DATA SCIENCE &
ARTIFICIAL INTELLIGENCE

SCIENTIFIC &
DATA-INTENSIVE COMPUTING

UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

Descriptive Data Analysis

Introductory course on Statistics and Probability

Nicholas A. Pearson

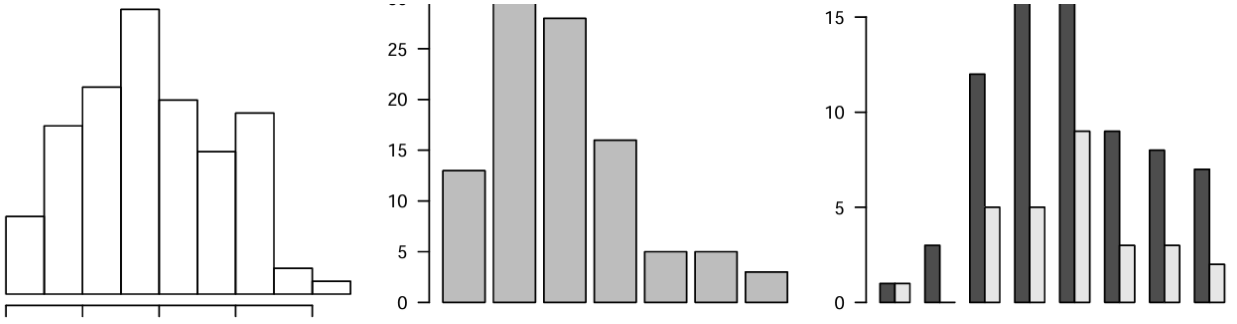
Università degli Studi di Trieste

September 9, 2025

Indexes

So far, we have seen...

- Data
  - Data organized as a matrix
  - List of observations: $y_1, ..., y_n$
- Frequency distributions
  - List of modalities and frequencies
  - List of class of modalities and frequencies
- Visualizations



But what are distributions and plots used for?

Sum up the data

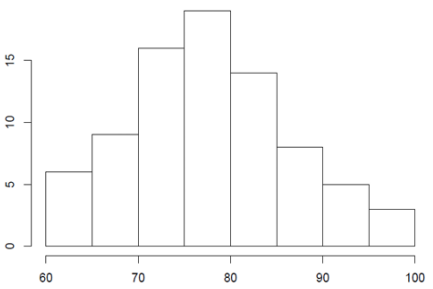
List

75 81 77 88 72 78 71 66
82 74 72 80 72 79 84 73
100 77 60 74 87 88 64 82
83 85 96 86 77 84 93 75
85 90 74 77 81 75 78 80
75 61 98 66 82 68 60 85
80 76 63 80 68 72 70 93
87 90 76 79 70 92 77 70
89 81 71 83 78 80 75 95
68 64 70 83 77 77 94 72

Classes distribution

y_i	n_i
[60,70]	15
(70,80]	35
(80,90]	22
(90,100]	8

Graphical representation



The aim is:

- Summarize data
- Shed light on some specific aspects

Distributions and plots help us gain a quick understanding of our data. However, it's important to remember that when you summarize data you also lose some detailed information.

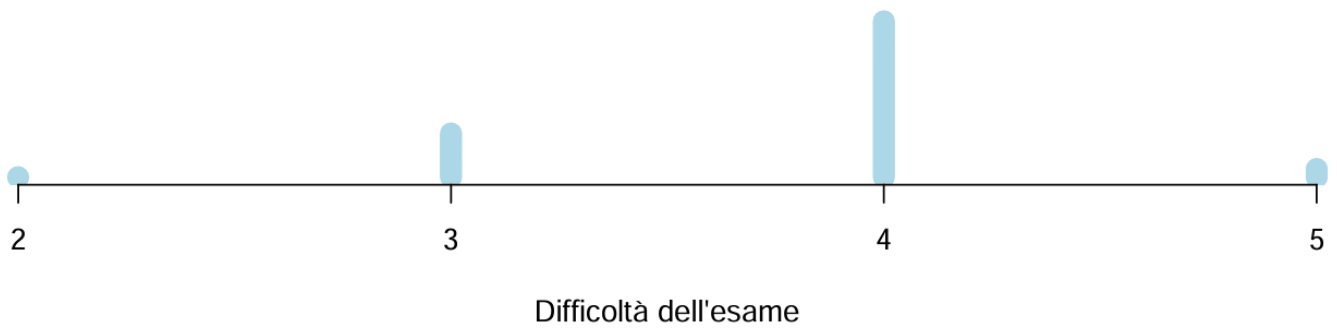
New tools

There are other tools available to summarize data. In particular, the aim is to summarize 3 different aspects of data distribution:

- Central tendency
- Variability
- Shape

Example: perceived difficulty of exam

	2	3	4	5
Freq	1	17	58	4

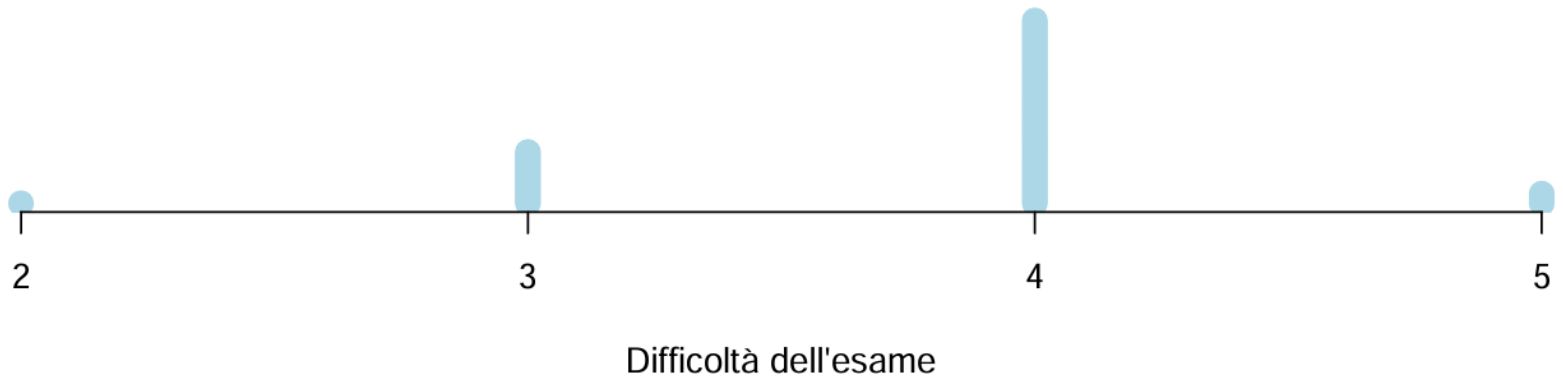


How would you describe this distribution? In particular, around what values is the distribution positioned? In other words, where is the center of the distribution?

"Position" of the distribution

The previous question asks us to **summarize the entire distribution into a single value** which, in some way, indicates where the distribution itself is "positioned".

It could be said that the distribution is positioned on the value that appears **most frequently**.



This value is called **mode** of the distribution.

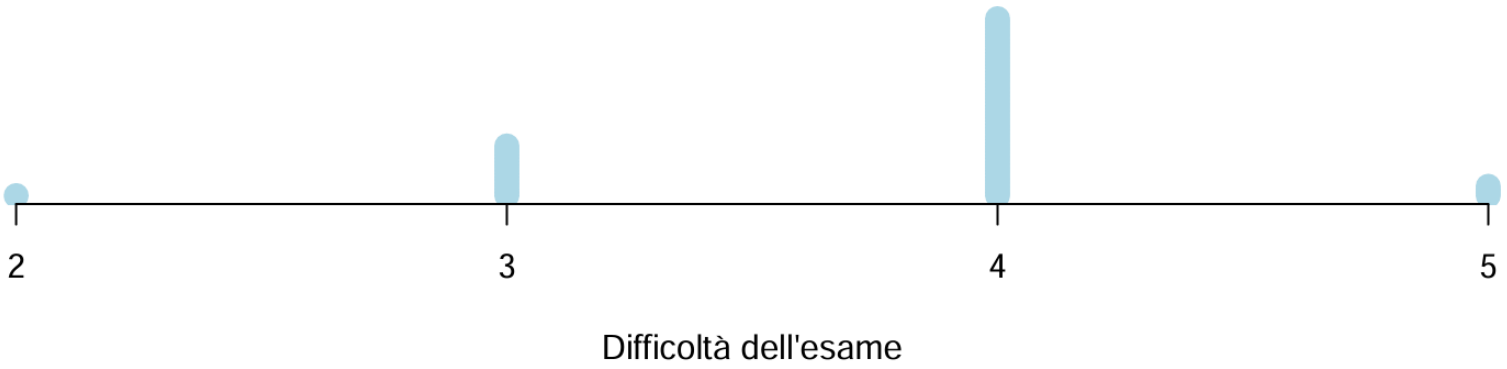
Central tendency measure: the mode

The **mode** of a distribution is the value that presents the highest relative frequency.

- The **mode** expresses the most frequent value in the distribution.
- It is defined for both qualitative and quantitative variables.

Mode as summarizing tool

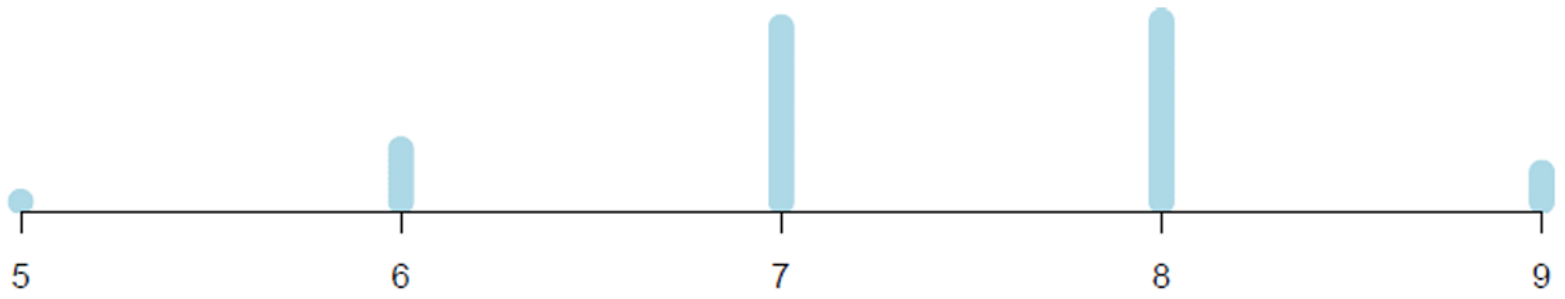
	2	3	4	5
Freq	1	17	58	4



In this example, the mode provides a clear summary of the overall distribution of the perceived exam difficulty.

