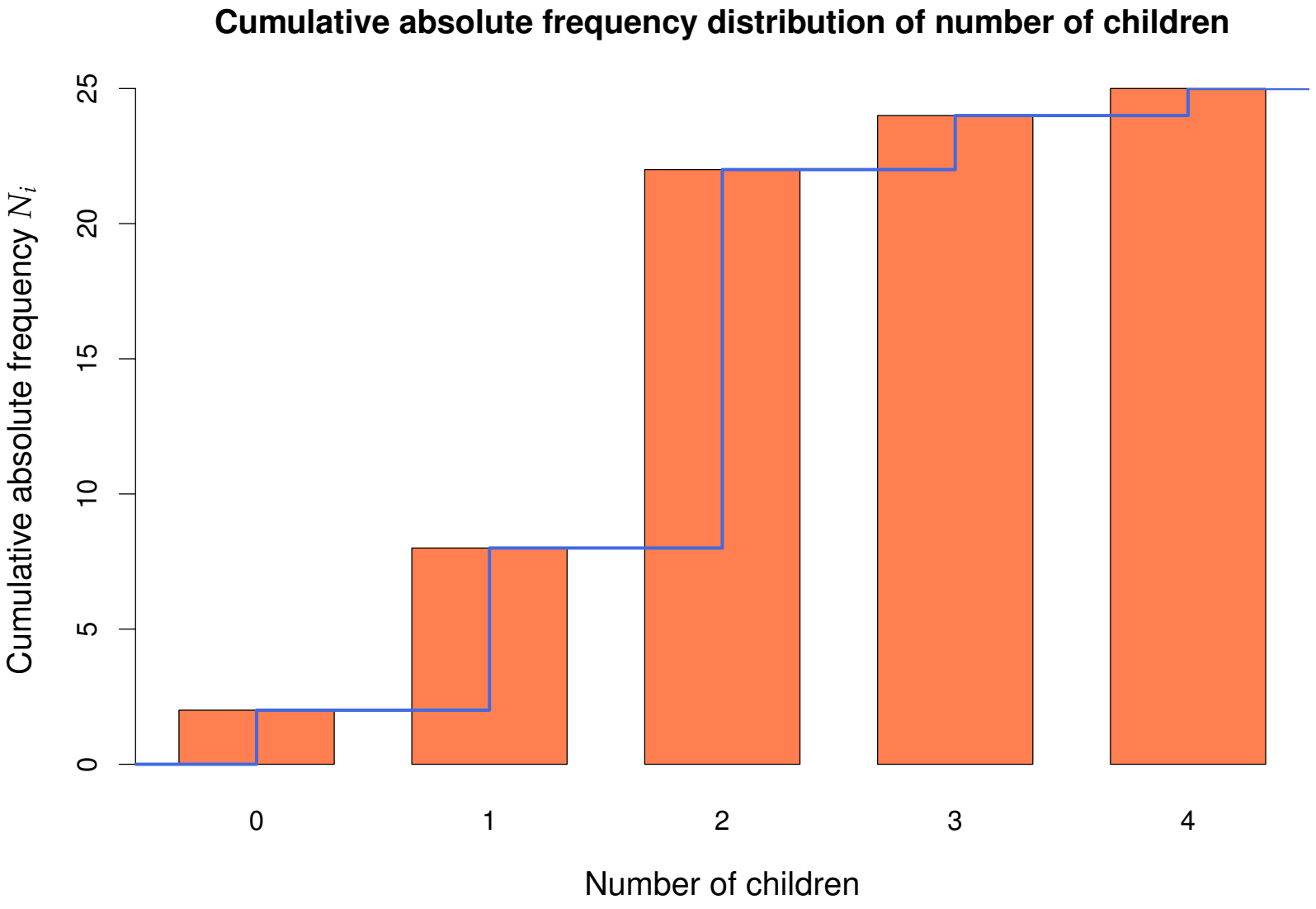
# Cumulative absolute frequency bar chart

