Statistics: descriptive statistics, summarizing the data

S. Maset

Dipartimento di Matematica e Geoscienze, Università di Trieste

PEM 2017-2018

Outline

1. Introduction
2. Mean
3. Median
4. Percentiles
5. Mean and percentiles of a symmetric data
6. Mode
7. Variance
8. Standard deviation
9. Interquartile range
10. Box plot
11. Mean and percentiles of an histogram
12. Normal data
    - Other types of data
    - Some considerations about normal data
13. Correlation coefficient
    - The regression line
    - Causation and association

Introduction

- We are interested in determining certain summary measures about the data.

  These summary measures are called **statistics**: a statistic is a rule (a function, in mathematical terms) that associates a number to the data $\boldsymbol{x}$.

  These summary measures are important for data of large size.

- A famous example of data of large size:
  - The medical statisticians R. Doll and A. B. Hill sent questionnaires in 1951 to all doctors in the United Kingdom and received 40000 replies.
  - Their questionnaire dealt with age, eating habits, exercise habits and smoking habits.
  - These doctors were then monitored for 10 years and the causes of death of those who died were determined.

  Even if we just focus only on one component of the study, such as the age of the doctors, the resulting data is huge: $n = 40000$.

  The Doll–Hill study yielded the results:

  - only about 1 in 1000 nonsmoking doctors died of lung cancer; for heavy smokers the ratio was 1 in 8;
  - the death rates from heart attacks were 50 percent higher for smokers.

- We are interested in statistics that describe the **central tendency** of the data, i.e. statistics that say where is the center of the data.

  We present three of these:

  - the **mean**;
  - the **median**;
  - the **mode**.

  Once we have identified where is the center of the data, the following question can be raised: how much variation is there in the data with respect to the center?

  Are most of the components of the data close to the center, or do they vary widely with respect to the center?

  We discuss the **standard deviation** and the **interquartile range**, which are statistics for measuring such a variation.

Mean

- The **mean** of the data $\boldsymbol{x} \in \mathbb{R}^n$ is the arithmetic average of the components of $\boldsymbol{x}$:

$$\overline{x} := \frac{\sum\limits_{i=1}^{n} x_i}{n}.$$

Example. The data

$\boldsymbol{x}$*=(3,4,1,0,0,1,-2,-2,-2,-4,-3,1,-2,1,0,0,1,1,-2,-1,0,-1,-1,3,1,0,5,2)*

gives the minimum temperatures (unit: Celsius degree) in February 2013 in Pordenone (from February 1 to February 28). The mean is

$$\overline{x} = \frac{4}{28} = 0.14.$$

Observe that

$$\min_{i\in\{1,\ldots,n\}} x_i \leq \overline{x} \leq \max_{i\in\{1,\ldots,n\}} x_i.$$

Exercise. Prove this property.

- Here are two important properties of the mean. Let $\boldsymbol{x} \in \mathbb{R}^n$ be a data.

  Let $c \in \mathbb{R}$. Consider the $n-$tuple $\boldsymbol{y} = c + \boldsymbol{x}$ whose components are

  $$y_i = c + x_i,\ i \in \{1, \ldots, n\}.$$

  We have

  $$\overline{y} = c + \overline{x}.$$

  Let $c \in \mathbb{R}$. Consider the $n-$tuple $\boldsymbol{y} = c\boldsymbol{x}$ whose components are

  $$y_i = cx_i,\ i \in \{1, \ldots, n\}.$$

  We have

  $$\overline{y} = c\overline{x}.$$

  Exercise. Prove these two properties.

The first property can be used to simplify the computations by hand.

Example. The scores in the last 13 years of the seria A winners (tournaments with 20 teams) are

$$\boldsymbol{y} = (91, 91, 87, 102, 87, 84, 82, 82, 84, 85, 97, 91, 86)$$

(from season 2016-2017 down to to season 2004-2005).

Since

$$\boldsymbol{y} = 87 + (4, 4, 0, 15, 0, -3, -5, -5, -3, 10, -1)$$

we have

$$\overline{y} = 87 + \frac{16}{13} = 88 + \frac{3}{13}.$$

- In MATLAB, the mean of the data in the vector x is computed by

$$\mathrm{sum(x)/length(x)} \text{ or } \mathrm{mean(x)}.$$

Exercise. By using MATLAB, compute the mean of the temperatures and the mean of the scores of the two previous examples.

- Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$ and let $\boldsymbol{y} \in \mathbb{R}^n$ obtained by replacing in $\boldsymbol{x}$ one component equal to a with b. Determine the difference of the means $\overline{y} - \overline{x}$ in terms of the difference $b - a$ and n.

  Exercise. Let $\boldsymbol{y} = (\boldsymbol{x}, a) \in \mathbb{R}^{n+1}$. Determine the difference of the means $\overline{y} - \overline{x}$ in terms of the difference $a - \overline{x}$ and n.

  Exercise. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Consider the $n-$tuple $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$ whose components are

  $$z_i = x_i + y_i,\ i \in \{1, \ldots, n\}.$$

  Prove that

  $$\overline{z} = \overline{x} + \overline{y}.$$

  This property along with $\overline{y} = c\overline{x}$ for $\boldsymbol{y} = c\boldsymbol{x}$ says that the function

  $$\mathbb{R}^n \to \mathbb{R},\ \boldsymbol{x} \mapsto \overline{x},$$

  is linear.

- Now, we consider the computation of the mean when the data $\boldsymbol{x}$ is arranged in a frequency table with data values v_j, $j \in \{1, \ldots, l\}$. We have

$$\overline{x} = \frac{\sum\limits_{j=1}^{l} f_j v_j}{\sum\limits_{j=1}^{l} f_j} = \sum_{j=1}^{l} \widehat{f}_j v_j$$

where f_j and $\widehat{f}_j$ are the frequency and the relative frequency, respectively, of v_j in $\boldsymbol{x}$, $j \in \{1, \ldots, l\}$.

In fact, we have

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{l} f_j v_j \text{ and } n = \sum_{j=1}^{l} f_j$$

and then

$$\overline{x} = \frac{\sum\limits_{i=1}^{n} x_i}{n} = \frac{\sum\limits_{j=1}^{l} f_j v_j}{\sum\limits_{j=1}^{l} f_j} \quad \text{and} \quad \overline{x} = \frac{\sum\limits_{i=1}^{n} x_i}{n} = \frac{\sum\limits_{j=1}^{l} f_j v_j}{n} = \sum_{j=1}^{l} \frac{f_j}{n} v_j = \sum_{j=1}^{l} \widehat{f}_j v_j.$$

Example: consider the frequency table

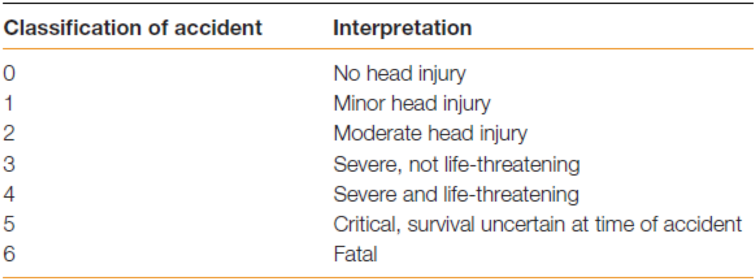
v_j	f_j
3	2
4	1
5	3

We have

$$\overline{x} = \frac{\sum_{j=1}^{l} f_j v_j}{\sum_{j=1}^{l} f_j} = \frac{2 \cdot 3 + 1 \cdot 4 + 3 \cdot 5}{2 + 1 + 3} = \frac{25}{6} = 4 + \frac{1}{6}.$$

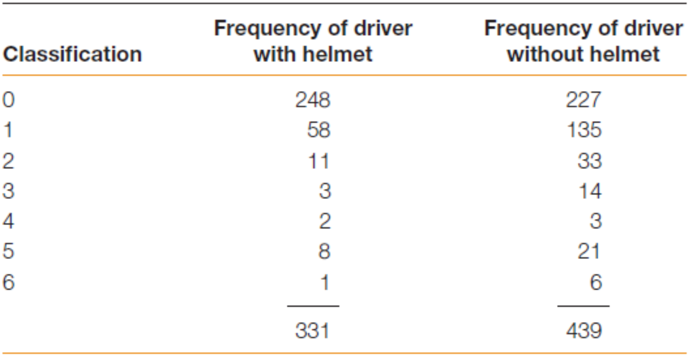
Example: we analyze informations about 770 *similar motorcycle accidents in the Los Angeles area between 1976 and 1977.*

Each accident was classified according to the severity of the head injury suffered by the motorcycle driver:



Classification of accident	Interpretation
0	No head injury
1	Minor head injury
2	Moderate head injury
3	Severe, not life-threatening
4	Severe and life-threatening
5	Critical, survival uncertain at time of accident
6	Fatal

Frequency tables giving the severities of the accidents that occurred when the driver was wearing and was not wearing a helmet



Classification	Frequency of driver with helmet	Frequency of driver without helmet
0	248	227
1	58	135
2	11	33
3	3	14
4	2	3
5	8	21
6	1	6
	331	439

Exercise. Here we have two data ***x*** *and* ***y****: the first for drivers with helmet and the second for drivers without helmet. Describe* ***x*** *and* ***y****.*

The mean of the severities for drivers that wore a helmet is

$$\begin{aligned}\overline{x} &= \frac{\sum_{j=1}^{I} f_j v_j}{\sum_{j=1}^{I} f_j} = \frac{248 \cdot 0 + 58 \cdot 1 + 11 \cdot 2 + 3 \cdot 3 + 2 \cdot 4 + 8 \cdot 5 + 1 \cdot 6}{331} \\ &= 0.432.\end{aligned}$$

The mean of the severities for drivers that did not wear a helmet is

$$\begin{aligned}\overline{y} &= \frac{\sum_{j=1}^{I} f_j v_j}{\sum_{j=1}^{I} f_j} = \frac{227 \cdot 0 + 135 \cdot 1 + 33 \cdot 2 + 14 \cdot 3 + 3 \cdot 4 + 21 \cdot 5 + 6 \cdot 6}{439} \\ &= 0.902.\end{aligned}$$

Conclusion: those drivers who were wearing a helmet suffered, by looking at the means, less severe head injuries than those who were not wearing a helmet.

- In MATLAB, given the frequencies in the vector f and the data values in the vector v, both column vectors, the mean is given by

$$\mathrm{f}' * \mathrm{v}/\mathrm{sum}(\mathrm{f}).$$

Exercise. By using MATLAB, find the two means for drivers wearing or not wearing the helmet in the previous example.

- Given numbers $w_j \in [0,1]$, $j \in \{1,\dots,l\}$, such that

$$\sum_{j=1}^{l} w_j = 1,$$

the quantity

$$\sum_{j=1}^{l} w_j v_j$$

is called the **weighted average** of the values v_j, $j \in \{1,\dots,l\}$, with **weights** w_j.

The mean is the weighted average of the data values with weights the relative frequencies: we have

$$\overline{x} = \sum_{j=1}^{l} \widehat{f}_j v_j \ \text{ and } \ \sum_{j=1}^{l} \widehat{f}_j = 1.$$

- Let $\boldsymbol{x} \in \mathbb{R}^n$ be a data. The quantities

$$x_i - \overline{x},\ i \in \{1, \ldots, n\},$$

are called **deviations** from the mean.

We have

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0.$$

Exercise. Prove this property of the deviations by proving that, looking for $a \in \mathbb{R}$ such that

$$\sum_{i=1}^{n} (x_i - a) = 0$$

we have the unique solution $a = \overline{x}$.

- Now we can give a physical interpretation of the mean $\overline{x}$ as the center of the data $\boldsymbol{x}$

  If n objects of the same mass m are placed in a rod at the abscissas x_i, $i \in \{1, \ldots, n\}$, then $\overline{x}$ is the position where a fulcrum has to be placed under the rod for having the rod in equilibrium.

  In fact, the fulcrum has to placed at a point P such that the sum of torques with respect to P is zero, i.e.

  $$\sum_{i=1}^{n} mg\,(x_i - a) = mg \sum_{i=1}^{n} (x_i - a) = 0$$

  where a is the abscissa of P. This is equivalent to

  $$\sum_{i=1}^{n} (x_i - a) = 0$$

  which gives $a = \overline{x}$.

We can also say that $\overline{x}$ is the position of the center of mass:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n} m x_i}{\sum_{i=1}^{n} m}.$$

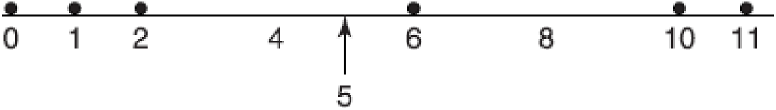
Example: consider

$$\boldsymbol{x} = (0, 1, 2, 6, 10, 11)\,.$$

The mean

$$\overline{x} = \frac{0+1+2+6+10+11}{6} = \frac{30}{6} = 5$$

in this interpretation is shown below



Median

- The following data

$$\boldsymbol{x} = (24, 23, 24, 24, 23, 23, 26, 89, 24, 23, 24, 23)$$

represents the ages of the twelve students attending a degree master.

The mean is $\overline{x} = 29.17$. By summarizing the data with the mean, one would think that the students have not a typical student age.

But, eleven of the twelve students have a typical student age and the other student is very old.

This points out a weakness of the mean as an indicator of the center of the data: the mean is greatly affected by extreme components of the data.

- A statistic that is also used to indicate the center of a data, but that is not affected by extreme components, is the **median.**

  Given a data $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, let

  $$\boldsymbol{x}^{\text{ord}} := \left(x_1^{\text{ord}}, x_2^{\text{ord}}, \ldots, x_n^{\text{ord}}\right).$$

  be the $n-$tuple of the components $x_1, x_2, \ldots, x_n$ ordered from the smallest to the largest.

  As an example, for

  $$\boldsymbol{x} = (0, -1, 0, 1, 2, 1, 0, -1)$$

  we have

  $$\boldsymbol{x}^{\text{ord}} = (-1, -1, 0, 0, 0, 1, 1, 2).$$

Informally, the median is defined as the value v that divide the $n-$tuple $\boldsymbol{x}^{\mathrm{ord}}$ in two parts with the same number of components: one part has all components $\leq v$ and the other part has all component $\geq v$.

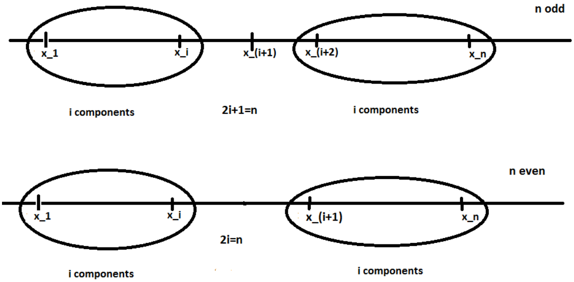
Formally, the median m_x of $\boldsymbol{x}$ is defined as

$$m_x := \begin{cases} x^{\mathrm{ord}}_{\frac{n-1}{2}+1} = x^{\mathrm{ord}}_{\frac{n+1}{2}} \text{ if } n \text{ is odd} \\ \\ \frac{1}{2}\left(x^{\mathrm{ord}}_{\frac{n}{2}} + x^{\mathrm{ord}}_{\frac{n}{2}+1}\right) \text{ if } n \text{ is even.} \end{cases}$$

By this formal definition: if $n = 3$, the median is the second component of $\boldsymbol{x}^{\mathrm{ord}}$; if $n = 4$, it is the average of the second and the third components of $\boldsymbol{x}^{\mathrm{ord}}$.

Exercise. For the data $\boldsymbol{x} = (0, -1, 0, 1, 2, 1, 0, -1)$ previously seen, compute the median.

Now, we show that the median m_x given by the formal definition satisfies the informal definition.