Mode as summarizing tool

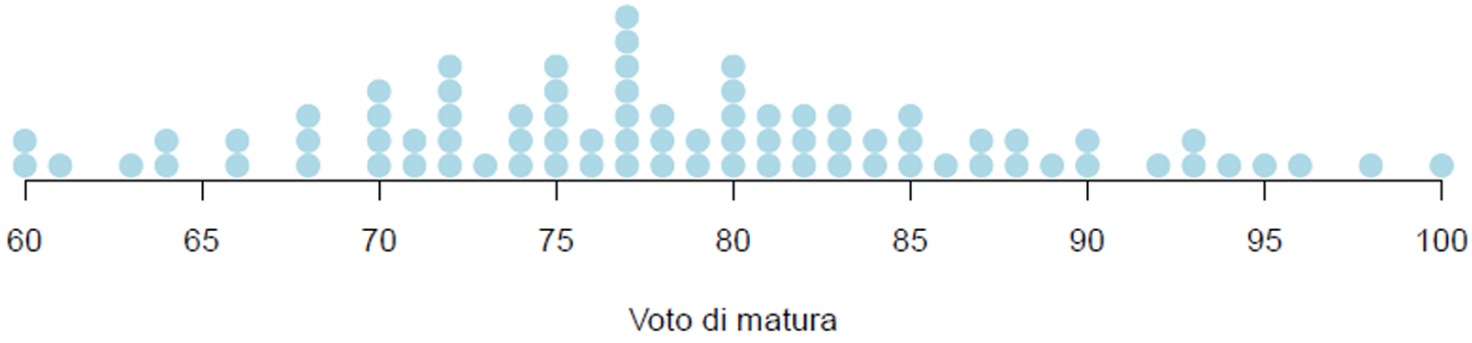
	5	6	7	8	9
Freq	1	10	31	32	6



However, in this example, the mode does not seem to work as well as before...

Mode as summarizing tool

	60	61	63	64	66	68	70	71	72	73	74	75	76	77	78	79	80	81
Freq	2	1	1	2	2	3	4	2	5	1	3	5	2	7	3	2	5	3



And it is even worse in this example focusing on final high school marks...

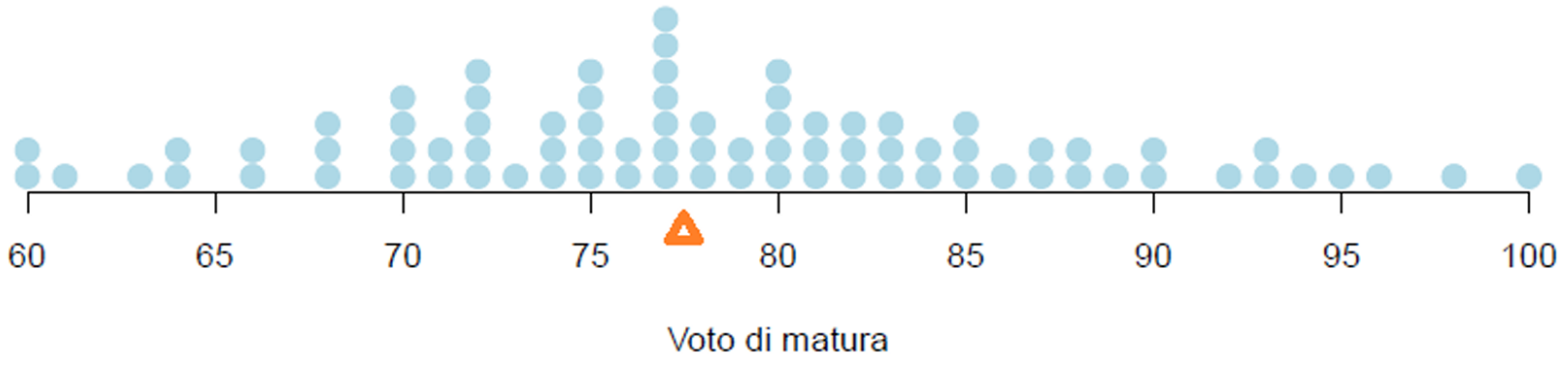
Central tendency measures

The center of a distribution could also be thought of as that value that leaves to both its right and to its left exactly 50% of the observations.

60	60	61	63	64	64	66	66	68	68
68	70	70	70	70	71	71	72	72	72
72	72	73	74	74	74	75	75	75	75
75	76	76	77	77	77	77	77	77	77
78	78	78	79	79	80	80	80	80	80
81	81	81	82	82	82	83	83	83	84
84	85	85	85	86	87	87	88	88	89
90	90	92	93	93	94	95	96	98	100

Central tendency measures

The center of a distribution could also be thought of as that value that leaves to its right and to its left exactly 50% of the observations.



Other central tendency measure: the median

Let $y_1, y_2, \cdots, y_N$ be a disaggregated statistical distribution. Let $y_{(1)}, y_{(2)}, \cdots, y_{(N)}$ the corresponding distribution of the ordered (sorted) values.