## Non-grouped data



# Cumulative absolute frequency line chart or polygon

Non-grouped data



# Histogram

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

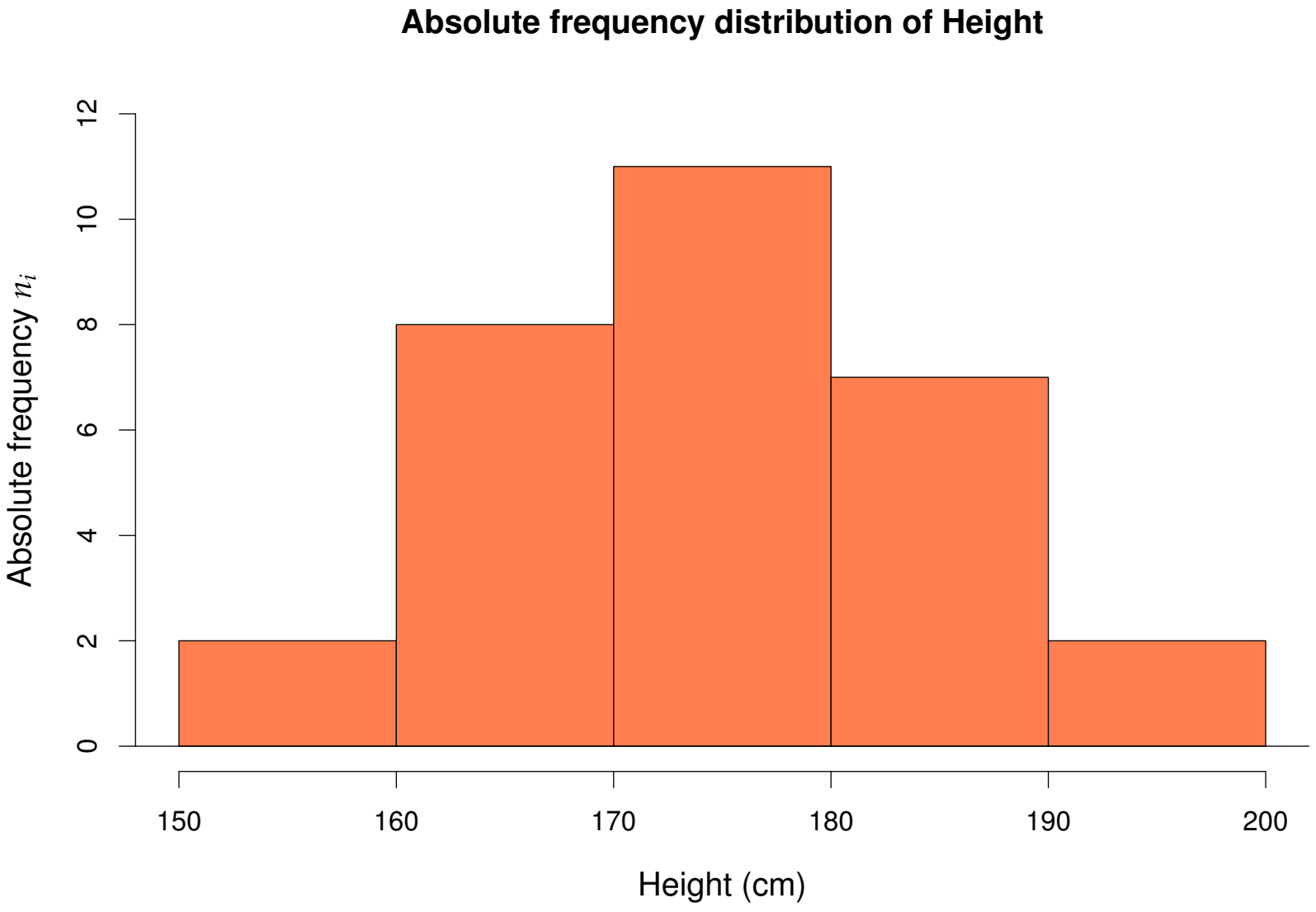
Usually the classes or grouping intervals are represented on the $x$-axis, and the frequencies on the $y$-axis. For each class, a bar is draw to the height of its frequency. Contrary to bar charts, the width of bars coincides with the width of classes, and there are no space between two consecutive bars.