If n is odd, the median is the middle component x_{i+1}^{ord}, with $i = \frac{n-1}{2}$, of $\boldsymbol{x}^{\mathrm{ord}}$: the first i components of $\boldsymbol{x}^{\mathrm{ord}}$ are all $\leq x_{i+1}^{\mathrm{ord}}$ and the last i components of $\boldsymbol{x}^{\mathrm{ord}}$ are all $\geq x_{i+1}^{\mathrm{ord}}$.

If n is even, the median is the average a of the two middle components x_i^{ord} and x_{i+1}^{ord}, with $i = \frac{n}{2}$: the first i components of $\boldsymbol{x}^{\mathrm{ord}}$ are all $\leq a$ and the last i components of $\boldsymbol{x}^{\mathrm{ord}}$ are all $\geq a$.

For the previous data giving the ages of the students, where $n = 12$, the ordered $n-$tuple is

$$\boldsymbol{x}^{\mathrm{ord}} = (23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 26, 89)$$

and then the median is

$$m_x = \frac{x_6^{\mathrm{ord}} + x_7^{\mathrm{ord}}}{2} = 24.$$

With respect to the mean $\overline{x} = 29.17$, this is a better measure of the central tendency of the data.

- The mean, being the arithmetic average, is computed by using all the components of the data and then it is affected by extreme components.

  On the other hand, the median is computed by using the middle components and then it is not affected by extreme components.

  Exercise. Consider the data $\boldsymbol{x} = (0, -1, 0, 1, 2, 1, 0, -1)$ and $\boldsymbol{y} = (0, -1, 0, 1, 2000, 1, 0, -1)$. Are the means different? Are the medians different?

In general, consider $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$ obtained by replacing in $\boldsymbol{x}$ the last component of $\boldsymbol{x}^{\mathrm{ord}}$ equal to a with b larger than a. We have (see a previous exercise on the mean)

$$\overline{y} - \overline{x} \quad = \quad \frac{b-a}{n}$$

The change in the mean depends on b and it can be arbitrarily large.

On the other hand, there is no change in the median since $\boldsymbol{x}^{\mathrm{ord}}$ and $\boldsymbol{y}^{\mathrm{ord}}$ differ only in the last component.

Now, consider this other situation where $\boldsymbol{x} \in \mathbb{R}^n$ and

$$\boldsymbol{y} = (\boldsymbol{x}, a) \in \mathbb{R}^{n+1},$$

where a is larger than all components of $\boldsymbol{x}$. We have (see a previous exercise on the mean)

$$\overline{y} - \overline{x} \quad = \quad \frac{a - \overline{x}}{n+1}$$

The change in the mean depends on a and it can be arbitrarily large.

On the other hand, since $\boldsymbol{y}^{\mathrm{ord}} = \left(\boldsymbol{x}^{\mathrm{ord}}, a\right)$, as for the medians we have

$$
\begin{aligned}
m_y - m_x &= \begin{cases} y^{\mathrm{ord}}_{\frac{n+1+1}{2}} - \frac{1}{2}\left(x^{\mathrm{ord}}_{\frac{n}{2}+1} + x^{\mathrm{ord}}_{\frac{n}{2}}\right) \text{ if } n+1 \text{ is odd} \\ \\ \frac{1}{2}\left(y^{\mathrm{ord}}_{\frac{n+1}{2}} + y^{\mathrm{ord}}_{\frac{n+1}{2}+1}\right) - x^{\mathrm{ord}}_{\frac{n+1}{2}} \text{ if } n+1 \text{ is even} \end{cases} \\
&= \begin{cases} x^{\mathrm{ord}}_{\frac{n}{2}+1} - \frac{1}{2}\left(x^{\mathrm{ord}}_{\frac{n}{2}+1} + x^{\mathrm{ord}}_{\frac{n}{2}}\right) \text{ if } n+1 \text{ is odd} \\ \\ \frac{1}{2}\left(x^{\mathrm{ord}}_{\frac{n+1}{2}} + x^{\mathrm{ord}}_{\frac{n+1}{2}+1}\right) - x^{\mathrm{ord}}_{\frac{n+1}{2}} \text{ if } n+1 \text{ is even} \end{cases} \\
&= \begin{cases} \frac{1}{2}\left(x^{\mathrm{ord}}_{\frac{n}{2}+1} - x^{\mathrm{ord}}_{\frac{n}{2}}\right) \text{ if } n+1 \text{ is odd} \\ \\ \frac{1}{2}\left(x^{\mathrm{ord}}_{\frac{n+1}{2}+1} - x^{\mathrm{ord}}_{\frac{n+1}{2}}\right) \text{ if } n+1 \text{ is even} \end{cases}
\end{aligned}
$$

The change in the median is independent of a.

- *Example. Reconsider the scores of the winners of the seria A in the last 13 years*

$$\boldsymbol{y} = (91, 91, 87, 102, 87, 84, 82, 82, 84, 85, 97, 91, 86)\,,$$

where the mean is

$$\overline{y} = 88 + \frac{3}{13}.$$

On the other hand, the ordered n−tuple is

$$\boldsymbol{y}^{\mathrm{ord}} = (82, 82, 84, 84, 85, 86, 87, 87, 91, 91, 91, 97, 102)$$

and the median is

$$m_y = y_7^{\mathrm{ord}} = 87.$$

Exercise. Compute the median in the previous example of the minimum temperatures during February 2013 in Pordenone.

- Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$, let $c \in \mathbb{R}$ and let $\boldsymbol{y} = c + \boldsymbol{x}$. Prove that $m_y = c + m_x$.

  Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$, let $c > 0$ and let $\boldsymbol{y} = c\boldsymbol{x}$. Prove that $m_y = cm_x$.

  Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$ and let $\boldsymbol{y} = -\boldsymbol{x}$. Prove that $m_y = -m_x$.

  Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$, let $c \in \mathbb{R}$ and let $\boldsymbol{y} = c\boldsymbol{x}$. Prove that $m_y = cm_x$.

  Exercise. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $\boldsymbol{x} = \boldsymbol{x}^{\mathrm{ord}}$ and $\boldsymbol{y} = \boldsymbol{y}^{\mathrm{ord}}$, and let $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$. Prove that $m_z = m_x + m_y$. Moreover, find data $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3$ such that $m_z \neq m_x + m_y$.

  Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$. Prove that the sum $\sum_{i=1}^{n}(x_i - m_x)$ of the deviations from the median is $n(\overline{x} - m_x)$.

- The question regarding which of mean and median is the more informative summarizing statistics of the center of the data depends on what one is interested in learning from the data.

  If one is interested in:

  - something related to the sum of the components of the data, it is better the mean;
  - something related to the value dividing the ordered data in two parts with the same number of components, it is better the median.

Example. Consider a city with many poor people and very few very rich people. In this city, due to the very few very rich people, the mean of the incomes is much larger than the median.

If the city government

- *introduces a new flat-rate income tax (flat-rate means that the amount is a fixed precentage of the income) and it is trying to figure out how much money it can get, the mean of the incomes of the inhabitants is more informative than the median;*
- *plans to build some middle-income housing and it is trying to figure out the selling price for these houses, the median of the incomes is more informative than the mean; if the selling price was based on the mean, the very few rich people can move the mean at a level such that few people can buy the houses.*

Exercise. In the common speech, when we say that someone is "above the mean", do we intend to say that she/he is "above the median"?

- Exercise. Often, sports rankings are presented in TV by two columns of the same size: the ranking is from the top to the bottom of the the first column and then it continues from the top to the bottom in the second column. On 19th October 2017, the italian sports newspaper "Gazzetta dello Sport" published a ranking of the 32 teams participating to the Champions League 2017-2018. The ranking was based on the results of the teams after three turns. The team Athletic Madrid was placed in 16th position and Gazzetta wrote about this: "Athletic Madrid is almost in the right column". Try to rewrite this sentence in a more appropriate statistical wording.

- In MATLAB, the median of the data in the vector x is computed by

$$\mathrm{median}(\mathrm{x}).$$

Exercise. By using MATLAB, compute the medians in the examples of the students attending a degree master, the scores for the serie A winners and the minimum temperatures during February 2013 in Pordenone.

- IN MATLAB, given a data with frequencies in the vector f and data values in the vector v,

  $$\mathrm{x} = \mathrm{construct}(\mathrm{f}, \mathrm{v})$$

  constructs a vector x with those frequencies and data values.

  In this situation where the frequency table is given, the median is obtained by

  $$\mathrm{x} = \mathrm{construct}(\mathrm{f}, \mathrm{v});\ \mathrm{median}(\mathrm{x}).$$

  Exercise. Compute the medians for drivers wearing or not wearing the helmet in the previous example of the motorcycle accidents.

Percentiles

- The median is a particular case of a more general statistic known as **percentile** or **quantile**. For any $p \in (0,1)$, we define what is called the $100p$th percentile, or the quantile p.

  Informally, given a data $\boldsymbol{x} \in \mathbb{R}^n$, the $100p$th percentile is defined as the value v that divides the n-tuple $\boldsymbol{x}^{\mathrm{ord}}$ in two parts with number of components proportional to p and $1-p$: the part with number of components proportional to p has components $\leq v$ and the other part has components $\geq v$.

  Formally, the $100p$th percentile of $\boldsymbol{x}$ is defined as

  $$100p\text{th percentile of } \boldsymbol{x} = \begin{cases} x^{\mathrm{ord}}_{\lceil np \rceil} \text{if } np \text{ is not an integer} \\ \\ \frac{1}{2}\left(x^{\mathrm{ord}}_{np} + x^{\mathrm{ord}}_{np+1}\right) \text{ if } np \text{ is an integer.} \end{cases}$$

  The median corresponds to $p = \frac{1}{2}$. Exercise. Prove this fact.

Now, we show that the 100p-th percentile given by the formal definition satisfies the informal definition.

Observe that:

- when np is an integer, the 100pth percentile $v = \frac{1}{2}\left(x_{np}^{\text{ord}} + x_{np+1}^{\text{ord}}\right)$ has the property

  $$x_1^{\text{ord}} \leq \cdots \leq x_{np}^{\text{ord}} \leq v \leq x_{np+1}^{\text{ord}} \leq \cdots \leq x_n^{\text{ord}}$$

  and the proportion of the components $x_1^{\text{ord}}, \ldots, x_{np}^{\text{ord}}$ over all components is

  $$\frac{np}{n} = p.$$

- when np is not an integer, the 100pth percentile $v = x_{\lceil np \rceil}^{\text{ord}}$ has the property

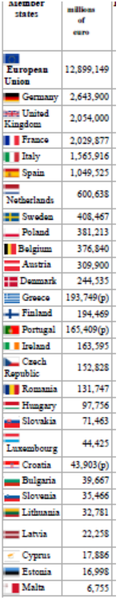
  $$x_1^{\text{ord}} \leq \cdots \leq x_{\lceil np \rceil - 1}^{\text{ord}} \leq v = x_{\lceil np \rceil}^{\text{ord}} \leq x_{\lceil np \rceil + 1}^{\text{ord}} \leq \cdots \leq x_n^{\text{ord}}$$

  and the proportion of the components $x_1^{\text{ord}}, \ldots, x_{\lceil np \rceil - 1}^{\text{ord}}$ over all components is

  $$\frac{\lceil np \rceil - 1}{n} \in \left(\frac{np - 1}{n}, \frac{np}{n} \right) = \left(p - \frac{1}{n}, p \right)$$

  (since $\lceil np \rceil - 1 < np < \lceil np \rceil$ and then $np - 1 < \lceil np \rceil - 1 < np$) and so this proportion is close to p.

Example: consider the 2012 Gross Domestic Product for countries in EU



Member states	millions of euro
European Union	12,899,149
Germany	2,643,900
United Kingdom	2,054,000
France	2,029,877
Italy	1,565,916
Spain	1,049,525
Netherlands	600,638
Sweden	408,467
Poland	381,213
Belgium	376,840
Austria	309,900
Denmark	244,535
Greece	193,749(p)
Finland	194,469
Portugal	165,409(p)
Ireland	163,595
Czech Republic	152,828
Romania	131,747
Hungary	97,756
Slovakia	71,463
Luxembourg	44,425
Croatia	43,903(p)
Bulgaria	39,667
Slovenia	35,466
Lithuania	32,781
Latvia	22,258
Cyprus	17,886
Estonia	16,998
Malta	6,755

We determine the 90*th percentile and the* 20*th percentile of the GDPs.*

We have: $n = 28$ *and*

$$np = \begin{cases} 25.2 \text{ if } p = 0.9 \\ 5.6 \text{ if } p = 0.2 \end{cases}$$

and so

$$\lceil np \rceil = \begin{cases} 26 \text{ if } p = 0.9 \\ 6 \text{ if } p = 0.2. \end{cases}$$

The 90*th percentile is 2,029.877 millions of euro (France's GDP) and the* 20*th percentile is 35.466 millions of euro (Slovenia's GDP).*

- The **quartiles** of $\boldsymbol{x}$ are the 25th, 50th and 75th percentiles of $\boldsymbol{x}$:
  - the 25th percentile is the **first quartile**;
  - the 50th percentile (the median) is the **second quartile**;
  - the 75th percentile is the **third quartile**.

  The quartiles break up $\boldsymbol{x}^{\mathrm{ord}}$ into four parts with:
  - about 25 percent of the components up to the first quartile; this part is known as the first lower quartile or the fourth upper quartile;
  - about 25 percent of the components between the first quartile and the second quartile; this part is known as the second lower quartile or the third upper quartile;
  - about 25 percent of the components between the second quartile and the third quartile; this part is known as the third lower quartile or the second upper quartile;
  - about 25 percent of the components after the third quartile; this part is known as the fourth lower quartile or the first upper quartile.

Exercise. Find on the web the FIFA World ranking for football national teams and determine the quartiles.

Exercise. What are the quintiles and the deciles and how many are they?

- Exercise. A mother takes her daughter to the pediatrician for a check up. The doctor measures the height of the daughter and notes that she is above the 95th percentile. What does it mean?

- In MATLAB, the 100*p*th percentile of the data in the vector x is computed by

  $$\mathrm{prctile}(\mathrm{x}, 100\mathrm{p}) \text{ or } \mathrm{quantile}(\mathrm{x}, \mathrm{p}).$$

  These two MATLAB functions $\mathrm{prctile}$ and $\mathrm{quantile}$ use a definition of percentile 100*p*th which is different from the definition given above. However, the values given by the two definitions are close since both are values which divide the data in two parts with number of components close to be proportional to p and $1 - p$.

  Use

  $$\mathrm{percentile}(\mathrm{x}, 100\mathrm{p})$$

  for obtaining the right values of our definition.

  Exercise. For $\boldsymbol{x} = (1, 2, 3, 4)$, compute $\mathrm{prctile}(\mathrm{x}, 30)$ and $\mathrm{percentile}(\mathrm{x}, 30)$.

Exercise. By using MATLAB, generate by

$$\mathrm{x} = \mathrm{rand}(1000, 1);$$

a data x of $n = 1000$ independent random numbers uniformly distributed on $[0, 1]$. Compute the 100pth percentile of x for $p = 0.1, 0.2, \ldots, 0.9$.

- Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$, let $p \in (0,1)$, let $c \in \mathbb{R}$ and let $\boldsymbol{y} = c + \boldsymbol{x}$. Prove that

  $$100p\text{th percentile of } \boldsymbol{y} = c + 100p\text{th percentile of } \boldsymbol{x}.$$

  Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$, let $p \in (0,1)$, let $c > 0$ and let $\boldsymbol{y} = c\boldsymbol{x}$. Prove that

  $$100p\text{th percentile of } \boldsymbol{y} = c \cdot 100p\text{th percentile of } \boldsymbol{x}.$$

  Exercise. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $\boldsymbol{x} = \boldsymbol{x}^{\mathrm{ord}}$ and $\boldsymbol{y} = \boldsymbol{y}^{\mathrm{ord}}$, let $p \in (0,1)$ and let $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$. Prove that

  $$100p\text{th percentile of } \boldsymbol{z} = 100p\text{th percentile of } \boldsymbol{x} + 100p\text{th percentile of } \boldsymbol{y}.$$

  Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$, let $p \in (0,1)$ and let $\boldsymbol{y} = -\boldsymbol{x}$. Prove that

  $$100p\text{th percentile of } \boldsymbol{y} = -100(1-p)\text{th percentile of } \boldsymbol{x}.$$

Mean and percentiles of a symmetric data

- Let $\boldsymbol{x} \in \mathbb{R}^n$ be a symmetric data about a number c.