- $y_{(1)} = \min(y_1, \ldots, y_N)$, $y_{(N)} = \max(y_1, \ldots, y_N)$;
- $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)}$.

The **median**, indicated with m, is computed as:

$$m = \begin{cases} y_{(\frac{N+1}{2})} & \text{if N is odd} \\ \frac{y_{(\frac{N}{2})} + y_{(\frac{N}{2}+1)}}{2} & \text{if N is even} \end{cases}$$

The median is a particular **quantile**.

Quantiles

Quantiles are statistical measures that partition an ordered dataset into equal-sized subsets. They provide a summary of the distribution and spread of the data.

Types of Quantiles:

- **Percentile**: divide the data in 100 equal parts
- **Quartiles**: divide the data in 4 equal parts
  - Q_1: 25^{th} percentile, 25% of data below this value
  - Q_2: 50^{th} percentile, **Median**
  - Q_3: 75^{th} percentile

Central tendency measures: arithmetic mean

- The **arithmetic mean** $\bar{y}$ is calculated as:

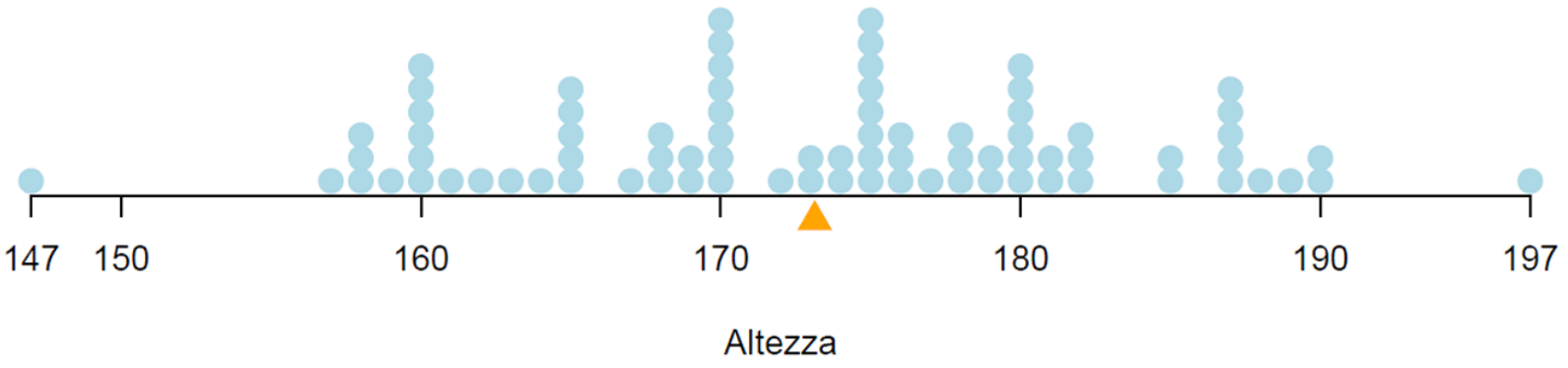
$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{1}{N} \sum_{i=1}^{N} y_i,$$

  where $(y_1, y_2, \cdots, y_N)$ represents the sample of N observed values of the variable Y.
- There different types of "means", but the Arithmetic one is undoubtedly the most commonly used. Other examples include:
  - Harmonic Mean
  - Geometric Mean

Example: height

Sample size is N = 80. Therefore:

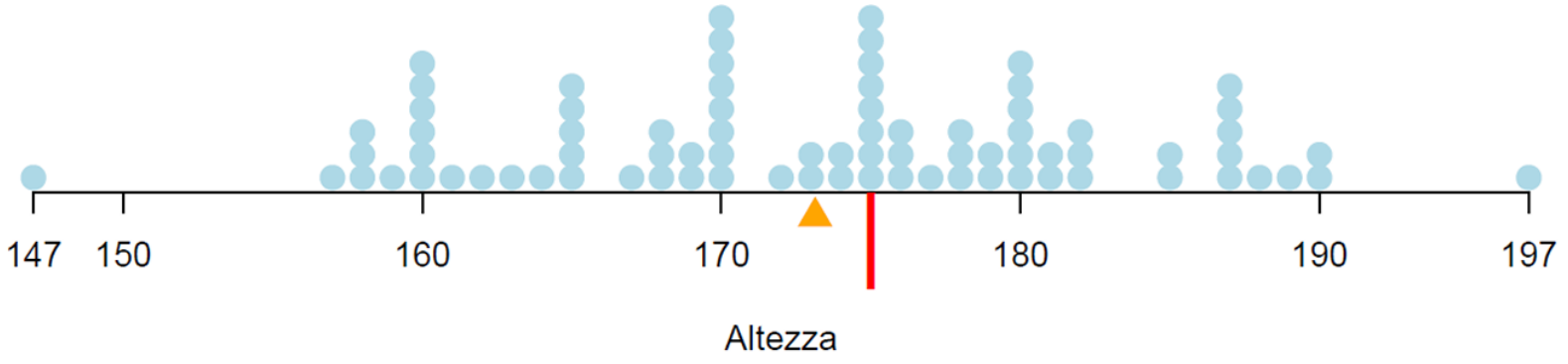
$$\frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} y_{(i)} = \frac{13851}{80} = 173.$$