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

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

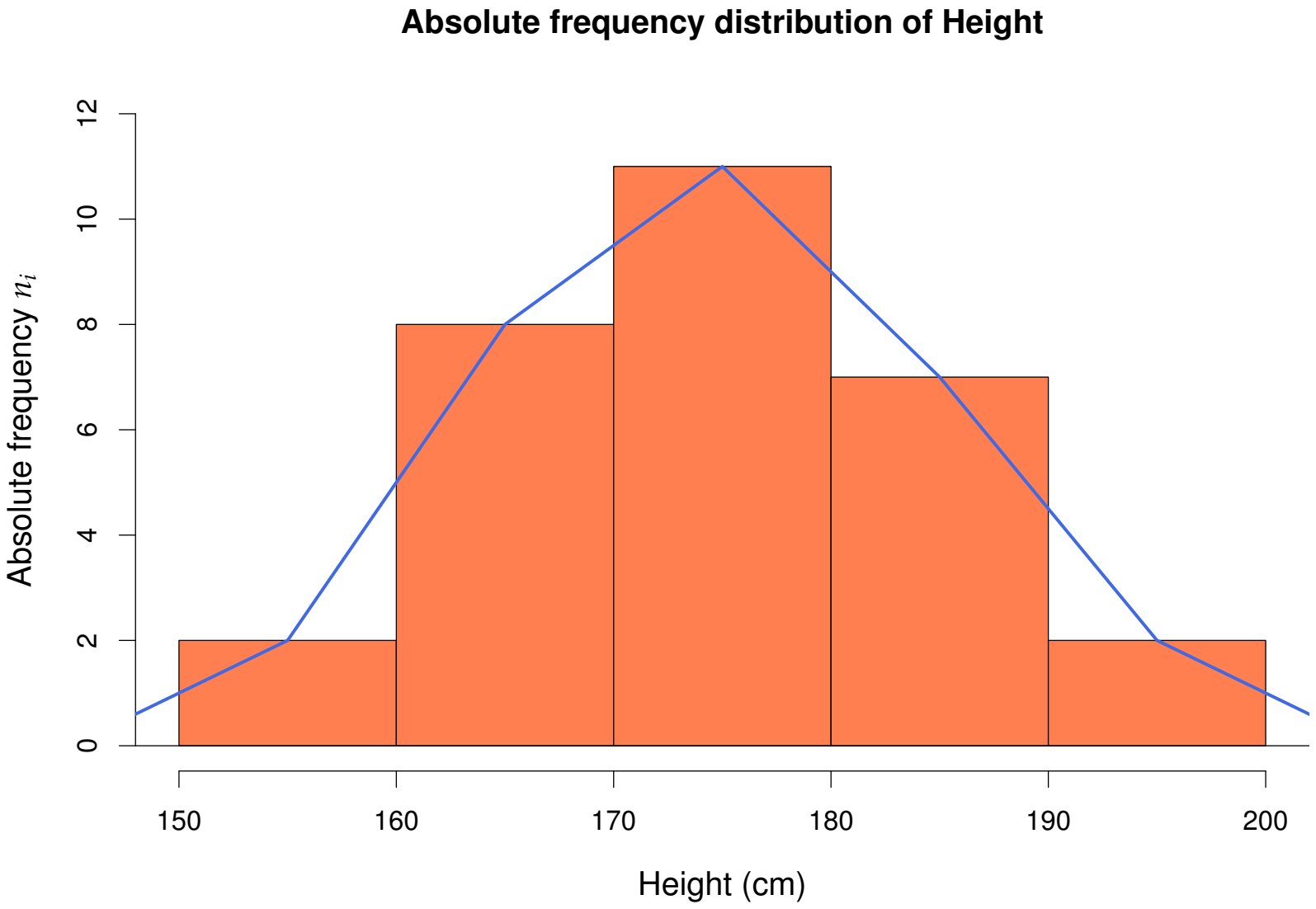
# Absolute frequency histogram

## Grouped data



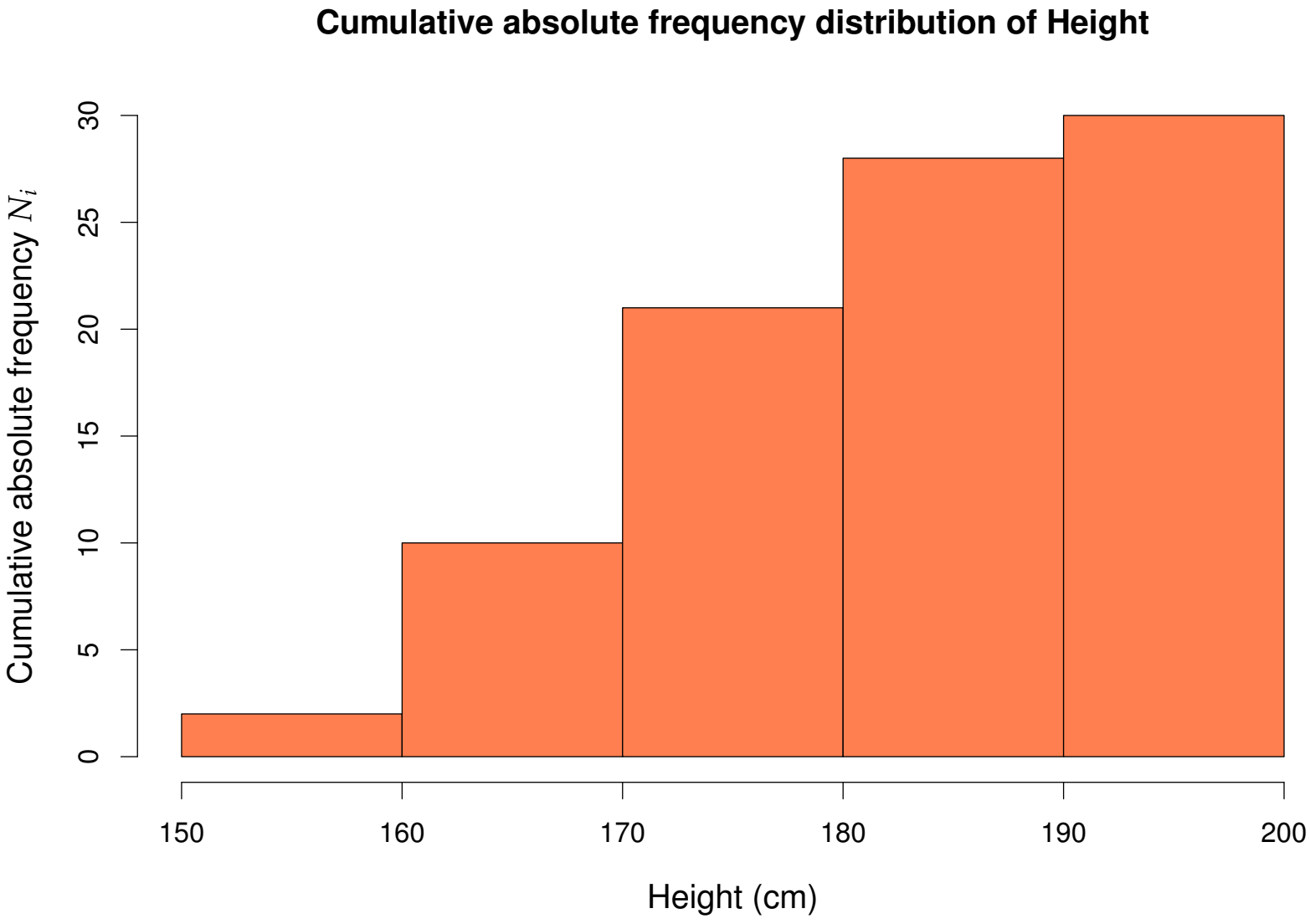