  $\boldsymbol{x}^{\text{ord}}$ has the form

$$\boldsymbol{x}^{\text{ord}} = \Bigg(\underbrace{c - d_1, \ldots, c - d_1}_{f_1 \quad \text{times}}, \ldots, \underbrace{c - d_m, \ldots, c - d_m}_{f_m \text{ times}}, \underbrace{c, \ldots, c}_{f_{m+1} \quad \text{times}}$$

$$\underbrace{c + d_m, \ldots, c + d_m}_{f_m \quad \text{times}}, \ldots, \underbrace{c + d_1, \ldots, c + d_1}_{f_1 \quad \text{times}} \Bigg)$$

  For $i \in \{1, \ldots, n\}$, if $x_i^{\text{ord}} = c \mp d_{j_i}$, where $j_i \in \{1, \ldots, m+1\}$ with $d_{m+1} = 0$, then

$$x_{n+1-i}^{\text{ord}} = c \pm d_{j_i} = c \mp d_{j_i} \pm 2d_{j_i} = x_i^{\text{ord}} \pm 2d_{j_i} = x_i^{\text{ord}} + 2\left(c - x_i^{\text{ord}}\right)$$

  where the last equality follows by $\pm d_{j_i} = c - x_i^{\text{ord}}$.

- The mean of $\boldsymbol{x}$ is c.

  In fact,

$$\begin{aligned}
\overline{x} = \overline{x}^{\text{ord}} &= \frac{\sum_{i=1}^{n} x_i^{\text{ord}}}{n} \\
&= \frac{f_1\,(c - d_1) + \cdots + f_m\,(c - d_m) + f_{m+1}\,c + f_m\,(c + d_m) + \cdots + f_1\,(c + d_1)}{n} \\
&= \frac{(2f_1 + \cdots + 2f_m + f_{m+1})\,c}{n} = \frac{nc}{n} = c.
\end{aligned}$$

- Let $p \in (0,1)$. The $100p$-th and $100\,(1-p)$-th percentiles of $\boldsymbol{x}$ are symmetric about c, i.e. .

$$\begin{aligned} &100\,(1-p)\text{-th percentile of } \boldsymbol{x} \\ &= 100p\text{-th percentile of } \boldsymbol{x} + 2\,(c - 100p\text{-th percentile of } \boldsymbol{x}) \end{aligned}$$

In fact, if np is an integer, then $n\,(1-p) = n - np$ is an integer and so

$$\begin{aligned} &100\,(1-p)\text{-th percentile of } \boldsymbol{x} \\ &= \frac{1}{2}\left(x^{\text{ord}}_{n-np} + x^{\text{ord}}_{n-np+1}\right) \\ &= \frac{1}{2}\left(x^{\text{ord}}_{n+1-(np+1)} + x^{\text{ord}}_{n+1-np}\right) \\ &= \frac{1}{2}\left(x^{\text{ord}}_{np+1} + 2\left(c - x^{\text{ord}}_{np+1}\right) + x^{\text{ord}}_{np} + 2\left(c - x^{\text{ord}}_{np}\right)\right) \\ &= \frac{1}{2}\left(x^{\text{ord}}_{np} + x^{\text{ord}}_{np+1}\right) + 2\left(c - \frac{1}{2}\left(x^{\text{ord}}_{np} + x^{\text{ord}}_{np+1}\right)\right) \\ &= 100p\text{-th percentile of } \boldsymbol{x} + 2\,(c - 100p\text{-th percentile of } \boldsymbol{x})\,. \end{aligned}$$

If np is not an integer, then $n(1-p) = n - np$ is not an integer. We have

$$\lceil np \rceil - 1 < np < \lceil np \rceil$$

and then

$$n - \lceil np \rceil < n - np < n - (\lceil np \rceil - 1) = n - \lceil np \rceil + 1.$$

This means that

$$\lceil n - np \rceil = n - \lceil np \rceil + 1$$

and so

$$\begin{aligned} & 100(1-p)\text{-th percentile of } \boldsymbol{x} \\ & = x^{\text{ord}}_{\lceil n-np \rceil} = x^{\text{ord}}_{n+1-\lceil np \rceil} \\ & = x^{\text{ord}}_{\lceil np \rceil} + 2\left(c - x^{\text{ord}}_{\lceil np \rceil}\right) \\ & = 100p\text{-th percentile of } \boldsymbol{x} + 2(c - 100p\text{-th percentile of } \boldsymbol{x}). \end{aligned}$$

Exercise. Prove that the median of $\boldsymbol{x}$ is c.

- *Example: consider the symmetric data around* 5 *with frequency table*

Value v_j	*Frequency f_j*
2	2
3	1
4	2
5	2
6	2
7	1
8	2

The ordered n-tuple, where $n = 12$, is

$$\boldsymbol{x}^{\mathrm{ord}} = (2, 2, 3, 4, 4, 5, 5, 6, 6, 7, 8, 8)\,.$$

The mean is

$$\overline{x} = 60/12 = 5.$$

The quartiles are

$$\textit{first quartile, } np = 3\textit{:}\ \frac{1}{2}\left(x_3^{\mathrm{ord}} + x_4^{\mathrm{ord}}\right) = \frac{1}{2}(3+4) = 3.5 = 5 - 1.5$$

$$\textit{median, } np = 6\textit{:}\ \frac{1}{2}\left(x_6^{\mathrm{ord}} + x_7^{\mathrm{ord}}\right) = \frac{1}{2}(5+5) = 5$$

$$\textit{third quartile, } np = 9\textit{:}\ \frac{1}{2}\left(x_9^{\mathrm{ord}} + x_{10}^{\mathrm{ord}}\right) = \frac{1}{2}(6+7) = 6.5 = 5 + 1.5.$$

Mode

- Let $\{v_1, \ldots, v_l\}$ be the (non-necessarily numeric) data values of the data $\boldsymbol{x}$. The **mode** of $\boldsymbol{x}$ is the data value with the highest frequency.

  Example: consider the weather in last two weeks in a given town:

  $$\boldsymbol{x} = (C, C, S, S, S, C, S, R, R, R, C, S, S, S)$$

  where C stands for "Cloudy", S for "Sunny" and R for "Rainy". The frequency table is

v_j	f_j
C	4
S	7
R	3

  and so the mode is S.

If there are more data values that occur most frequently, then all these values are called **modal values.**

Example: consider

$$\boldsymbol{x} = (4, 3, 5, 6, 5, 4, 5, 6, 3, 3, 4, 5, 3)\,.$$

The frequency table is

v_j	f_j
3	4
4	3
5	4
6	2

and so 3 *and* 5 *are modal values.*

- Similarly to the mean and the median, the mode is a measure of the central tendency of the data.

  But, unlike the mean and the median, the mode is defined also in case of non-numeric data values.

- In MATLAB, the mode of the data in the vector x, whose components are numbers, is computed by
$$\mathrm{mode(x)}.$$
If there are more than one modal values, the smallest one is returned.

  Given a data with frequencies in the vector f and the data values in the vector v, all the modal values can be obtained in MATLAB by
$$\mathrm{I = find(f == max(f));\ v(I)}.$$

  Exercise. Compute $\mathrm{mode(x)}$ and find the modal values for the data ***x*** of the previous example.

- Exercise. Throw a normal die fifty times and find the mean, the median and the mode of the obtained scores. If a die is not available, use the MATLAB function die that simulates a die.

  Exercise. For $\boldsymbol{x} \in \mathbb{R}^n$ symmetric data around c, can we say that the mode of $\boldsymbol{x}$ is equal to c?

  Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$ be a symmetric data around c. Prove that c is a modal value of $\boldsymbol{x}$ if and only if $\boldsymbol{x}$ has a odd number of modal values.

Variance

- Up to now, we have introduced statistics that measure the central tendency of the data.

  Now, we pass to consider statistics that measure the spread, or variability, or dispersion, of the data.

  Example: consider the two data

  $$\boldsymbol{a} = (1, 2, 5, 6, 6)\,,\ \boldsymbol{b} = (-40, 0, 5, 20, 35)\,,$$

  both of size $n = 5$. We have

  $$\overline{a} = \overline{b} = 4,\ m_a = m_b = 5$$

  but there is clearly more spread in ***b*** *than in* ***a****.*

When the mean $\overline{x}$ is used as a measure of the central tendency of a data $\boldsymbol{x} \in \mathbb{R}^n$, one way of measuring the variability of $\boldsymbol{x}$ is to consider the deviations

$$x_i - \overline{x}, \ i \in \{1, \ldots, n\},$$

from the mean.

But, we cannot use

$$\sum_{i=1}^{n} (x_i - \overline{x}),$$

or the average

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}),$$

as a measure of the variability, since we have

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0.$$

We have to consider the absolute values of the deviations.

So, measures of the variability of $\boldsymbol{x}$ could be

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$$

and

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 .$$

We prefer the second measure, because it is mathematically more tractable.

However, for technical reasons (which will become clear in the following), we divide the sum of the squares of the deviations by $n-1$, rather than n.

Observe that $n-1$ is the number of degrees of freedom of the deviations $x_i - \overline{x}$, $i \in \{1, \dots, n\}$: $n-1$ deviations can be arbitrarily fixed and then the last is determined by $\sum\limits_{i=1}^{n}(x_i - \overline{x}) = 0$.

The **variance** of the data $\boldsymbol{x} \in \mathbb{R}^n$ is given by

$$s_x^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \,.$$

Example:

- *for the data* $\boldsymbol{a} = (1, 2, 5, 6, 6)$ *we have*

$$\begin{aligned} \overline{a} &= 4 \\ \boldsymbol{a} - \overline{a} &= (-3, -2, 1, 2, 2) \\ s_a^2 &= \frac{1}{4}\left(3^2 + 2^2 + 1^2 + 2^2 + 2^2\right) = \frac{22}{4} = 5.5; \end{aligned}$$

- *for the data* $\boldsymbol{b} = (-40, 0, 5, 20, 35)$ *we have*

$$\begin{aligned} \overline{b} &= 4 \\ \boldsymbol{b} - \overline{b} &= (-44, -4, 1, 16, 31) \\ s_b^2 &= \frac{1}{4}\left(44^2 + 4^2 + 1^2 + 16^2 + 31^2\right) = \frac{3170}{4} = 792.5. \end{aligned}$$

- A physical interpretation of the variance: if n objects of the same mass m are placed in a rod at the positions x_i, $i \in \{1, \ldots, n\}$, then

  $$I = \sum_{i=1}^{n} m\,(x_i - \overline{x})^2 = (n-1)ms_x^2$$

  is the moment of inertia of the system with respect to an axis perpendicular to the rod and passing through the center of mass $\overline{x}$.

  So, the variance

  $$s_x^2 = \frac{1}{(n-1)m} \cdot I$$

  is proportional to the moment of inertia I.

- An useful relation for computing variances is

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2.$$

In the physical interpretation of variances as moments of inertia, this relation corresponds to the "Parallel Axis Theorem" or "Steiner's Theorem" for the moments of inertia.

Exercise. Explain this correspondence.

Here is the proof of the previous relation:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\overline{x})^2 &= \sum_{i=1}^{n}\left(x_i^2-2x_i\overline{x}+\overline{x}^2\right)\\
&= \sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}(-2x_i\overline{x})+\sum_{i=1}^{n}\overline{x}^2\\
&= \sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}^2\\
&= \sum_{i=1}^{n}x_i^2-2\overline{x}\cdot n\overline{x}+n\overline{x}^2\\
&= \sum_{i=1}^{n}x_i^2-n\overline{x}^2.
\end{aligned}$$

With this relation we have

$$s_x^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right).$$

- Here are two important properties of the variance. Let $\boldsymbol{x} \in \mathbb{R}^n$ be a data.

  Let $c \in \mathbb{R}$. Consider the $n-$tuple $\boldsymbol{y} = c + \boldsymbol{x}$. We have

  $$s_y^2 = s_x^2.$$

  Let $c \in \mathbb{R}$. Consider the $n-$tuple $\boldsymbol{y} = c\boldsymbol{x}$. We have

  $$s_y^2 = c^2 s_x^2.$$

  Exercise. Prove these properties.

The first property can be used to simplify computations.

Example: let

$$\boldsymbol{y} = (175, 184, 186, 183, 178, 180, 177, 179, 189, 185)$$

be the heights in cm of $n = 10$ *males.*

By writing

$$\begin{aligned} \boldsymbol{y} &= 180 + (-5, 4, 6, 3, -2, 0, -3, -1, 9, 5) \\ &= 180 + \boldsymbol{x} \end{aligned}$$

we have

$$\begin{aligned} \overline{y} &= 180 + \overline{x} = 180 + \frac{16}{10} = 180 + 1.6 = 181.6 \\ s_y^2 &= s_x^2 = \frac{1}{9}\Big(5^2 + 4^2 + 6^2 + 3^2 + 2^2 + 0^2 + 3^2 + 1^2 + 9^2 + 5^2 \\ &\qquad -10 \cdot 1.6^2\Big) \text{ (we are using the Parallel Axis Theorem)} \\ &= 20.04. \end{aligned}$$

- Exercise. Find a formula for the variance of a data in terms of its data values v_j and frequencies f_j, $j \in \{1, \ldots, l\}$.

  Exercise. When a data has zero variance?

  Exercise. Sometimes, in descriptive Statistics, the definition of variance is given by dividing by n, rather than $n - 1$, the sum of the squares of the deviations. What is the relative error

  $$\frac{\widehat{s}_x^2 - s_x^2}{s_x^2}$$

  of $\widehat{s}_x^2$, the variance with division by n, with respect to s_x^2, the variance with division by $n - 1$?

- In MATLAB, the variance of the data in the vector x is computed by

$$\mathrm{var(x)}.$$

Exercise. By using MATLAB compute the variance of the heights in the previous example.

- Exercise. Compute the variance of the scores obtained by throwing a die fifty times.

  Then, for fifty times, throw a die fifty times. For each time, compute the mean of the scores. Then, compute the variance of the fifty computed means.

  Try to explain why the second variance is much smaller than the first one.

Standard deviation

- The **standard deviation** (from the mean) of the data $\boldsymbol{x} \in \mathbb{R}^n$ is given by

$$s_x := \sqrt{s_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Unlike s_x^2, s_x has the same dimensions as the components of the data.

Example. The standard deviation for the example of the heights is $s_y = \sqrt{20.04 \text{ cm}^2} = 4.48 \text{ cm}$.

The standard deviation is more appropriate than the variance as a measure of the variability of the data. Since standard deviation and mean have the same dimensions, they can be compared. On the other hand, it is not possible to compare variance and standard deviation because they have different dimensions.

- Properties of the standard deviation: for a data $\boldsymbol{x} \in \mathbb{R}^n$ and $c \in \mathbb{R}$, we have

$$s_{c+\boldsymbol{x}} = s_x \quad \text{and} \quad s_{c\boldsymbol{x}} = |c| \, s_x .$$

Exercise. Prove these properties of the standard deviation.

- Exercise. What is the relative error

  $$\frac{\widehat{s}_x - s_x}{s_x}$$

  of $\widehat{s}_x$, the standard deviation with division by n, with respect to s_x, the standard deviation with division by $n-1$?

- In MATLAB, the standard deviation of the data in the vector x is computed by
$$\mathrm{std(x)} \ \text{ or } \ \mathrm{sqrt(var(x))}.$$

  Exercise. By using MATLAB compute the standard deviation for the example of the heights.