Example: height

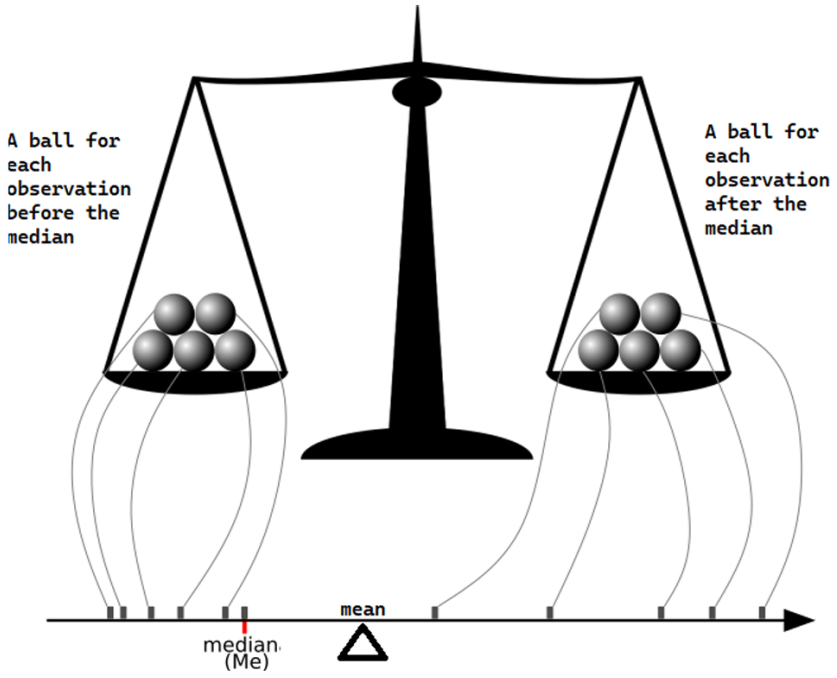
Sample size is N = 80. Therefore:

$$\frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} y_{(i)} = \frac{13851}{80} = 173.$$



The median (174.5) is very close to it.

Mean and Median



The median provides a balance point, with half of the observations below it and half above it, without regard to their distance from the center.

The mean, instead, is the centroid: like in physics, it balances the observations as if they were masses, taking their distances into account.

In a nutshell...

- The **mode**, the **median** and the **arithmetic mean** are the most used measures for the position (central tendency) of a distribution.
- If we deal with the entire population (we have a census), the measures are called **of the population** (it is traditional to indicate them with different symbols, often Greek letters).
- As we have said, it is rare to collect the data of the whole population.
- If we deal with a sample (most of the time, this is the real case), the measurements are called **sampling measures**. If the sample is representative, in general the sampling measures are good "indications" of the measures calculated on the entire population.

Marginal and conditional measures

The central tendency measures computed for conditional variables are, for simplicity, referred to as **conditional central tendency measures**, to distinguish them from those calculated on the unconditional (marginal) distribution.

Example:
Let Y being the height and let X the sex (let's assume, for simplicity, only the values M e F). We can calculate sex-conditional height measures and marginal measures

- Median of $Y|X = M \rightarrow 180$ (conditional median)
- Mean of $Y|X = M \rightarrow 180.2$ (conditional mean)
- Median of $Y|X = F \rightarrow 165$ (conditional median)
- Mean of $Y|X = F \rightarrow 165.7$ (conditional mean)
- Median of $Y \rightarrow 174.5$ (marginal median)
- Mean of $Y \rightarrow 173.1$ (marginal mean)

Some, just some, formulas

So far, we have already introduced some formulas for calculating central tendency measurements, in the case of having raw data available (i.e. the disaggregated statistical distribution).

Sometimes, even starting from the raw data, there may be ambiguities in the calculation of the measures (or indicators).

More generally, the data can be provided in aggregate form.

Now we will see what to do in these cases.

Median: for classes frequency distribution

Suppose we have the following frequency distribution:

	(0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]
absolute frequency	1	4	4	2	1

The data has size N = 12. The median should be chosen from 6th and the 7th observation from below. So m $\in$ (2, 3].

- Suppose (arbitrarily) that the four data belonging to the third interval are equally distributed. Under this assumption, the median is the mean of the values attributed to the 6° and to the 7° observation from below.
- Therefore:
  - y(6) = 2.25, y(7) = 2.50 $\rightarrow$ m = $\frac{2{,}25+2{,}50}{2} = 2.375$
  - y(6) = 2.20, y(7) = 2.40 $\rightarrow$ m = $\frac{2{,}20+2{,}40}{2} = 2.30$

Mean: for classes frequency distribution

Suppose we have a frequency distribution for classes of the following type:

intervals	$(c_0,\ c_1]$	$(c_1,\ c_2]$	$\cdots$	$(c_{k-1},\ c_k]$
absolute frequency	n_1	n_2	$\cdots$	n_k

where k indicates the number of classes. The **mean** can not be calculated directly in an exact way. A proxy often used in this case is:

$$\frac{\sum_{i=1}^{k} y_i n_i}{\sum_{i=1}^{k} n_i} = \frac{1}{N} \sum_{i=1}^{k} y_i n_i$$

where y_i is the central value of the class i, that is:

$$y_i = \frac{c_{i-1} + c_i}{2}$$

Example: high school mark

mark (class) $(c_{i-1}, c_i]$	frequency absolute n_i	central value of the class y_i	$y_i n_i$
[60,70]	15	65.0	975.0
(70,80]	35	75.5	2642.5
(80,90]	22	85.5	1881.0
(90,100]	8	95.5	764.0
		Total	6262.5