# Absolute frequency histogram

## Grouped data



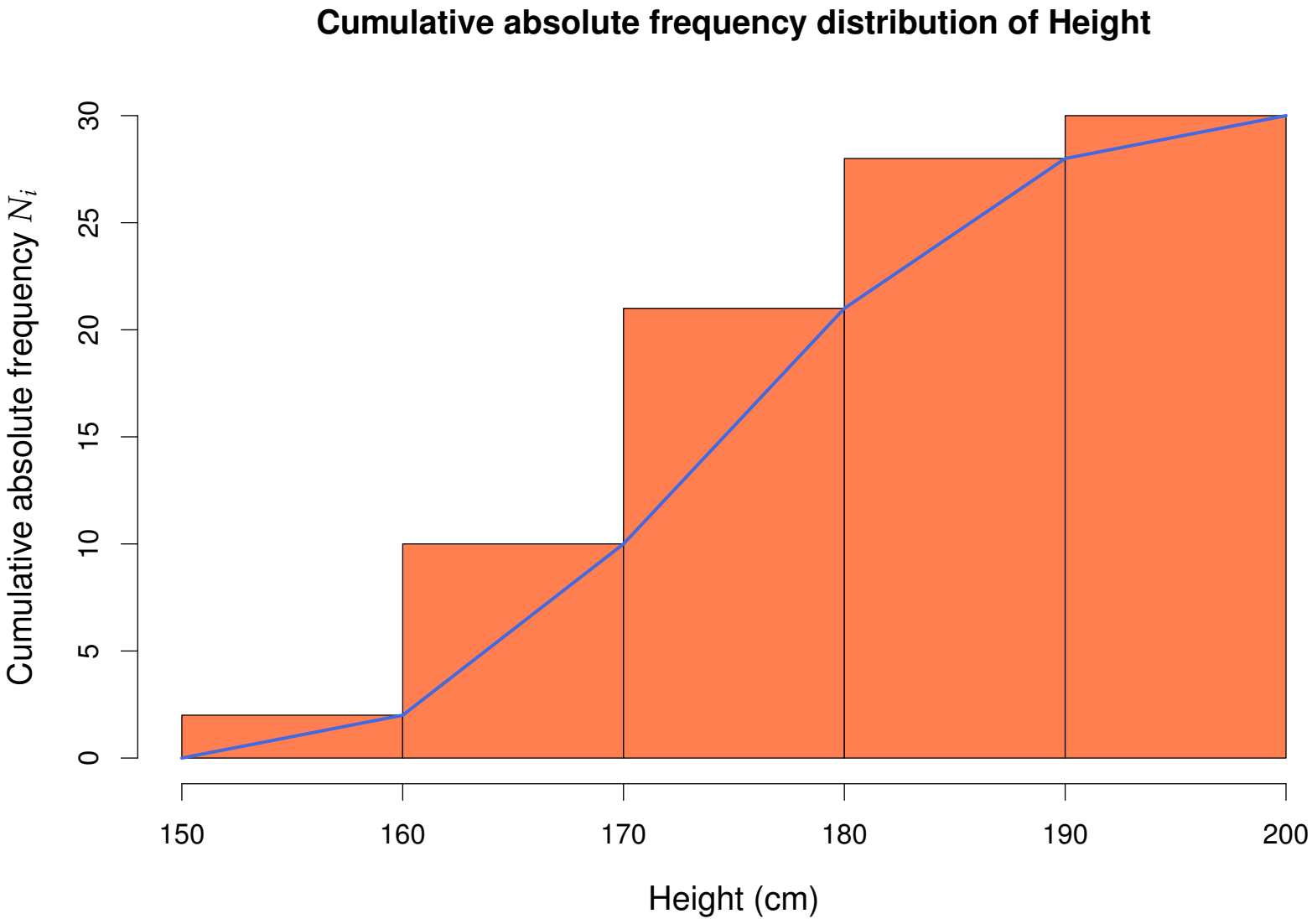
# Cumulative absolute frequency histogram