Interquartile range

- Another indicator of the variability of the data $\boldsymbol{x} \in \mathbb{R}^n$ is the **interquartile range** of $\boldsymbol{x}$ given by

$$\text{interquartile range of } \boldsymbol{x} := \text{third quartile of } \boldsymbol{x} - \text{first quartile of } \boldsymbol{x}.$$

Roughly speaking, the interquartile range of $\boldsymbol{x}$ is the length of the interval in which the middle half of the components of $\boldsymbol{x}$ lie: 25% of the components between the first quartile and the median and other 25% between the median and the third quartile.

The interquartile range should be used as indicator of the variability when the median is used as indicator of the center, whereas the standard deviation should be used as indicator of the variability when the mean is used as indicator of the center.

Example: consider the data

$$\boldsymbol{y} = (175, 184, 186, 183, 178, 180, 177, 179, 189, 185)$$

of the heights of size $n = 10$*. We have*

$$\boldsymbol{y}^{\mathrm{ord}} = (175, 177, 178, 179, 180, 183, 184, 185, 186, 189)\,.$$

The first, second and third quartiles are

$$\begin{aligned}
&\textit{for } p = 0.25 : y^{\mathrm{ord}}(\lceil 2.5 \rceil) = y^{\mathrm{ord}}(3) = 178\\
&\textit{for } p = 0.5\ \ : \frac{1}{2}\left(y^{\mathrm{ord}}(5) + y^{\mathrm{ord}}(6)\right) = \frac{1}{2}(180 + 183) = 181.5\\
&\textit{for } p = 0.75 : y^{\mathrm{ord}}(\lceil 7.5 \rceil) = y^{\mathrm{ord}}(8) = 185.
\end{aligned}$$

The median is 181.5 *and the interquartile range is* $185 - 178 = 7$*.*

The median $181.5\ \mathrm{cm}$ *and the interquartile range* $7\ \mathrm{cm}$ *should be compared with the mean* $181.6\ \mathrm{cm}$ *and the standard deviation* $4.48\ \mathrm{cm}$*.*

- In MATLAB, the interquartile range of a data in the vector x is computed by

$$\mathrm{iqr}(\mathrm{x}) \text{ or } \mathrm{prctile}(\mathrm{x}, 75) - \mathrm{prctile}(\mathrm{x}, 25).$$

Use

$$\mathrm{percentile}(\mathrm{x}, 75) - \mathrm{percentile}(\mathrm{x}, 25)$$

for percentiles computed by following our definition.

Exercise. By using MATLAB compute the interquartile range of the heights in the previous example.

- Exercise. Find on

  http://www.ilmeteo.it/portale/archivio-meteo

  the maximum temperatures in Pordenone during January, Febraury, March and April 2017. Determine for each month the mean, the median, the standard deviation and the interquartile range.

- Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$, let $c \in \mathbb{R}$ and let $\boldsymbol{y} = c + \boldsymbol{x}$. Prove that

  $$\text{interquartile range of } \boldsymbol{y} = \text{ interquartile range of } \boldsymbol{x}.$$

  Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$, let $c > 0$ and let $\boldsymbol{y} = c\boldsymbol{x}$. Prove that

  $$\text{interquartile range of } \boldsymbol{y} = c \cdot \text{ interquartile range of } \boldsymbol{x}.$$

  Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$ and let $\boldsymbol{y} = -\boldsymbol{x}$. Prove that

  $$\text{interquartile range of } \boldsymbol{y} = \text{ interquartile range of } \boldsymbol{x}.$$

  Exercise. Let $\boldsymbol{x} \in \mathbb{R}^n$, let $c \in \mathbb{R}$ and let $\boldsymbol{y} = c\boldsymbol{x}$. Prove that

  $$\text{interquartile range of } \boldsymbol{y} = |c| \cdot \text{ interquartile range of } \boldsymbol{x}.$$

  Observe that the properties in the first and in the last of the previous exercises are also satisfied by the standard deviation.

Exercise. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $\boldsymbol{x} = \boldsymbol{x}^{\mathrm{ord}}$ and $\boldsymbol{y} = \boldsymbol{y}^{\mathrm{ord}}$, and let $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$. Prove that

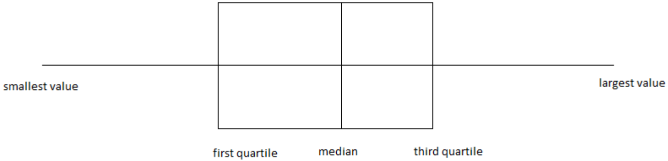
interquartile range of $\boldsymbol{z} =$ interquartile range of $\boldsymbol{x}+$ interquartile range of $\boldsymbol{y}$.

This property is not satisfied by the standard deviation: there exists data $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $s_z \neq s_x + s_y$. Check this by considering $\boldsymbol{x} \in \mathbb{R}^n$ with $s_x \neq 0$ and $\boldsymbol{y} = -\boldsymbol{x}$.

Box plot

- A **box plot** is used to plot some of the summarizing statistics of a data $\boldsymbol{x} \in \mathbb{R}^n$.

  It is composed by a segment on the real line with extremes the smallest data value and the largest data value. Imposed on this segment, there is a "box", that starts at the first quartile and finishes at the third quartile, with the value of the median indicated by a line perpendicular to the segment.

  A box plot is also called a **box and whiskers plot**: the "whiskers" are the segments exiting from the box that reach the extremes.

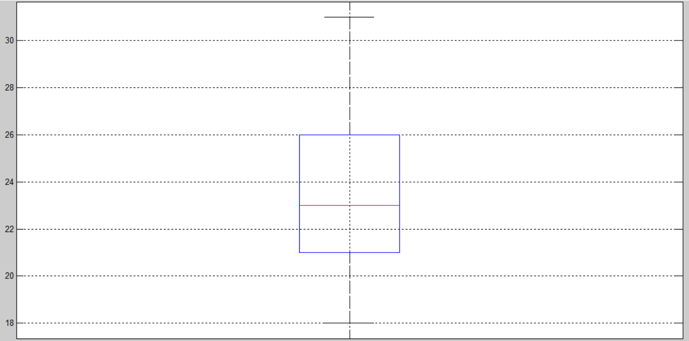
Example: consider the following frequency table for positive marks in a Trieste University exam (Numerical Analysis).

v_j	f_j
18	2
19	2
20	4
21	3
22	5
23	4
24	4
25	3
26	3
27	4
28	2
29	0
30	1
$31 = 30L$	1

Since $n = 38$*, the quartiles are*

$$p = 0.25 : x^{\text{ord}}\left(\lceil 9.5 \rceil\right) = x^{\text{ord}}\left(10\right) = 21$$
$$p = 0.5 : \frac{x^{\text{ord}}\left(19\right) + x^{\text{ord}}\left(20\right)}{2} = 23$$
$$p = 0.75 : x^{\text{ord}}\left(\lceil 28.5 \rceil\right) = x^{\text{ord}}\left(29\right) = 26.$$

Box plot:



- In MATLAB, the box plot of the data in the vector x is created by

$$\mathrm{boxplot(x)}.$$

Exercise. By using MATLAB create the box plot of the marks in the previous example.

- Exercise. Find on Wikipedia the list of the US presidents and then construct a box plot for the ages when they became president (for the first time).

Mean and percentiles of an histogram

- Consider a data $\boldsymbol{x} \in \mathbb{R}^n$ ordered in increasing order, i.e. $\boldsymbol{x} = \boldsymbol{x}^{\mathrm{ord}}$, and a histogram for this data with K class intervals.

  Based on the histogram, we introduce the data $\boldsymbol{x}^{\mathrm{hist}}$ ordered in increasing order, whose frequency table is given by the pairs

  $$(v_k, f_k)\,,\ k \in \{1, \dots, K\}\,,$$

  where v_k and f_k are the the middle point and the frequency, respectively, of the k-th class interval.

  Exercise. What is the number of components of $\boldsymbol{x}^{\mathrm{hist}}$?

Example: consider as an histogram the following stem-and-leaf plot

0	1	2	3
1	5	6	
2	2	9	
3	4	6	

The data is

$$\boldsymbol{x} = (1, 2, 3, 15, 16, 22, 29, 34, 36)\,.$$

The data $\boldsymbol{x}^{\text{hist}}$ has frequency table

v_j	f
5	3
15	2
25	2
35	2

and so

$$\boldsymbol{x}^{\text{hist}} = (5, 5, 5, 15, 15, 25, 25, 35, 35)\,.$$

- Observe that $\boldsymbol{x}^{\mathrm{hist}}$ is obtained from $\boldsymbol{x}$ by substituting each components x_i, $i \in \{1, \ldots, n\}$, with v_{k_i}, i.e. $x_i^{\mathrm{hist}} = v_{k_i}$, where v_{k_i} is the middle point of the class interval containing x_i with $k_i \in \{1, \ldots, K\}$ the index of this class interval.

  For $i \in \{1, \ldots, n\}$, since v_{k_i} is the middle point of the class interval containing x_i we have

$$\left|x_i^{\mathrm{hist}} - x_i\right| = \left|v_{k_i} - x_i\right| \leq \frac{h}{2},$$

  where h is the length of the class intervals.

- The mean of $\boldsymbol{x}^{\mathrm{hist}}$ is close to the mean of $\boldsymbol{x}$.

  In fact

$$\begin{aligned}\left|\overline{x}^{\mathrm{hist}} - \overline{x}\right| &= \left|\frac{\sum_{i=1}^{n} x_i^{\mathrm{hist}} - \sum_{i=1}^{n} x_i}{n}\right| = \left|\frac{\sum_{i=1}^{n} \left(x_i^{\mathrm{hist}} - x_i\right)}{n}\right| \\ &\le \frac{\sum_{i=1}^{n} \left|x_i^{\mathrm{hist}} - x_i\right|}{n} \le \frac{\sum_{i=1}^{n} \frac{h}{2}}{n} = \frac{h}{2}.\end{aligned}$$

- The $100p$th percentile, $p \in (0,1)$, of $\boldsymbol{x}^{\text{hist}}$ is close to the $100p$-th percentile of $\boldsymbol{x}$.

  In fact, if np is an integer, then (recall that $\boldsymbol{x}$ and $\boldsymbol{x}^{\text{hist}}$ are ordered in increasing order)

$$
\begin{aligned}
&\left|100p\text{th percentile of } \boldsymbol{x}^{\text{hist}} - 100p\text{-th percentile of } \boldsymbol{x}\right| \\
&= \left|\frac{1}{2}\left(x_{np}^{\text{hist}} + x_{np+1}^{\text{hist}}\right) - \frac{1}{2}\left(x_{np} + x_{np+1}\right)\right| \\
&= \left|\frac{1}{2}\left(x_{np}^{\text{hist}} - x_{np}\right) + \frac{1}{2}\left(x_{np+1}^{\text{hist}} - x_{np+1}\right)\right| \\
&\leq \frac{1}{2}\left|x_{np}^{\text{hist}} - x_{np}\right| + \frac{1}{2}\left|x_{np+1}^{\text{hist}} - x_{np}\right| \\
&\leq \frac{1}{2}\cdot\frac{h}{2} + \frac{1}{2}\cdot\frac{h}{2} = \frac{h}{2}.
\end{aligned}
$$

  if np is not an integer, then

$$
\begin{aligned}
&\left|100p\text{th percentile of } \boldsymbol{x}^{\text{hist}} - 100p\text{th percentile of } \boldsymbol{x}\right| \\
&= \left|x_{\lceil np \rceil}^{\text{hist}} - x_{\lceil np \rceil}\right| \leq \frac{h}{2}.
\end{aligned}
$$

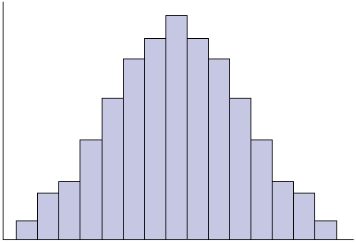
- Exercise. Consider as an histogram the following stem-and-leaf plot of a data $\boldsymbol{x}$

0	2	2	2	2	4	7	7			
1	0	3	5	5	6	8				
2	2	4	4	6	8	8	9			
3	2	3	3	7	8					
4	2	5	5	8						
5	1	1	2	3	4	5	6	6	7	9
6	0	1	3	5	6	7	9	9	9	
7	3	5	6	8	8	9				
8	4	5	7	7						
9	2	7	8							

Compute the mean, the quartiles and the modal values of $\boldsymbol{x}$ and $\boldsymbol{x}^{\text{hist}}$.

Normal data

- A data $\boldsymbol{x}$ is said **normal** (or **gaussian**) if it has a histogram with the following characteristics:
  - the histogram is symmetric with respect to the middle interval, i.e. $\boldsymbol{x}^{\text{hist}}$ is symmetric about the middle point c of the middle interval;
  - the middle interval has the highest frequency, i.e. $\boldsymbol{x}^{\text{hist}}$ has mode c;
  - the frequencies of the histogram decrease from the middle interval in a bell-shaped fashion (explained below).

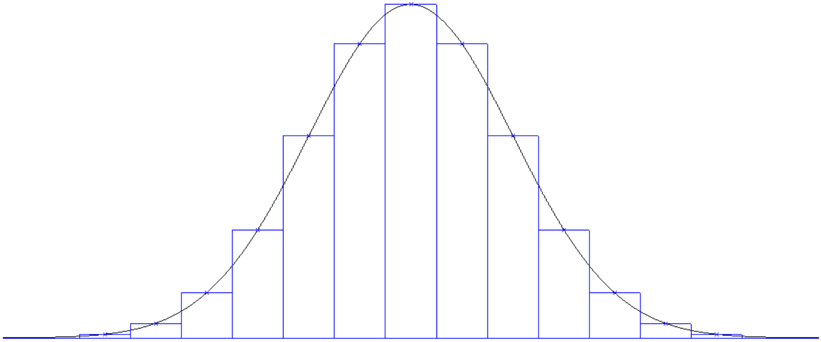


As a consequence of the first characteristic, we have that mean and median are close each other.

In fact, $\boldsymbol{x}^{\mathrm{hist}}$ is symmetric about c and then mean and median of $\boldsymbol{x}^{\mathrm{hist}}$ are equal to c. Moreover, the mean of $\boldsymbol{x}$ is close to the mean of $\boldsymbol{x}^{\mathrm{hist}}$ and the median of $\boldsymbol{x}$ is close to the median of $\boldsymbol{x}^{\mathrm{hist}}$.

So, mean and median of $\boldsymbol{x}$ are both close to c and then they are close each other.

Qualitative description of the bell-shaped fashion decrease: in both sides, the curve interpolating the frequencies at the middle points of the intervals starts with an horizontal tangent at the middle point of the middle interval, then it is concave and, after a flex point, it becomes convex and goes asymptotically to zero.



Quantitative description of the bell-shaped fashion decrease: for any $k > 0$, we can give a percentage $C(k)\%$ such that

> in the bell-shaped fashion decrease, approximately $C(k)\%$ of the components of $\boldsymbol{x}$ lie between $\overline{x} - ks_x$ and $\overline{x} + ks_x$, i.e. within k standard deviations s_x from the mean $\overline{x}$.

The percentages $C(k)\%$ for all $k > 0$ will be given later in the course. Here, we only observe that in the bell-shaped fashion decrease:

1) approximately $C(1)\% = 68\%$ of the components of $\boldsymbol{x}$ lie between $\overline{x} - s_x$ and $\overline{x} + s_x$, i.e. within one standard deviation from the mean;

2) approximately $C(2)\% = 95\%$ of the components of $\boldsymbol{x}$ lie between $\overline{x} - 2s_x$ and $\overline{x} + 2s_x$, i.e. within two standard deviations from the mean;

3) approximately $C(3)\% = 99.7\%$ of the components of $\boldsymbol{x}$ lie between $\overline{x} - 3s_x$ and $\overline{x} + 3s_x$, i.e. within three standard deviations from the mean.

When we are checking for the bell-shaped fashion decrease, we can check only the three previous points 1), 2) and 3), that constitute the so-called **empirical rule** for the bell-shaped fashion decrease.