From which

$$\bar{y} = \frac{6262.5}{80} = 78.28$$

Mean computed from raw data is: $\bar{y} = 80$

Weighted mean calculation

The arithmetic mean calculated for grouped data is an example of **weighted arithmetic mean**

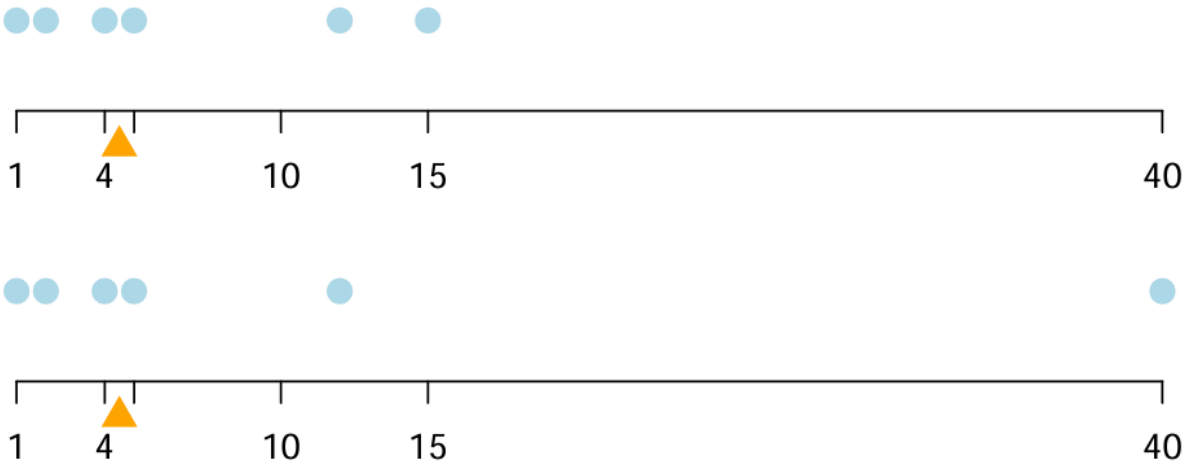
$$\bar{y}_w = \frac{\sum_{i=1}^{k} y_i w_i}{\sum_{i=1}^{k} w_i}$$

where to each modality y_i a non-negative weight w_i is assigned .

Median as a summarizing measure

Consider these two different samples which share the same value for the median...

- 1, 2, 4, 5, 12, 15
- 1, 2, 4, 5, 12, **40**

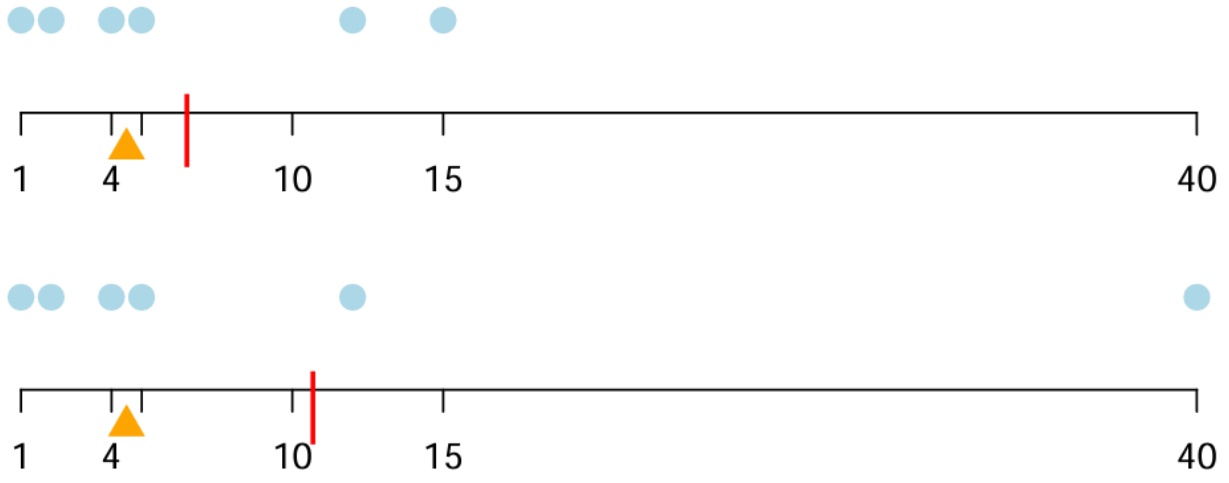


The median is not affected by the last extreme value (often called as outlier).

Median as a summarizing measure

Consider these two different samples which share the **same value for the median**...

- 1, 2, 4, 5, 12, 15
- 1, 2, 4, 5, 12, **40**



The arithmetic mean is very sensitive to extreme values:

- 1, 2, 4, 5, 12, 15 $\rightarrow \bar{y} = 6.5$
- 1, 2, 4, 5, 12, 40 $\rightarrow \bar{y} = 10.67$

Mean is not enough...