## Grouped data



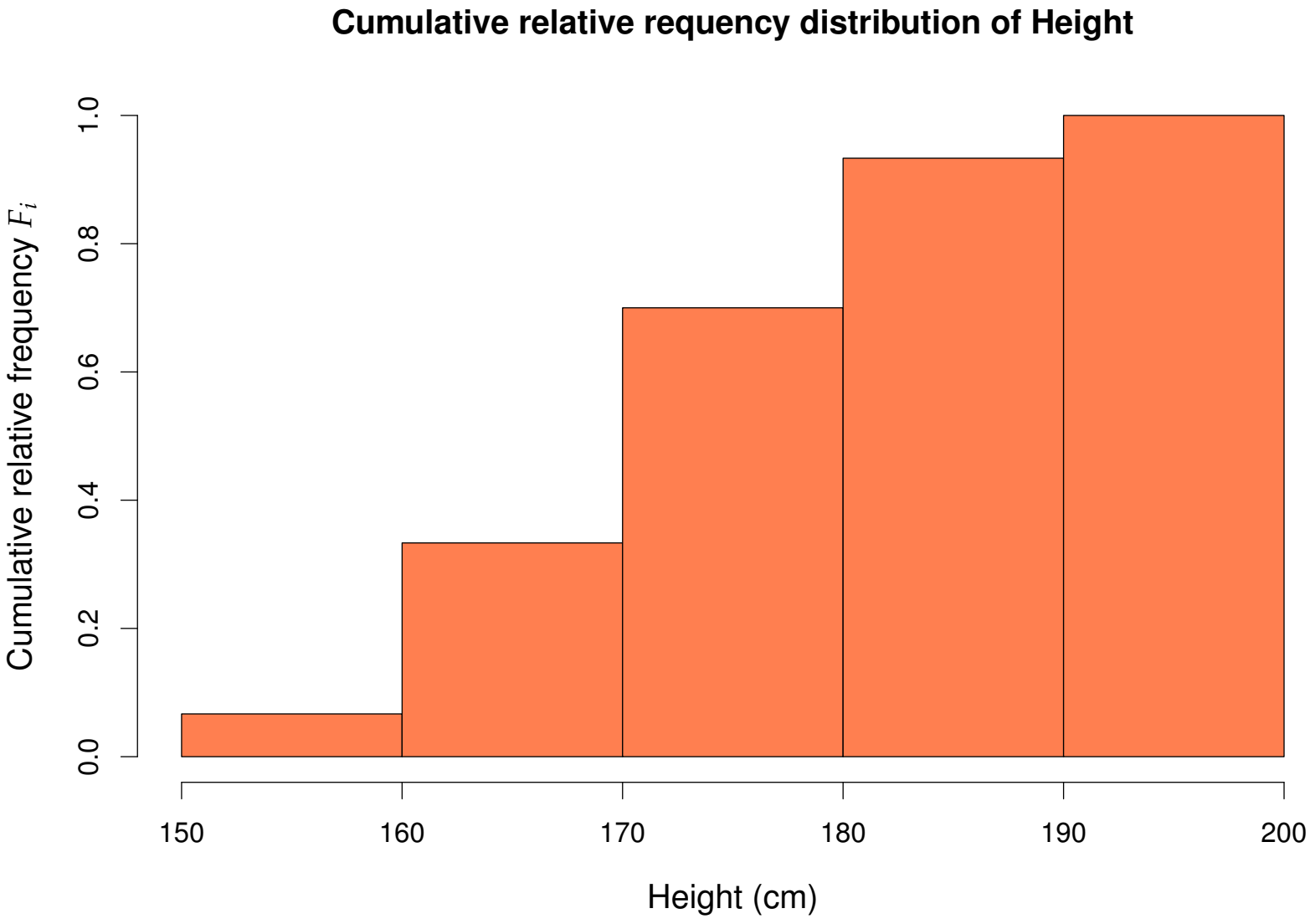
# Cumulative absolute frequency line chart or ogive