- Saying that

  approximately $C(k)\%$ of the components of $\boldsymbol{x}$ lie between $\overline{x} - ks_x$ and $\overline{x} + ks_x$

  is equivalent to saying that

  the 100pth and $100(1-p)$th percentiles of $\boldsymbol{x}$, where

  $$100p = \frac{100 - C(k)}{2},$$

  are approximately $\overline{x} - ks_x$ and $\overline{x} + ks_x$, respectively.

We prove this equivalence.

Consider a data $\boldsymbol{x}$ with the first characteristic required for a normal data.

Since $\boldsymbol{x}^{\mathrm{hist}}$ is symmetric about c, the percentiles $100p$th and $100(1-p)$th of $\boldsymbol{x}^{\mathrm{hist}}$, $p \in (0, \frac{1}{2}]$, have the form $c - w_p$ and $c + w_p$, respectively, for some nonnegative number w_p depending on p

Since the mean c of $\boldsymbol{x}^{\mathrm{hist}}$ is close to the mean $\overline{x}$ of $\boldsymbol{x}$ and the percentiles of $\boldsymbol{x}^{\mathrm{hist}}$ are close to the percentiles of $\boldsymbol{x}$, we have that the percentiles $100p$th and $100(1-p)$th of $\boldsymbol{x}$, $p \in (0, \frac{1}{2}]$, have approximately the form $\overline{x} - w_p$ and $\overline{x} + w_p$, respectively, for some nonnegative number w_p.

Now, for a given $p \in (0, \frac{1}{2}]$, let v_1 and v_2 be the 100pth and $100(1-p)$th percentiles of $\boldsymbol{x}$. These percentiles are approximately of the form $\overline{x} - w_p$ and $\overline{x} + w_p$, respectively, for some nonnegative number w_p.

We have that approximately $100p\%$ of the components of $\boldsymbol{x}$ are $\leq v_1$ and $100\,(1-p)\,\%$ of the components are $\leq v_2$.

So, approximately

$$C\% = 100\,(1-p)\,\% - 100p\% = (100 - 2 \cdot 100p)\,\%$$

of the components of $\boldsymbol{x}$ are between v_1 and v_2 and then approximately $C\%$ of the components of $\boldsymbol{x}$ are between $\overline{x} - w_p$ and $\overline{x} + w_p$.

Now, we are ready to prove the equivalence. Assume that approximately $C(k)\%$ of the components of $\boldsymbol{x}$ lie between $\overline{x} - ks_x$ and $\overline{x} + ks_x$.

Let $p \in (0, \frac{1}{2}]$ such that

$$100p = \frac{100 - C(k)}{2}.$$

Then, approximately

$$C\% = (100 - 2 \cdot 100p)\,\% = C(k)\%$$

of the components of $\boldsymbol{x}$ are between $\overline{x} - w_p$ and $\overline{x} + w_p$.

Since we have approximately $C(k)\%$ of the components of $\boldsymbol{x}$ between $\overline{x} - w$ and $\overline{x} + w$ only for w approximately equal to ks_x, w_p has to be approximately equal to ks_x.

Since the percentiles $100p$th and $100(1 - p)$th of $\boldsymbol{x}$ are approximately $\overline{x} - w_p$ and $\overline{x} + w_p$, respectively, we conclude that they are also approximately $\overline{x} - ks_x$ and $\overline{x} + ks_x$, respectively.

Viceversa, assume that the $100p$th and $100(1-p)$th percentiles of $\boldsymbol{x}$, where

$$100p = \frac{100 - C(k)}{2},$$

are approximately $\overline{x} - ks_x$ and $\overline{x} + ks_x$, respectively.

Let v_1 be the $100p$th percentile of $\boldsymbol{x}$ and let v_2 be the $100(1-p)$th percentile of $\boldsymbol{x}$.

Approximately

$$C\% = (100 - 2 \cdot 100p)\,\% = C(k)\%$$

of the components of $\boldsymbol{x}$ are between v_1 and v_2.

Since v_1 and v_2 are approximately $\overline{x} - ks_x$ and $\overline{x} + ks_x$, respectively, we conclude that approximately $C(k)\%$ of the components of $\boldsymbol{x}$ are between $\overline{x} - ks_x$ and $\overline{x} + ks_x$.

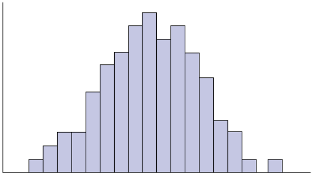
By this equivalence, we can restate the points 1), 2) and 3) of the empirical rule as

1bis) The $\frac{100-C(1)}{2} = \frac{100-68}{2} = 16$th and $100 - 16 = 84$th percentiles of $\boldsymbol{x}$ are approximately $\overline{x} - s_x$ and $\overline{x} + s_x$, respectively;

2bis) The $\frac{100-C(2)}{2} = \frac{100-95}{2} = 2.5$th and $100 - 2.5 = 97.5$th percentiles of $\boldsymbol{x}$ are approximately $\overline{x} - 2s_x$ and $\overline{x} + 2s_x$, respectively;

3bis) The $\frac{100-C(3)}{2} = \frac{100-99.7}{2} = 0.15$th and $100 - 0.15 = 99.85$th percentiles of $\boldsymbol{x}$ are approximately $\overline{x} - 3s_x$ and $\overline{x} + 3s_x$, respectively.

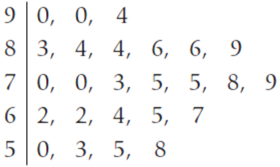
- A data $\boldsymbol{x}$ is said **approximately normal** if it has a histogram with the following characteristics:
  - the histogram is approximately symmetric with respect to the middle interval;
  - the highest frequency is approximately in the middle interval;



  - the percentages $C(k)\%$, $k > 0$, of the components of $\boldsymbol{x}$ between $\overline{x} - k \cdot s_x$ and $\overline{x} - k \cdot s_x$ are approximately satisfied worse than a normal data.

A normal data is only a theoretical notion. Only approximately normal data can be encountered in the real world.

Example: the following stem-and-leaf plot shows the scores of $n = 25$ candidates in a public concourse in Italy:



9	0, 0, 4
8	3, 4, 4, 6, 6, 9
7	0, 0, 3, 5, 5, 8, 9
6	2, 2, 4, 5, 7
5	0, 3, 5, 8

By turning this plot on its side, we can see that the corresponding histogram satisfies the first two characteristics required to an approximately normal data.

A confirmation that the data $\boldsymbol{x}$ *satisfies the first characteristic is the closeness of mean and median. In fact:*

$$m_x = x_{13}^{\mathrm{ord}} = 75, \; \overline{x} = 73.68.$$

We use the empirical rule to check the bell-shaped fashion decrease. We have $s_x = 12.80$.

Between $\overline{x} - s_x$ and $\overline{x} + s_x$, i.e. between 60.88 and 86.48, we should find approximately the 68% of the components : indeed $\frac{17}{25} = 68\%$ of the components is between these numbers and

$$16th\ percentile = \frac{1}{2}\left(x_4^{\mathrm{ord}} + x_5^{\mathrm{ord}}\right) = 60, \quad 25 \cdot 0.16 = 4 \ is\ an\ integer,$$

and

$$84th\ percentile = x^{\mathrm{ord}}_{\lceil 25 \cdot 0.84 \rceil} = x^{\mathrm{ord}}_{21} = 86.$$

Between $\overline{x} - 2s_x$ and $\overline{x} + 2s_x$, i.e. between 48.08 and 99.28, we should find approximately the 95% of the components: indeed, all the components lie within these numbers.

Exercise. Check whether the maximum temperatures in Pordenone during January, February, March and April 2017 found in

http://www.ilmeteo.it/portale/archivio-meteo

are approximately normal.

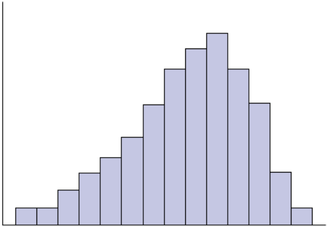
Exercise. Find on the web site

http://www.accessoprogrammato.miur.it

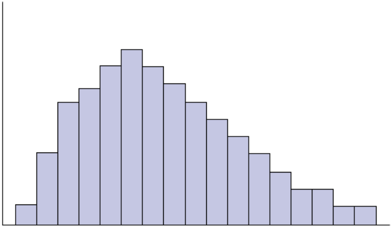
the results of the admission test to the Doctor of Medicine degree for the academic year 2017-2018 at the University of Trieste and check whether they are approximately normal.

Other types of data

- A data $\boldsymbol{x} \in \mathbb{R}^n$ is said **skewed** if it has a histogram with the following characteristics:
  - there is a class interval with the highest frequency, that is shifted with respect to the middle interval;
  - the histogram decreases in the longest side (the skewed side) more slowly than the other side, with a long tail.

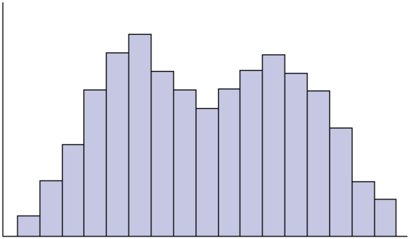


skewed to the left.



skewed to the right.

- A data $\boldsymbol{x} \in \mathbb{R}^n$ is said **bimodal** if it has a histogram



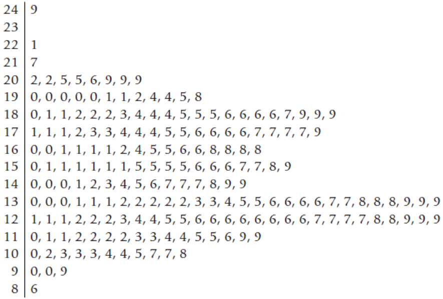
with two local peaks (local maximum) of frequency.

The use of the adjective "bimodal" comes from the fact these two local peaks of frequency are "local" modal values of $\boldsymbol{x}^{\text{hist}}$.

Exercise. What is the definition of m-modal data, where $m \in \{1, 2, 3, \dots\}$?

A bimodal data ***z*** appears when there is a superposition of two approximately normal data ***x*** and ***y***, i.e. the components of ***z*** can be partitioned in two parts that constitute ***x*** and ***y***.

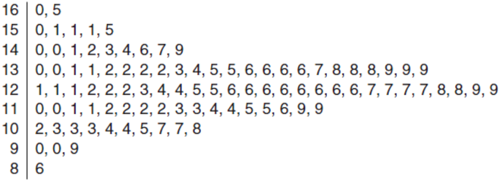
Example : the following stem-and-leaf plot gives the weights in pounds ($1\text{pound} = 0.45\text{kg}$*) of* 200 *members of a fitness center.*



24	9
23	
22	1
21	7
20	2, 2, 5, 5, 6, 9, 9, 9
19	0, 0, 0, 0, 0, 1, 1, 2, 4, 4, 5, 8
18	0, 1, 1, 2, 2, 2, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 9, 9, 9
17	1, 1, 1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 9
16	0, 0, 1, 1, 1, 1, 2, 4, 5, 5, 6, 6, 8, 8, 8, 8
15	0, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9
14	0, 0, 0, 1, 2, 3, 4, 5, 6, 7, 7, 7, 8, 9, 9
13	0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9, 9
12	1, 1, 1, 2, 2, 2, 3, 4, 4, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 9, 9
11	0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 9, 9
10	0, 2, 3, 3, 3, 4, 4, 5, 7, 7, 8
9	0, 0, 9
8	6

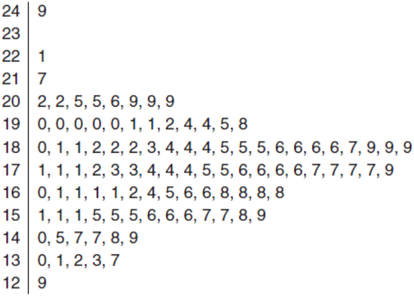
By turning the stem-and-leaf plot, we see that the data is bimodal.

This data is the superposition of the following two approximately normal data: the weights $\boldsymbol{x}$ *of* 97 *women*



16	0, 5
15	0, 1, 1, 1, 5
14	0, 0, 1, 2, 3, 4, 6, 7, 9
13	0, 0, 1, 1, 2, 2, 2, 2, 3, 4, 5, 5, 6, 6, 6, 6, 7, 8, 8, 8, 9, 9, 9
12	1, 1, 1, 2, 2, 2, 3, 4, 4, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 9
11	0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 9, 9
10	2, 3, 3, 3, 4, 4, 5, 7, 7, 8
9	0, 0, 9
8	6

and the weights $\boldsymbol{y}$ *of* 103 *men*



24	9
23	
22	1
21	7
20	2, 2, 5, 5, 6, 9, 9, 9
19	0, 0, 0, 0, 0, 1, 1, 2, 4, 4, 5, 8
18	0, 1, 1, 2, 2, 2, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 9, 9, 9
17	1, 1, 1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 9
16	0, 1, 1, 1, 1, 2, 4, 5, 6, 6, 8, 8, 8, 8
15	1, 1, 1, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9
14	0, 5, 7, 7, 8, 9
13	0, 1, 2, 3, 7
12	9

*We check that **x** and **y** are approximately normal.*

*For both **x** and **y**, the mean and the median are close:*

$$\overline{x} = 125.70,\ m_x = x^{\mathrm{ord}}_{49} = 126,\ \overline{y} = 174.69,\ m_y = y^{\mathrm{ord}}_{52} = 176.$$

*For the data **x**, we have* $s_x = 15.58$ *and*

$$\overline{x} - 2s_x = 94.54,\ \ 2.5\textit{th percentile } = x^{\mathrm{ord}}_{\lceil 97\cdot 0.025\rceil} = x^{\mathrm{ord}}_{3} = 90$$

$$\overline{x} - s_x = 110.12,\ \ 16\textit{th percentile } = x^{\mathrm{ord}}_{\lceil 97\cdot 0.16\rceil} = x^{\mathrm{ord}}_{16} = 110$$

$$\overline{x} + s_x = 141.28,\ \ 84\textit{th percentile } = x^{\mathrm{ord}}_{\lceil 97\cdot 0.84\rceil} = x^{\mathrm{ord}}_{82} = 140$$

$$\overline{x} + 2s_x = 156.86,\ \ 97.5\textit{th percentile } = x^{\mathrm{ord}}_{\lceil 97\cdot 0.975\rceil} = x^{\mathrm{ord}}_{95} = 155.$$

*All the components of **x** are between* $\overline{x} - 3s_x$ *and* $\overline{x} + 3s_x$*, i.e. between* 78.96 *and* 172.44*.*

For the data $\mathbf{y}$*, we have* $s_y = 21.23$ *and*

$$\overline{y} - 2s_y = 132.23, \ \ 2.5\textit{th percentile } = y^{\mathrm{ord}}_{\lceil 103 \cdot 0.025 \rceil} = y^{\mathrm{ord}}_{3} = 131$$

$$\overline{y} - s_y = 153.46, \ \ 16\textit{th percentile } = y^{\mathrm{ord}}_{\lceil 103 \cdot 0.16 \rceil} = y^{\mathrm{ord}}_{17} = 155$$

$$\overline{y} + s_y = 195.92, \ \ 84\textit{th percentile } = x^{\mathrm{ord}}_{\lceil 103 \cdot 0.84 \rceil} = y^{\mathrm{ord}}_{87} = 191$$

$$\overline{y} + 2s_y = 217.15, \ \ 97.5\textit{th percentile } = x^{\mathrm{ord}}_{\lceil 103 \cdot 0.975 \rceil} = y^{\mathrm{ord}}_{101} = 217.$$

All the components of $\mathbf{y}$ *except one, i.e.* $\frac{102}{103} = 99.03\%$ *of the components, are between* $\overline{y} - 3s_y$ *and* $\overline{x} + 3s_x$*, i.e. between* 111 *and* 238.38*.*

Some considerations about normal data

- It is a fact that:
  - if the data $\boldsymbol{x} \in \mathbb{R}^n$ represents some biological characteristic (for example heights, weights, blood pressure,...) of a sample taken from an homogeneous population of human beings, or other living beings, and its size n is large, then it is approximately normal, and it becomes normal, with the percentages $c(k)\%$, $k > 0$, exactly satisfied, as $n \to \infty$ and $h \to 0$, h being the length of the class intervals.