Given two different groups of five individuals, we analyze the observed heights in cm:

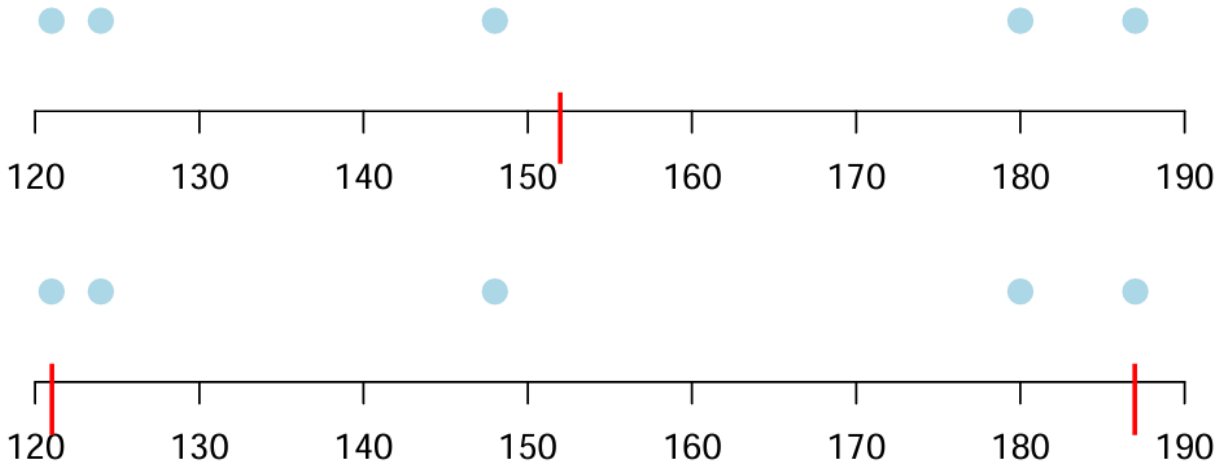
- Group 1: 150, 151, 156, 146, 157
- Group 2: 121, 150, 190, 180, 119

The mean is equal to 152cm for both groups, but the groups are pretty different!

Elementary Measures of variability: *range*

Two samples with same mean

- 146, 150, 151, 156, 157
- 121, 124, 148, 180, 187

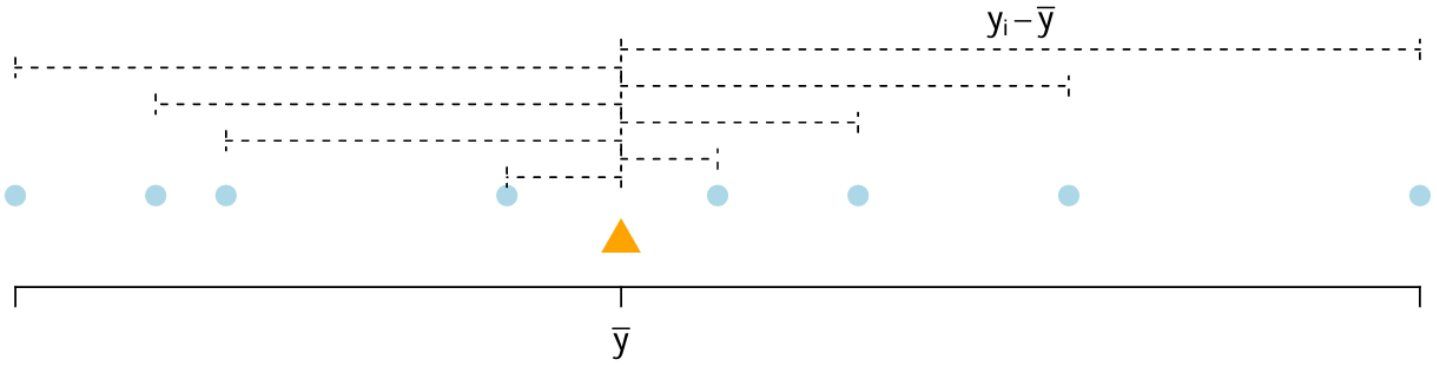


An intuitive measure of the variability of a set of data is the **range**, the difference/distance between maximum and minimum.

$$\text{Range} = y_{(N)} - y_{(1)}$$

Distance from the center

Another way to measure variability: distance from a center.



We consider as center the arithmetic mean $\bar{y}$.

- We compute the distance of each observation from the center (mean) as: $(y_i - \bar{y})^2$
- As last step, we make the mean of such quantities:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

Variance

The **Variance** of the observations $y_1, \ldots, y_N$ is the mean of the squares of the deviation of each observation from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N}$$

The variance of the variable Y is usually expressed in symbol as σ^2_Y or V(Y).

Variance: an example

Example:
variance for 5 observations, the mean is $\bar{y} = 2.8$

Observations y_i	deviations $y_i - \bar{y}$	(deviations)^2 $(y_i - \bar{y})^2$
-1	-3.80	14.44
1	-1.80	3.24
3	0.20	0.04
4	1.20	1.44
7	4.20	17.64
	Total	36.8

The variance is:

$$\sigma^2 = \frac{36.8}{5} = 7.36$$

Variance with frequency distribution

If the variable Y has modalities $y_1, \ldots, y_k$ with absolute frequencies $n_1, \ldots, n_k$ ($\sum_{i=1}^{k} n_i = N$) and relative frequencies $f_1, \ldots, f_k$ ($f_i = n_i/N$) the **variance** is calculated as:

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i (y_i - \bar{y})^2}{N} = \sum_{i=1}^{k} f_i (y_i - \bar{y})^2$$

Example: hours of sleep per night, N = 80, $\bar{y} = 7.4$

Modality y_i	Frequency n_i	deviation $y_i - \bar{y}$	(deviation)^2 $(y_i - \bar{y})^2$	weighted deviations $n_i(y_i - \bar{y})^2$
5	1	-2.40	5.7600	5.7600
6	10	-1.40	1.9600	19.6000
7	31	-0.40	0.1600	4.9600
8	32	0.60	0.3600	11.5200
9	6	1.60	2.5600	15.3600
			Total	57.2

The variance is: $\sigma^2 = \frac{57.2}{80} = 0.72$

Standard deviation

The **Standard deviation** is the square root of the variance:

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

Its main advantage is that it is expressed in the same unit of measures as the original variable.
Standard deviation for hours of sleep is as follows:

$$\sigma = \sqrt{0.72} = 0.85$$

Quantile Range

The **Interquartile Range** measures the spread of the middle 50% of the data.

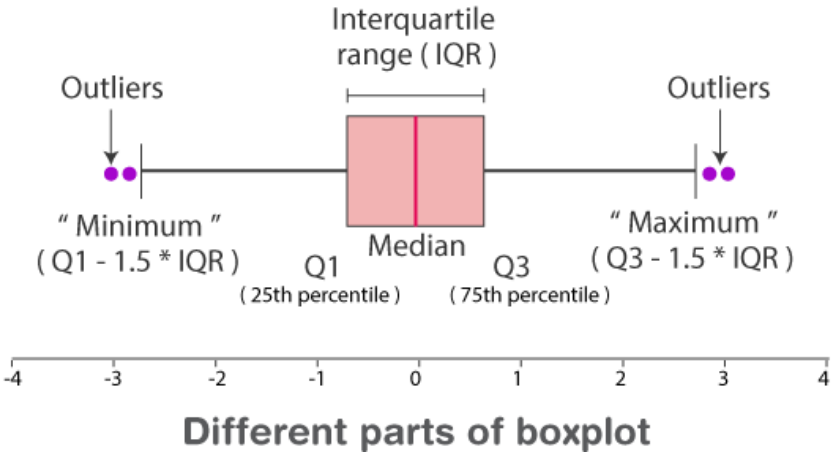
$$IQR = Q_3 - Q_1$$

Properties of IQR:

- Robust to outliers (unlike range or standard deviation)
- Contains the middle 50% of observations
- Useful for identifying outliers: values beyond $Q_1 - 1.5 \times IQR$ or $Q_3 + 1.5 \times IQR$

A **Box Plot** provides a visual summary using quantiles and identifies outliers.

Box Plot: Visual summary



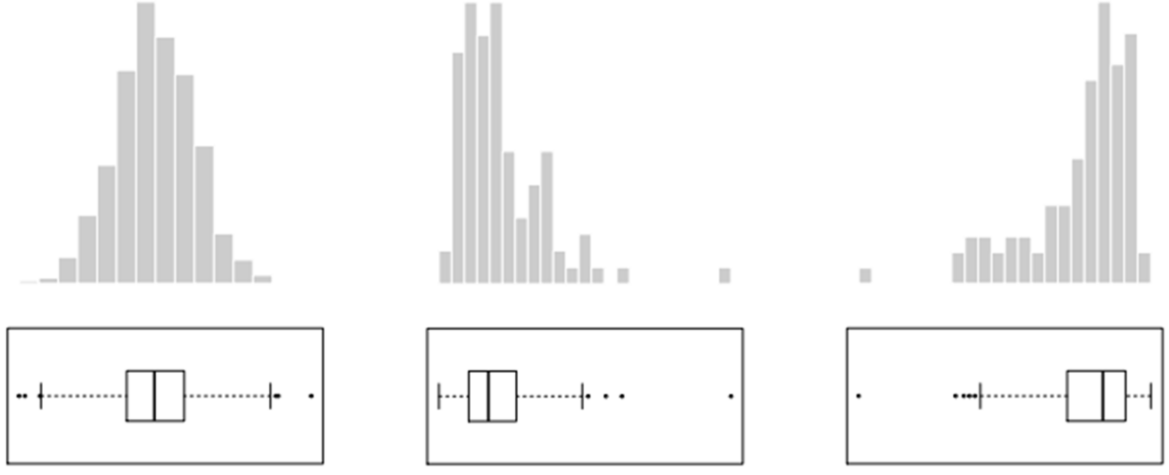
Different parts of boxplot

Box Plot Components:

- **Box**: From Q_1 to Q_3 (contains middle 50% of data)
- **Median line**: Vertical line at Q_2 inside the box
- **Whiskers**: Extend to furthest points within $1.5 \times IQR$ from box edges
- **Outliers**: Points beyond the whiskers (plotted individually)

Skewness

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.



Most software programs compute the *Adjusted Fisher-Pearson coefficient*:

$$G_1 = \frac{\sqrt{N(N-1)}}{N-2} \frac{\sum_{i=1}^{N}(y_i - \bar{y})^3/N}{s^3}$$

Skewness

- The skewness of symmetric data should be near zero.
- Negative values for the skewness indicate that the data are skewed left (left tail is long relative to the right tail).
- Positive values for the skewness indicate that the data are skewed right (right tail is long relative to the left tail).