## Grouped data



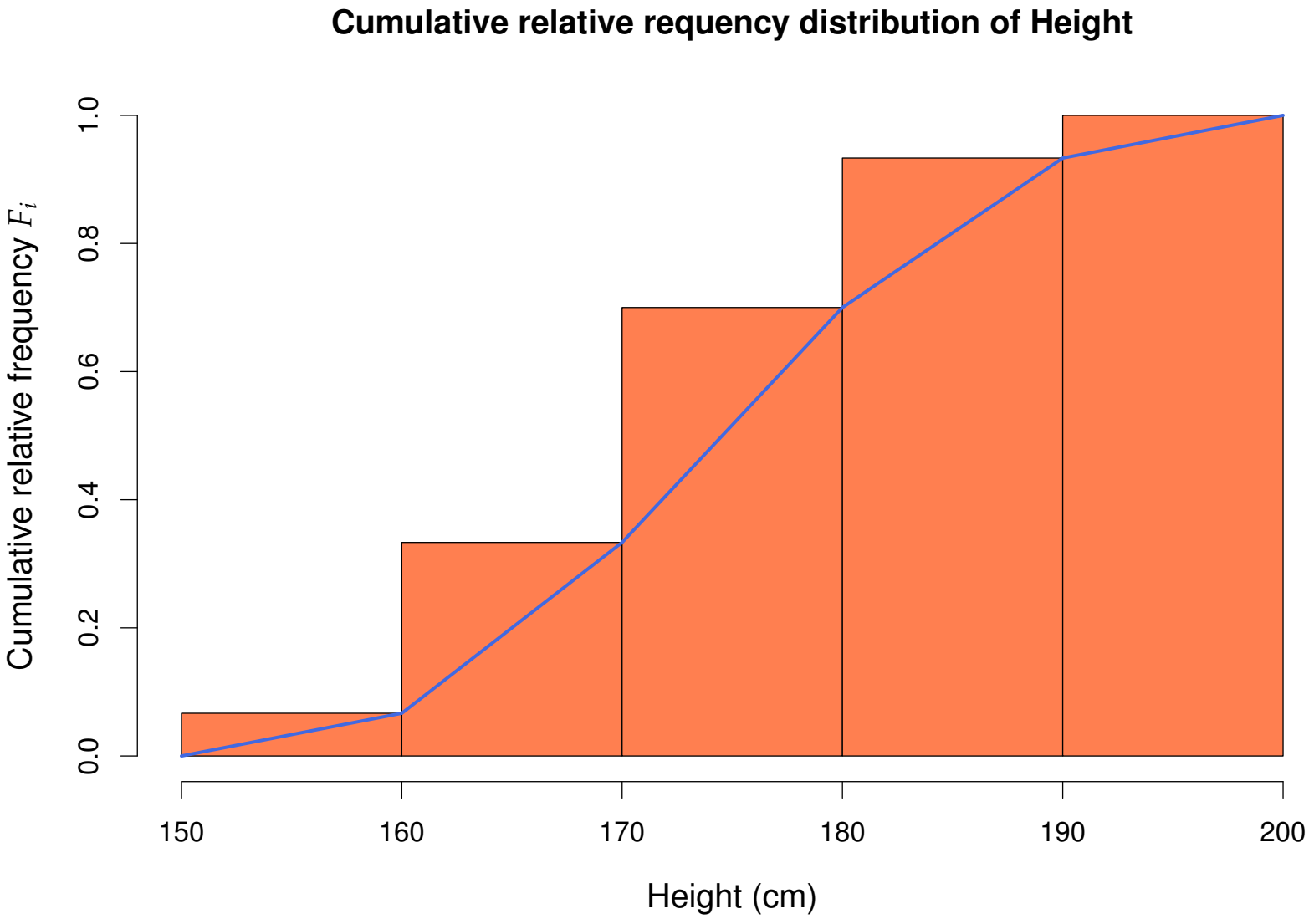
# Cumulative relative frequency histogram