Here, homogeneous population means that all the individuals in the population are of the same type, not a mixture as in the previous example of the fitness center.

- A historical remark about normal data.

  The Belgian social scientist and statistician Adolphe Quetelet (1796-1874) (inventor of the "Body Mass Index")

Adolphe Quetelet

was the first to observe that data representing biological characteristics are normal.

He measured the chests of 5738 Scottish soldiers, plotted the resulting data in a histogram and concluded that it was normal.

In another study, he considered the heights of a huge sample of 100 000 conscripts in the French army and he uncovered a fraud.

In fact, he plotted the data in a histogram with class intervals of length 1 $\mathrm{inch} = 2.54\ \mathrm{cm}$ and he found that, with the exception of the class intervals around 62 $\mathrm{inch} = 151.9\ \mathrm{cm}$, the data appeared to be normal.

In particular, there were fewer components in the interval 62-63 inch and more components in the interval 61-62 inch than it was expected in a perfect normal data.

But, 62 inch was the minimum height required for soldiers in the French army.

Correlation coefficient

- Consider the paired data $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. We assume that both $\boldsymbol{x}$ and $\boldsymbol{y}$ have not all the components equal.

  We present a statistic, called the **correlation coefficient** that associates to $\boldsymbol{x}$ and $\boldsymbol{y}$ a measure of their "degree of correlation".

  We say that:

  - there is a **positive correlation** between $\boldsymbol{x}$ and $\boldsymbol{y}$ if smaller values (in the components x_i) of $\boldsymbol{x}$ go with smaller values (in the components y_i of the same index i) of $\boldsymbol{y}$ and larger values of $\boldsymbol{x}$ go with larger values of $\boldsymbol{y}$;

  - there is a **negative correlation** between $\boldsymbol{x}$ and $\boldsymbol{y}$ if smaller values of $\boldsymbol{x}$ go with larger values of $\boldsymbol{y}$ and larger values of $\boldsymbol{x}$ go with smaller values of $\boldsymbol{y}$.

*Example: consider the average daily number of cigarettes smoked (**x**) and the number of free radicals (**y**), in a suitable unit, found in the lungs of* $n = 10$ *smokers.*

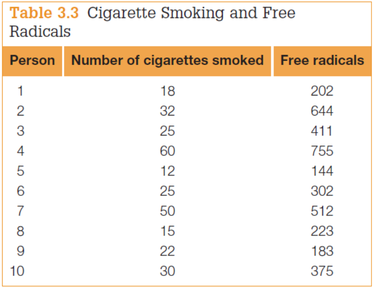
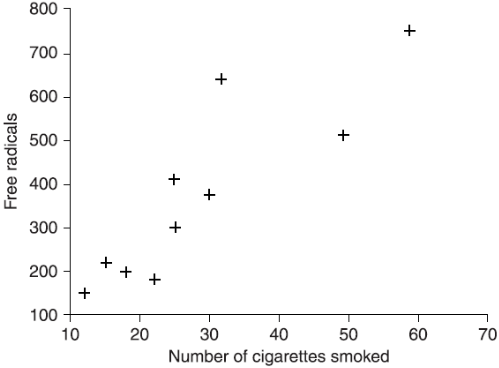


Table 3.3 Cigarette Smoking and Free Radicals

Person	Number of cigarettes smoked	Free radicals
1	18	202
2	32	644
3	25	411
4	60	755
5	12	144
6	25	302
7	50	512
8	15	223
9	22	183
10	30	375

A free radical is a molecule or atom presenting an unpaired electron, i.e. an electron that occupies an orbital of an atom singly. It is potentially harmful because it is highly reactive and has a strong tendency to combine with other molecules or atoms within the body.

Scatter diagram:



*There is a positive correlation between **x** and **y**.*

*Example: years of schooling (**x**) and the resting pulse rate in beats per minute (**y**) of $n = 10$ individuals.*

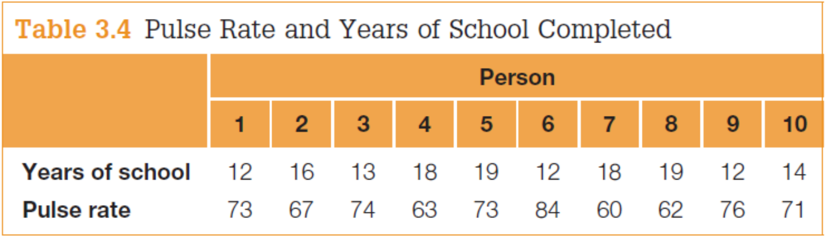
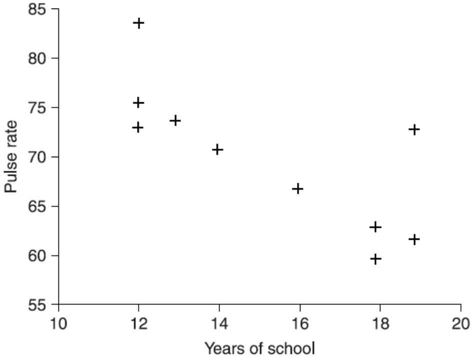


Table 3.4 Pulse Rate and Years of School Completed

	Person									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Years of school	12	16	13	18	19	12	18	19	12	14
Pulse rate	73	67	74	63	73	84	60	62	76	71

Scatter diagram:



*There is a negative correlation between **x** and **y**.*

- To obtain a statistic that can be used to measure the correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$, we observe that:
  - in case of a positive correlation, it is expected that, for many indeces $i \in \{1, 2, \ldots, n\}$, the deviations from the mean $x_i - \overline{x}$ and $y_i - \overline{y}$ have the same sign, i.e. $(x_i - \overline{x})(y_i - \overline{y})$ is positive;
  - in case of a negative correlation, it is expected that, for many indeces $i \in \{1, 2, \ldots, n\}$, $x_i - \overline{x}$ and $y_i - \overline{y}$ have different sign, i.e. $(x_i - \overline{x})(y_i - \overline{y})$ is negative.

Therefore, we can consider

$$\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

as a measure of the degree of correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Instead of the deviations $x_i - \overline{x}$ e $y_i - \overline{y}$, $i \in \{1, \dots, n\}$, it is better to consider the normalized deviations

$$\frac{x_i - \overline{x}}{\|\boldsymbol{x} - \overline{x}\|_2} = \frac{x_i - \overline{x}}{\sqrt{\sum\limits_{k=1}^{n} (x_k - \overline{x})^2}}$$

$$\frac{y_i - \overline{y}}{\|\boldsymbol{y} - \overline{y}\|_2} = \frac{y_i - \overline{y}}{\sqrt{\sum\limits_{k=1}^{n} (y_k - \overline{y})^2}}$$

where $\| \cdot \|_2$ is the euclidean norm in $\mathbb{R}^n$ and $\boldsymbol{x} - \overline{x}$ and $\boldsymbol{y} - \overline{y}$ are the $n-$tuples of components $x_k - \overline{x}$ and $y_k - \overline{y}$, $k \in \{1, \dots, n\}$, respectively.

Exercise. Prove that the normalized deviations are dimensionless, are included in the interval $[-1, 1]$ and the sum of their squares is 1.

Hence, we consider as a measure of the degree of correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$, the quantity

$$\begin{aligned} r_{x,y} \quad &:= \quad \sum_{i=1}^{n} \frac{x_i - \overline{x}}{\sqrt{\sum\limits_{k=1}^{n} (x_k - \overline{x})^2}} \cdot \frac{y_i - \overline{y}}{\sqrt{\sum\limits_{k=1}^{n} (y_k - \overline{y})^2}} \\ &= \quad \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum\limits_{k=1}^{n} (x_k - \overline{x})^2} \cdot \sqrt{\sum\limits_{k=1}^{n} (y_k - \overline{y})^2}} \\ &= \quad \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{n-1}s_x \cdot \sqrt{n-1}s_y} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{(n-1)\,s_x s_y}, \end{aligned}$$

called the **correlation coefficient** between $\boldsymbol{x}$ and $\boldsymbol{y}$.

- Previously, we have given an informal and qualitative definition of positive and negative correlations between the data $\boldsymbol{x}$ and $\boldsymbol{y}$. Now, we are ready to give a formal and quantitative definition of positive and negative correlations.

  We say that:

  - there is a **positive correlation** between $\boldsymbol{x}$ and $\boldsymbol{y}$ if $r_{x,y} > 0$;
  - there is a **negative correlation** between $\boldsymbol{x}$ and $\boldsymbol{y}$ if $r_{x,y} < 0$.

- The correlation coefficient can be written as

$$r_{x,y} = \frac{\langle \boldsymbol{x} - \overline{x}, \boldsymbol{y} - \overline{y} \rangle}{\|\boldsymbol{x} - \overline{x}\|_2 \cdot \|\boldsymbol{y} - \overline{y}\|_2}, \tag{1}$$

where $\langle \cdot \, , \, \cdot \rangle$ is the usual scalar product in $\mathbb{R}^n$.

Then, by using the Cauchy-Schwarz inequality we obtain

$$|r_{x,y}| \leq 1$$

and

$$r_{x,y} = 1 \;\Leftrightarrow\; \boldsymbol{y} - \overline{y} = a\,(\boldsymbol{x} - \overline{x}) \;\text{ for some } a > 0$$

and

$$r_{x,y} = -1 \;\Leftrightarrow\; \boldsymbol{y} - \overline{y} = a\,(\boldsymbol{x} - \overline{x}) \;\text{ for some } a < 0.$$

To better understand this, observe that the right-hand side of (1) is $\cos\theta$, where $\theta \in [0, \pi]$ is the angle between the vectors $\boldsymbol{x} - \overline{x}$ and $\boldsymbol{y} - \overline{y}$: so $|r_{x,y}| = |\cos\theta| \leq 1$, $r_{x,y} = \cos\theta = 1 \Leftrightarrow \theta = 0$ (i.e. $\boldsymbol{y} - \overline{y}$ is a positive multiple of $\boldsymbol{x} - \overline{x}$) and $r_{x,y} = \cos\theta = -1 \Leftrightarrow \theta = \pi$ (i.e. $\boldsymbol{y} - \overline{y}$ is a negative multiple of $\boldsymbol{x} - \overline{x}$).

We have $|r_{x,y}| = 1$ if and only if the pairs (x_i, y_i), $i \in \{1, \dots n\}$, lie on a straight line

$$y = mx + q,$$

with nonzero slope m. In fact, we have $|r_{x,y}| = 1$ if and only if

$$\boldsymbol{y} - \overline{y} = a\,(\boldsymbol{x} - \overline{x})$$

for some $a \neq 0$ and this means

$$y_i - \overline{y} = a\,(x_i - \overline{x})\,,\ i \in \{1, \dots, n\}\,,$$

i.e.

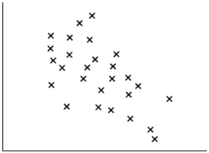
$$y_i = \underbrace{a}_{=m} x_i + \underbrace{\overline{y} - a\overline{x}}_{=q},\ i \in \{1, \dots, n\}\,.$$

The case where the pairs (x_i, y_i), $i \in \{1, \dots n\}$, lie on a straight line $y = mx + q$ is a situation of **perfect linear correlation** between $\boldsymbol{x}$ and $\boldsymbol{y}$. We have:

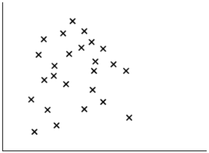
- **positive perfect linear correlation** when $m = a$ is positive and so $r_{x,y} = 1$;
- **negative perfect linear correlation** when $m = a$ is negative and so $r_{x,y} = -1$.

The absolute value of $r_{x,y}$ is a measure of the strength of correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$.

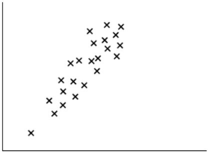
Scatter diagrams for paired data with various values of $r_{x,y}$:



$r = -0.50$



$r = 0$



$r = 0.90$

We say that $\boldsymbol{x}$ and $\boldsymbol{y}$ are **strongly correlated** if $|r_{x,y}| \geq 0.7$, **weakly correlated** if $|r_{x,y}| \leq 0.3$ and **uncorrelated** if $r_{x,y}$ is zero. When $|r_{x,y}| = 1$, they are perfectly linearly correlated.

- An important property of the correlation coefficients is

$$r_{p,q} = r_{x,y},$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are given by

$$\begin{aligned} \boldsymbol{p} &= \alpha\boldsymbol{x} + \beta \\ \boldsymbol{q} &= \gamma\boldsymbol{y} + \delta, \end{aligned}$$

with $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ and α, γ both positive or both negative.

In fact

$$\begin{aligned} r_{p,q} &= \frac{\sum_{i=1}^{n}(\alpha x_i + \beta - \overline{\alpha\boldsymbol{x} + \beta}) \cdot (\gamma y_i + \delta - \overline{\gamma\boldsymbol{y} + \delta})}{(n-1) \cdot s_{\alpha\boldsymbol{x}+\beta} \cdot s_{\gamma\boldsymbol{y}+\delta}} \\ &= \frac{\sum_{i=1}^{n}(\alpha x_i + \beta - \alpha\overline{x} - \beta) \cdot (\gamma y_i + \delta - \gamma\overline{y} - \delta)}{(n-1) \cdot |\alpha| s_x \cdot |\gamma| s_y} \\ &= \frac{\sum_{i=1}^{n} \alpha(x_i - \overline{x}) \cdot \gamma(y_i - \overline{y})}{(n-1) \cdot |\alpha| s_x \cdot |\gamma| s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1) s_x s_y} = r_{x,y}. \end{aligned}$$

This means that the correlation coefficient is invariant under a shift of the origin, or a change of scales in the axes, or an inversion of both axes, in the scatter diagram.

Exercise. What happens to the previous property if α and γ have different signs?

Exercise. Suppose that ***x*** and ***y*** are data of temperatures measured in F°. Does the correlation coefficient between ***x*** and ***y*** change if the temperature are measured in C° rather than F°?

We also have the property

$$r_{x,y} = r_{y,x}.$$

In fact

$$r_{x,y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{(n-1)\, s_x s_y} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{(n-1)\, s_y s_x} = r_{y,x}.$$

This means that the correlation coefficient is also invariant under an exchange of the axes in the scatter diagram.

The correlation coefficient can be expressed also in the following form

$$r_{x,y} = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{\sqrt{\left(\sum_{k=1}^{n} x_k^2 - n\overline{x}^2\right)\left(\sum_{k=1}^{n} y_k^2 - n\overline{y}^2\right)}}.$$

Exercise. Prove this new formula.

Example: consider the paired data

Year	*x (whole milk)*	*y (low-fat milk)*
1980	17.1	10.6
1984	14.7	11.5
1988	12.8	13.2

giving the US pro-capita consumption (in gallons: $1\ \mathrm{gallon} = 3.785\ \mathrm{ltr}$*) of whole milk and low-fat milk during the eighties.*

It appears that US people began in the eighties to consume low-fat milk instead of whole milk. So a negative correlation coefficient between these data is expected.

Since $n = 3$ is small, we can compute this coefficient by hands.

Instead of $\boldsymbol{x}$ *and* $\boldsymbol{y}$*, we use*

$$\boldsymbol{p} = \boldsymbol{x} - 12.8 = (4.3, 1.9, 0)\,,\ \boldsymbol{q} = \boldsymbol{y} - 10.6 = (0, 0.9, 2.6)\,,$$

for which $r_{p,q} = r_{x,y}$, and compute $r_{p,q}$ by the new formula.

We have

$$\overline{p} = \frac{4.3 + 1.9}{3} = \frac{6.2}{3} = 2.0667,\ \overline{q} = \frac{0.9 + 2.6}{3} = \frac{3.5}{3} = 1.1667,$$

$$\sum_{k=1}^{3} p_k q_k = 1.9 \cdot 0.9 = 1.71,\ \sum_{k=1}^{3} p_k^2 = 4.3^2 + 1.9^2 = 22.10,$$

$$\sum_{k=1}^{3} q_k^2 = 0.9^2 + 2.6^2 = 7.57$$

$$r_{x,y} = r_{p,q} = \frac{\sum\limits_{k=1}^{3} p_k q_k - 3 \cdot \overline{p} \cdot \overline{q}}{\sqrt{\left(\sum\limits_{k=1}^{3} p_k^2 - 3\overline{p}^2\right)\left(\sum\limits_{k=1}^{3} q_k^2 - 3\overline{q}^2\right)}}$$

$$= \frac{1.71 - 3 \cdot 2.0667 \cdot 1.1667}{\sqrt{(22.10 - 3 \cdot 2.0667^2)\,(7.57 - 3 \cdot 1.1667^2)}} = -0.97.$$

- In MATLAB, the correlation coefficient between paired data in the vectors x and y is computed by

$$\mathrm{corrcoef(x, y)}.$$

$\mathrm{corrcoef(x,y)}$ is a 2×2 symmetric matrix and the correlation coefficient is the off-diagonal element. If x and y are columns vector, the correlation coefficient can be also computed by

$$\mathrm{(x - mean(x))' * (y - mean(y))/((n - 1) * std(x) * std(y))}$$

Exercise. By using MATLAB compute the correlation coefficients for the example of cigarettes and free radicals, for the example of years of schooling and beats of pulse and for the example of IQ scores and salaries.

Exercise. By using MATLAB, find the correlation coefficient between $\boldsymbol{x} = (0, 0.1, 0.2, \dots, 1)$ and $\boldsymbol{y}$ with components

1) $y_i = \sqrt{x_i}$, $i \in \{1, 2, \dots, 11\}$;

2) $y_i = x_i^2$, $i \in \{1, 2, \dots, 11\}$;

3) $y_i = x_i^3$, $i \in \{1, 2, \dots, 11\}$;

4) $y_i = \frac{1}{1+x_i}$, $i \in \{1, 2, \dots, 11\}$;

5) $y_i = \sin(2\pi x_i)$, $i \in \{1, 2, \dots, 11\}$;

6) $y_i = x_i + \frac{1}{10}\sin(2\pi x_i)$, $i \in \{1, 2, \dots, 11\}$;

7) y_i, $i \in \{1, 2, \dots, 11\}$, are independently and uniformly distributed on $[0, 1]$, i.e. $\mathrm{y} = \mathrm{rand}(11, 1)$ in MATLAB.

- Exercise. Collect on Wikipedia for the twenty teams of Seria A in season 2016-2017, the number of goals scored (data $\boldsymbol{x}$), the number of goals conceded (data $\boldsymbol{y}$) and the number of points obtained (data $\boldsymbol{z}$). Study the correlation between $\boldsymbol{x}$ and $\boldsymbol{z}$, $\boldsymbol{y}$ and $\boldsymbol{z}$, and $\boldsymbol{x}$ and $\boldsymbol{y}$.

The regression line

- When $|r_{x,y}| = 1$, we have a perfect linear relation between $\boldsymbol{x}$ and $\boldsymbol{y}$: there is straight line passing through all the points (x_i, y_i) , $i \in \{1, \ldots, n\}$.

  When $|r_{x,y}| < 1$, there is not a straight line passing through all the points (x_i, y_i) , $i \in \{1, \ldots, n\}$, but the trend is given by the **regression line** of the paired data $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e. the nonvertical straight line

  $$y = mx + q$$

  that minimizes the **residual sum of the squares**

  $$rss = \sum_{i=1}^{n} (y_i - mx_i - q)^2$$

  among all the possible nonvertical straight lines in the plane, i.e. among all the possible pairs $(m, q) \in \mathbb{R}^2$.

In the next theorem we show that the slope m of the regression line is given by

$$m = \frac{s_y}{s_x} r_{x,y}.$$

So the correlation coefficient $r_{x,y}$ and the slope m of the regression line have the same sign.

The sign of $r_{x,y}$ and m gives the “direction” up/down of the correlation:

- positive sign: the regression line heads upward and smaller values of $\boldsymbol{x}$ tend to go with smaller values of $\boldsymbol{y}$ and larger values of $\boldsymbol{x}$ values tend to go with larger values of $\boldsymbol{y}$;
- negative sign: the regression line heads downward and smaller values of $\boldsymbol{x}$ tend to go with larger values of $\boldsymbol{y}$ and larger values of $\boldsymbol{x}$ tend to go with smaller values of $\boldsymbol{y}$.

- Now, we show how to determine the regression line.

Theorem

The regression line

$$y = mx + q$$

of the paired data $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, is given by

$$m = \frac{s_y}{s_x} r_{x,y} \quad \text{and} \quad q = \overline{y} - m\overline{x}.$$

Moreover, we have

$$rss = (n-1)s_x^2(1 - r_{x,y}^2).$$

for the regression line.

Proof.

We use the deviations

$$\Delta_i = y_i - \overline{y} \text{ and } \delta_i = x_i - \overline{x},\ i \in \{1, \ldots, n\}$$

and the vectors of the deviations

$$\boldsymbol{\Delta} = \boldsymbol{y} - \overline{y} = (\Delta_1, \ldots, \Delta_n),\ \boldsymbol{\delta} = \boldsymbol{x} - \overline{x} = (\delta_1, \ldots, \delta_n) \in \mathbb{R}^n.$$

Observe that *rss* can be expressed in terms of such deviations:

$$\begin{aligned} rss &= \sum_{i=1}^{n} (y_i - m x_i - q)^2 = \sum_{i=1}^{n} (\overline{y} + \Delta_i - m(\overline{x} + \delta_i) - q)^2 \\ &= \sum_{i=1}^{n} (\Delta_i - m\delta_i - (q - (\overline{y} - m\overline{x})))^2 = \sum_{i=1}^{n} (\Delta_i - m\delta_i - p)^2 \end{aligned}$$

where

$$p := q - (\overline{y} - m\overline{x}).$$

Proof.

Let

$$A = \begin{bmatrix} \delta_1 & 1 \\ \vdots & \vdots \\ \delta_n & 1 \end{bmatrix} = [\boldsymbol{\delta}\ \mathbf{1}] \in \mathbb{R}^{n\times 2},$$

where $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$, and let $\boldsymbol{z} = (m, p) \in \mathbb{R}^2$. We have that

$$\boldsymbol{\Delta} - A\boldsymbol{z} = \begin{bmatrix} \Delta_1 \\ \vdots \\ \Delta_n \end{bmatrix} - \begin{bmatrix} \delta_1 & 1 \\ \vdots & \vdots \\ \delta_n & 1 \end{bmatrix} \begin{bmatrix} m \\ p \end{bmatrix}$$

has components

$$\Delta_i - m\delta_i - p,\ i \in \{1, \dots, n\}\,.$$

So the regression line is the line $y = m_0 x + q_0$, with $q_0 = p_0 + (\overline{y} - m\overline{x})$ such that $\boldsymbol{z}_0 = (m_0, p_0)$ minimizes among all $\boldsymbol{z} = (m, p) \in \mathbb{R}^2$ the quantity

$$rss = \sum_{i=1}^{n} (\Delta_i - m\delta_i - p)^2 = \|\boldsymbol{\Delta} - A\boldsymbol{z}\|_2^2\,.$$

Proof.

Thus, the determination of the regression line is a particular instance of the **least squares problem**: given a matrix $B \in \mathbb{R}^{k\times\ell}$ and $\boldsymbol{c} \in \mathbb{R}^k$, find $\boldsymbol{u}_0 \in \mathbb{R}^\ell$ that minimizes among all $\boldsymbol{u} \in \mathbb{R}^\ell$ the quantity

$$\left\|\boldsymbol{c} - B\boldsymbol{u}\right\|_2^2 .$$

In our case $\boldsymbol{c} = \boldsymbol{\Delta} \in \mathbb{R}^n$, $B = A \in \mathbb{R}^{n\times 2}$ and $\boldsymbol{u} = \boldsymbol{z} \in \mathbb{R}^2$.

Such minimums $\boldsymbol{u}_0$ are the solutions of the **system of the normal equations**

$$B^T B\boldsymbol{u} = B^T\boldsymbol{c}$$

and for the minimum value $\left\|\boldsymbol{c} - B\boldsymbol{u}_0\right\|_2^2$ we have

$$\left\|\boldsymbol{c} - B\boldsymbol{u}_0\right\|_2^2 = \left\|\boldsymbol{c}\right\|_2^2 - \left\|B\boldsymbol{u}_0\right\|_2^2 .$$

Now, we prove this.

Proof.

We look for elements

$$\boldsymbol{b}_0 \in \operatorname{ran}(B) = \left\{ B\boldsymbol{u} : \boldsymbol{u} \in \mathbb{R}^\ell \right\}$$

minimizing $\|\boldsymbol{c} - \boldsymbol{b}\|_2^2$ among all $\boldsymbol{b} \in \operatorname{ran}(B)$.
Once we have found such minimizing elements $\boldsymbol{b}_0$, our minimums $\boldsymbol{u}_0$ are the vectors $\boldsymbol{u}_0 \in \mathbb{R}^\ell$ such that $B\boldsymbol{u}_0 = \boldsymbol{b}_0$.

Consider an element $\boldsymbol{b}_0 \in \operatorname{ran}(B)$ such that

$$\langle \boldsymbol{c} - \boldsymbol{b}_0, \boldsymbol{v} \rangle = 0 \text{ for any } \boldsymbol{v} \in \operatorname{ran}(B).$$

We have, for $\boldsymbol{b} \in \operatorname{ran}(B)$,

$$\begin{aligned} \|\boldsymbol{c} - \boldsymbol{b}\|_2^2 &= \|\boldsymbol{c} - \boldsymbol{b}_0 + \boldsymbol{b}_0 - \boldsymbol{b}\|_2^2 \\ &= \|\boldsymbol{c} - \boldsymbol{b}_0\|_2^2 + 2 \underbrace{\langle \boldsymbol{c} - \boldsymbol{b}_0, \boldsymbol{b}_0 - \boldsymbol{b} \rangle}_{=0 \text{ since } \boldsymbol{b}_0 - \boldsymbol{b} \in \operatorname{ran}(B)} + \|\boldsymbol{b}_0 - \boldsymbol{b}\|_2^2 \\ &= \|\boldsymbol{c} - \boldsymbol{b}_0\|_2^2 + \|\boldsymbol{b}_0 - \boldsymbol{b}\|_2^2 . \end{aligned}$$

Proof.

Since, for all $\boldsymbol{b} \in \operatorname{ran}(B)$, we have

$$\|\boldsymbol{c}-\boldsymbol{b}\|_2^2 = \|\boldsymbol{c}-\boldsymbol{b}_0\|_2^2 + \|\boldsymbol{b}_0-\boldsymbol{b}\|_2^2 \geq \|\boldsymbol{c}-\boldsymbol{b}_0\|_2^2 \,,$$

$\boldsymbol{b}_0$ is a minimum of $\|\boldsymbol{c}-\boldsymbol{b}\|^2$ among all $\boldsymbol{b} \in \operatorname{ran}(B)$.
It is the sole minimum. In fact, if $\boldsymbol{b} \in \operatorname{ran}(B)$ is a minimum, then

$$\|\boldsymbol{c}-\boldsymbol{b}_0\|_2^2 = \|\boldsymbol{c}-\boldsymbol{b}\|_2^2 = \|\boldsymbol{c}-\boldsymbol{b}_0\|_2^2 + \|\boldsymbol{b}_0-\boldsymbol{b}\|_2^2$$

and then $\|\boldsymbol{b}_0-\boldsymbol{b}\|_2^2 = 0$ and so $\boldsymbol{b}_0 = \boldsymbol{b}$.

The minimum value $\|\boldsymbol{c}-\boldsymbol{b}_0\|_2^2$ satisfies

$$\|\boldsymbol{c}-\boldsymbol{b}_0\|_2^2 = \|\boldsymbol{c}\|_2^2 - \|\boldsymbol{b}_0\|_2^2 \,.$$

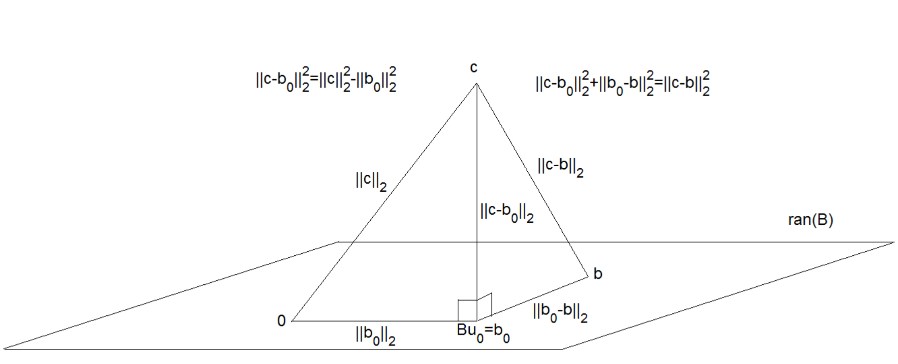
This is obtained by taking $\boldsymbol{b} = \mathbf{0}$ in

$$\|\boldsymbol{c}-\boldsymbol{b}\|_2^2 = \|\boldsymbol{c}-\boldsymbol{b}_0\|_2^2 + \|\boldsymbol{b}_0-\boldsymbol{b}\|_2^2 \,.$$

which holds for all $\boldsymbol{b} \in \operatorname{ran}(B)$.

Proof.

A picture when $k = 3$ (the space $\mathbb{R}^k$ is the three-dimensional space) and the subspace $\mathrm{ran}(B)$ has dimension 2, i.e. it is a plane passing through the origin.


$\|c-b_0\|_2^2=\|c\|_2^2-\|b_0\|_2^2$
c
$\|c-b_0\|_2^2+\|b_0-b\|_2^2=\|c-b\|_2^2$
$\|c-b\|_2$
$\|c\|_2$
$\|c-b_0\|_2$
ran(B)
b
0
$\|b_0-b\|_2$
$\|b_0\|_2$
$Bu_0=b_0$

Proof.

Now, the condition

$$\langle \boldsymbol{c} - \boldsymbol{b}_0, \boldsymbol{v} \rangle = 0 \text{ for any } \boldsymbol{v} \in \operatorname{ran}(B) \tag{2}$$

is equivalent to

$$\langle \boldsymbol{c} - \boldsymbol{b}_0, \boldsymbol{b}^{(i)} \rangle = 0 \text{ for any } i \in \{1, \ldots, \ell\}, \tag{3}$$

where $\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(\ell)}$ are the columns of B.

In fact: the columns of B are particular elements of $\operatorname{ran}(B)$ and so (2) $\Rightarrow$ (3); any $\boldsymbol{v} \in \operatorname{ran}(B)$ is a linear combination

$$\boldsymbol{v} = B\boldsymbol{u} = \sum_{i=1}^{\ell} u_i \boldsymbol{b}^{(i)},$$

for some $\boldsymbol{u} = (u_1, \ldots, u_\ell)$, of the columns $\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(\ell)}$ and then

$$\langle \boldsymbol{c} - \boldsymbol{b}_0, \boldsymbol{v} \rangle = \langle \boldsymbol{c} - \boldsymbol{b}_0, \sum_{i=1}^{\ell} u_i \boldsymbol{b}^{(i)} \rangle = \sum_{i=1}^{\ell} u_i \langle \boldsymbol{c} - \boldsymbol{b}_0, \boldsymbol{b}^{(i)} \rangle$$

and so (3) $\Rightarrow$ (2).

Proof.