## Grouped data



# Cumulative relative frequency line chart or ogive

## Grouped data



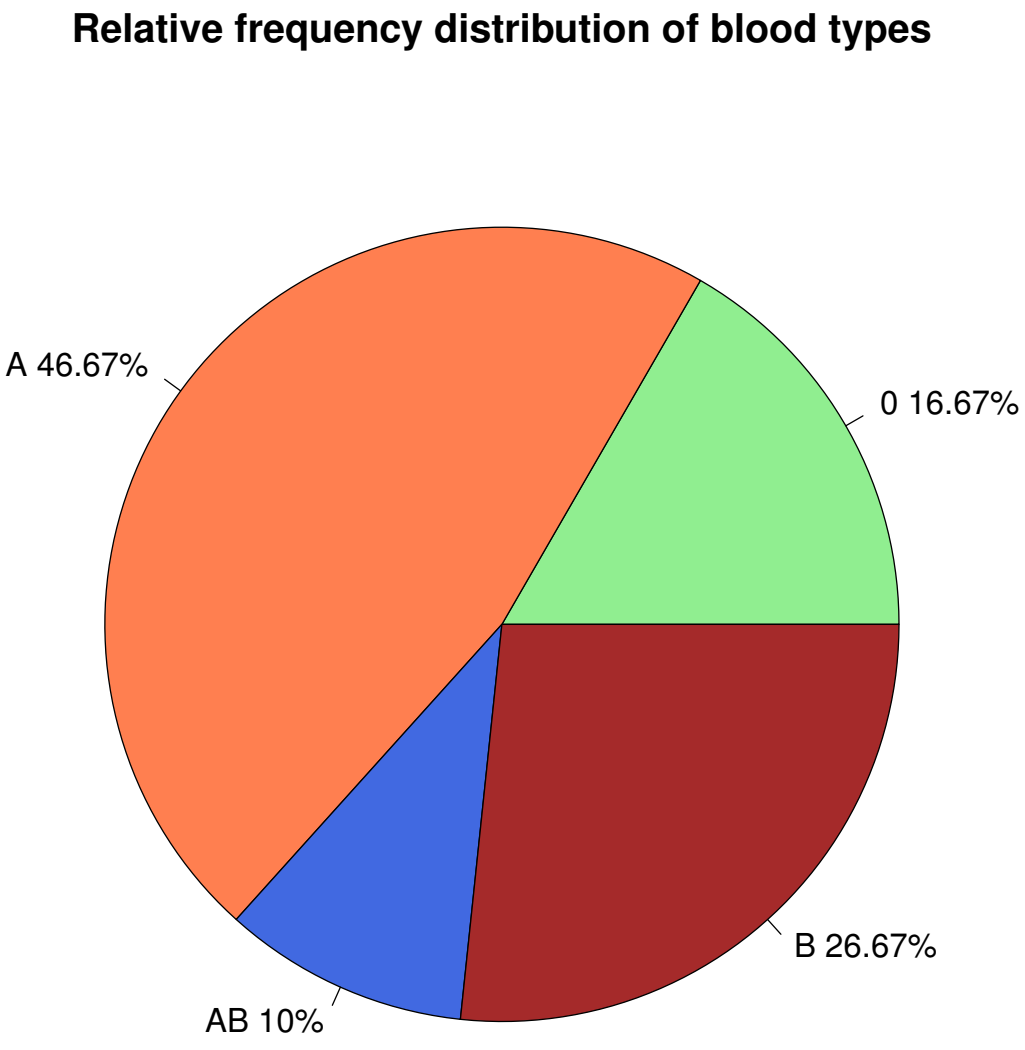
# Pie chart

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

Pie charts can represent absolute or relative frequencies, but not cumulative frequencies, and are used with nominal qualitative variables. For ordinal qualitative or quantitative variables is better to use bar charts or histograms, cause it's easy to perceive differences in one dimension (lenght of bars) than in two dimensions (areas of sectors).

# Pie chart

## Nominal variables



# Outliers

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

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

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

# Outliers management

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

With small samples we have several options:

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

# Sample statistics

The frequency table and charts summarize and give an overview of the distribution of values of the studied variable in the sample, but it's difficult to describe some aspects of the distribution from it.

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

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

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

Measures of dispersion: They measure the spread of data.

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

# Location statistics

There are two groups:

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

- Arithmetic mean
- Median
- Mode

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

- Quartiles.
- Deciles.
- Percentiles.