Finally, the condition

$$\langle \boldsymbol{c} - \boldsymbol{b}_0, \boldsymbol{b}^{(i)} \rangle = 0 \ \text{ for any } i \in \{1, \dots, \ell\},$$

can be expressed as

$$B^T (\boldsymbol{c} - \boldsymbol{b}_0) = \mathbf{0} \tag{4}$$

since the components of $B^T (\boldsymbol{c} - \boldsymbol{b}_0)$ are

$$\langle \boldsymbol{c} - \boldsymbol{b}_0, \boldsymbol{b}^{(i)} \rangle, \ i \in \{1, \dots, \ell\}.$$

We conclude that a solution of the least squares problem is an element $\boldsymbol{u}_0 \in \mathbb{R}^\ell$ such that $B\boldsymbol{u}_0 = \boldsymbol{b}_0$, where $\boldsymbol{b}_0$ satisfies (4) . So, we have to find the solutions $\boldsymbol{u}_0$ of the system

$$B^T (\boldsymbol{c} - B\boldsymbol{u}) = \mathbf{0}, \quad \text{i.e. } B^T B\boldsymbol{u} = B^T \boldsymbol{c}.$$

Moreover, we have

$$\|\boldsymbol{c} - B\boldsymbol{u}_0\|_2^2 = \|\boldsymbol{c}\|_2^2 - \|B\boldsymbol{u}_0\|_2^2.$$

We have proved the result about the least squares problem solution.

Proof.

Now, we apply this result to our case: we have the system of the normal equations

$$A^T A\mathbf{z} = A^T \mathbf{\Delta}$$

where

$$\begin{aligned} A^T A &= \begin{bmatrix} \boldsymbol{\delta}^T \\ \mathbf{1}^T \end{bmatrix} [\boldsymbol{\delta}\ \mathbf{1}] = \begin{bmatrix} \boldsymbol{\delta}^T\boldsymbol{\delta} & \boldsymbol{\delta}^T\mathbf{1} \\ \mathbf{1}^T\boldsymbol{\delta} & \mathbf{1}^T\mathbf{1} \end{bmatrix} = \begin{bmatrix} \sum\limits_{i=1}^{n} \delta_i^2 & \sum\limits_{i=1}^{n} \delta_i \\ \sum\limits_{i=1}^{n} \delta_i & n \end{bmatrix} \\ &= \begin{bmatrix} (n-1)\, s_x^2 & 0 \\ 0 & n \end{bmatrix} \end{aligned}$$

and

$$A^T\, \mathbf{\Delta} = \begin{bmatrix} \boldsymbol{\delta}^T \\ \mathbf{1}^T \end{bmatrix} \mathbf{\Delta} = \begin{bmatrix} \sum\limits_{i=1}^{n} \delta_i \Delta_i \\ \sum\limits_{i=1}^{n} \Delta_i \end{bmatrix} = \begin{bmatrix} (n-1)\, s_x s_y r_{x,y} \\ 0 \end{bmatrix}.$$

Proof.

As solution $\boldsymbol{z}_0 = (m_0, p_0)$ of this system, we obtain

$$\begin{aligned} m_0 &= \frac{(n-1)\, s_x s_y r_{x,y}}{(n-1)\, s_x^2} = \frac{s_y}{s_x} r_{x,y} \\ p_0 &= q_0 - (\overline{y} - m_0 \overline{x}) = 0. \end{aligned}$$

Observe that this is the unique solution of the system and so there is a unique line minimizing *rss*.

Moreover, since

$$rss = \|\Delta - A\boldsymbol{z}_0\|_2^2 = \|\boldsymbol{\Delta}\|_2^2 - \|A\boldsymbol{z}_0\|_2^2 \text{, with } A\boldsymbol{z}_0 = [\delta \ \mathbf{1}] \begin{bmatrix} m_0 \\ p_0 \end{bmatrix} = m_0 \boldsymbol{\delta},$$

we have

$$\begin{aligned} rss &= \|\boldsymbol{\Delta}\|_2^2 - \|m_0 \boldsymbol{\delta}\|_2^2 = \|\boldsymbol{\Delta}\|_2^2 - m_0^2 \|\boldsymbol{\delta}\|_2^2 = \sum_{i=1}^{n} \Delta_i^2 - m_0^2 \sum_{i=1}^{n} \delta_i^2 \\ &= (n-1)\, s_y^2 - \frac{s_y^2}{s_x^2} r_{x,y}^2 (n-1)\, s_x^2 = (n-1)\, s_y^2 \left(1 - r_{x,y}^2\right). \end{aligned}$$

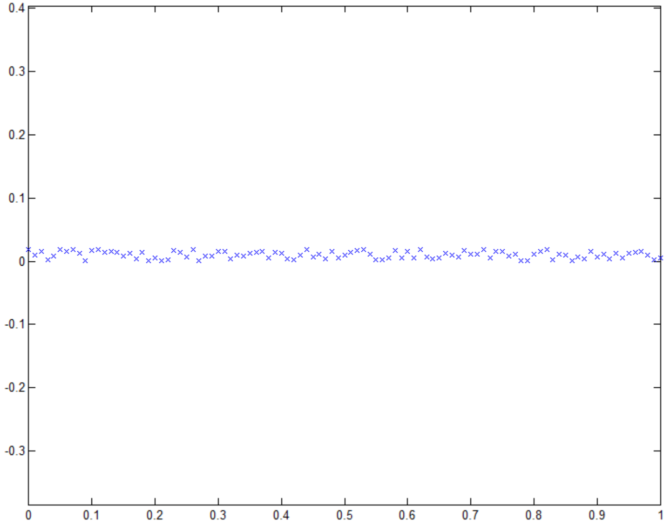
- The regression line is the line $y = mx + q$ minimizing the euclidean norm
$$\|\boldsymbol{y} - m\boldsymbol{x} - q\|_2 = \sqrt{rss}$$
of
$$\boldsymbol{y} - m\boldsymbol{x} - q = (y_1 - mx_1 - q, \ldots, y_n - mx_n - q)$$
and such minimal norm is
$$\begin{aligned} \sqrt{rss} &= \sqrt{(n-1)\, s_y^2 \left(1 - r_{x,y}^2\right)} = \sqrt{(n-1)\, s_y^2} \cdot \sqrt{1 - r_{x,y}^2} \\ &= \|\boldsymbol{y} - \overline{y}\|_2 \cdot \sqrt{1 - r_{x,y}^2} \end{aligned}$$

So $\sqrt{rss}$ is zero if and only if $|r_{x,y}| = 1$, i.e. there is a perfect linear correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Question: is $\sqrt{rss}$ a good measure of the error with respect to a perfect linear correlation?

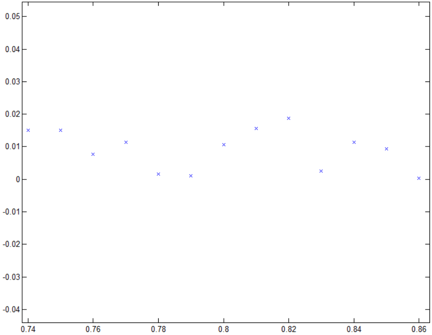
Observe that $\sqrt{rss}$ is small when $\|\boldsymbol{y} - \overline{y}\|_2$ is small, i.e. the components of $\boldsymbol{y}$ are close to the mean, or when $\sqrt{1 - r_{x,y}^2}$ is small, i.e. $r_{x,y}$ is close to one, i.e. we are close to have a perfect linear correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Consider the scatter diagram



In this case $\sqrt{rss}$ is small because $\|\boldsymbol{y} - \overline{y}\|_2$ is small.

But, a zoom on the stripe shows that the points are far from a perfect linear correlation



So, in order to describe the error with respect to a perfect linear correlation, it is better to use not $\sqrt{rss}$ but the dimensionless quantity

$$\frac{\sqrt{rss}}{\|\boldsymbol{y} - \overline{y}\|_2} = \sqrt{1 - r_{x,y}^2}.$$

Exercise. Consider the following two measures of the error with respect to a perfect linear correlation:

$$m_1 = \frac{\sqrt{rss}}{\|\boldsymbol{y} - \overline{y}\|_2} = \sqrt{1 - r_{x,y}^2} \text{ and } m_2 = 1 - |r_{x,y}|.$$

Express m_2 in terms of m_1 and m_1 in terms of m_2. Prove that m_2 is an increasing function of m_1 and m_1 is an increasing function of m_2. Find numbers a and b such that

$$|r_{x,y}| \geq 0.7 \text{ (strong correlation) } \Leftrightarrow m_1 \leq a$$

and

$$|r_{x,y}| \leq 0.3 \text{ (weak correlation) } \Leftrightarrow m_1 \geq b.$$

Finally, show that

$$m_2 \approx \frac{1}{2} m_1^2 \text{ for } m_1 \text{ small.}$$

So, m_2 is much smaller than m_1 when m_1 is small.

Exercise. Above, we have described a situation where $\sqrt{1 - r_{x,y}^2}$ is not small, and so we are not close to a perfect linear correlation, but $\sqrt{rss}$ is small because $\|\boldsymbol{y} - \overline{y}\|_2$ is small.

Now, describe a situation where $\sqrt{1 - r_{x,y}^2}$ is small, and so we are close to a perfect linear correlation, but $\sqrt{rss}$ is not small because $\|\boldsymbol{y} - \overline{y}\|_2$ is large.

- Instead of the regression line, we could consider the line $y = mx + q$ minimizing

$$f_1(m,q) = \sum_{i=1}^{n} |y_i - mx_i - q| = \|\boldsymbol{y} - m\boldsymbol{x} - q\|_1$$

or

$$f_\infty(m,q) = \max_{i\in\{1,\ldots,n\}} |y_i - mx_i - q| = \|\boldsymbol{y} - m\boldsymbol{x} - q\|_\infty.$$

But, these problems of minimization are mathematically more difficult than the previous one, where we minimize

$$\sqrt{\sum_{i=1}^{n}(y_i - mx_i - q)^2} = \|\boldsymbol{y} - m\boldsymbol{x} - q\|_2$$

i.e.

$$f_2(m,q) = rss = \sum_{i=1}^{n}(y_i - mx_i - q)^2.$$

Exercise. Try to explain why it is easier to find the minimum of f_2 rather than the minumum of f_1 or f_∞.

Exercise. Find the minimum of f_2 by using differential calculus.

- Exercise. Does the point $(\overline{x}, \overline{y})$ belong to the regression line?

  Exercise. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and let $\boldsymbol{p} = \alpha \boldsymbol{x} + \beta$ and $\boldsymbol{q} = \gamma \boldsymbol{y} + \delta$, where $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ and α, γ are both nonzero. Find the relation between the slope of the regression line for the paired data $\boldsymbol{x}, \boldsymbol{y}$ and the slope of the regression line for the paired data $\boldsymbol{p}, \boldsymbol{q}$.

  Exercise. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Find the relation between the slope of the regression line for the paired data $\boldsymbol{x}, \boldsymbol{y}$ and the slope of the regression line for the paired data $\boldsymbol{y}, \boldsymbol{x}$.

  Exercise. Explain why the regression line cannot lie in the scatter diagram above or below all points (x_i, y_i), $i \in \{1, \ldots, n\}$.

- The regression line is computed in MATLAB by the function $\mathrm{regline}$: for paired data in the vectors x and y

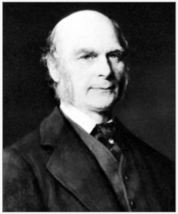
  $$[\mathrm{m}, \mathrm{q}] = \mathrm{regline}(\mathrm{x}, \mathrm{y})$$

  gives *m* and *q* of the regression line $y = mx + q$ and, in addition, plots the scatter diagram with the regression line imposed on it.

  Exercise. Find the regression line and plot the scatter diagram with the regression line for the example of cigarettes and free radicals, for the example of years of schooling and beats of pulse and for the example of IQ scores and salaries.

- A historical remark about the regression line and the correlation coefficient.

  The concepts of correlation coefficient and regression line were introduced by the English explorer and scientist Sir Francis Galton (1822-1911)



Francis Galton

who was trying to study the laws of inheritance parent-offspring from a quantitative point of view.

He wanted to quantify how a characteristic (e.g. the height) of an offspring is related to that of the parent: for $i \in \{1, \ldots, n\}$,

- x_i is the characteristic of the $i-$th parent;
- y_i is the characteristic of the offspring of the $i-$th parent.

Observe that if $s_x = s_y$ (and this fact can be observed in the example of the heights and in many other cases), then the regression line

$$y = mx + \overline{y} - m\overline{x} \ \text{ or } \ y - \overline{y} = m(x - \overline{x})$$

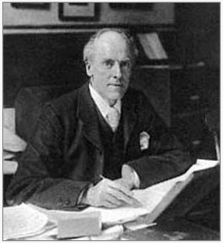
has $|m| = |r_{x,y}| \leq 1$ and so at a point (x, y) on the regression line we have

$$|y - \overline{y}| = |m(x - \overline{x})| = |m||x - \overline{x}| \leq |x - \overline{x}|$$

Therefore, the trend is such that there is a "regression towards the mean" of the offspring with respect to the parent. This was the reason for which Galton introduced the name "regression line".

Galton later realized that the correlation coefficient used for studying the laws of inheritance was also a method of quantifying the interrelation between any paired data.

However, his form of the correlation coefficient was different from the form that is presently in use. The present form is due to the English mathematician Karl Pearson (1857-1936)



Karl Pearson

and sometimes it is more properly called **Pearson’ s product-moment correlation coefficient**.

Causation and association

- Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be paired data. Suppose that there is a positive (negative) correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e. smaller values of $\boldsymbol{x}$ go with smaller (larger) values of $\boldsymbol{y}$ and larger values of $\boldsymbol{x}$ go with larger (smaller) values of $\boldsymbol{y}$.

  Can we say that the smaller values of $\boldsymbol{x}$ are the cause of the smaller (larger) values of $\boldsymbol{y}$ and the larger values of $\boldsymbol{x}$ are the cause of the larger (smaller) values of $\boldsymbol{y}$?

We saw that there is a strong positive correlation ($r_{x,y} = 0.876$) between the number of cigarettes smoked and the number of free radicals in the lungs.

In this case, one can guess that the large number of smoked cigarettes is the cause of the large number of free radicals. The explanation of this fact can come by biochemistry.

Exercise. Does it make sense to say that the large number of free radicals is the cause of the large number of smoked cigarettes?

We also saw that there is a strong negative correlation ($r_{x,y} = -0.764$) between the years of education and the resting pulse rate.

In this case, it does not make sense to say that the cause of a lower resting pulsing rate is given by additional years of school.

The true fact is that additional years of school tend to be associated with a lower resting pulse rate, it is an association not a causation.

- In general, the explanation for an association is given by an unexpressed factor that is related to both data ***x*** and ***y*** under consideration.

  In the example years of school versus pulse rate, this unexpressed factor could be the exercise and good nutrition:

  - *a person who has spent additional time in school has more knowledge about the health and thus she/he may be more aware of the importance of exercise and good nutrition;*
  - *or, perhaps, it is not knowledge that is making the difference but rather the fact that a person with more education tends to have a job that allow her/him more time for exercise and good nutrition.*

  So, we have

  additional years of schooling → *exercise and good nutrition*

  → *reduced resting pulse rate.*

Another example. In a study of US Air Force, it was found a positive correlation between the precision of the bombings in Europe during the II World War and the reaction of enemy air force.

This strange counter-intuitive correlation can be explained by the unexpressed factor given by a bad weather:

bad weather → *less precision in the bombings*
bad weather → *reduced reaction of the enemy air force.*

Exercise. In the following situations is there a causation or an association? For the associations, say what is the unexpressed factor.

- In some seaside places, it has been observed a positive correlation between the consumption of ice-creams and the number of drownings.
- The positive correlation between IQ scores and salaries in the example we have previously seen.
- In some countries, it has been observed a positive correlation between sales of cars and sales of tires, as well as sales of cars and sales of television sets.
- In some towns, it has been observed a negative correlation between number of mice and number of cats and a positive correlation between number of cats and number of dogs.
- In football goalkeepers, it has been observed a positive correlation between the number of goals conceded and the number of matches played as well as a positive correlation between the number of goals conceded and the number of expulsions.