

Statistics: descriptive statistics, summarizing the data

S. Maset

Dipartimento di Matematica e Geoscienze, Università di Trieste

PEM 2018-2019

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 The histogrammed data x^{hist}
- 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 \mathbf{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 $\mathbf{x} \in \mathbb{R}^n$ is the arithmetic average of the components of \mathbf{x} :

$$\bar{x} := \frac{\sum_{i=1}^n x_i}{n}.$$

Example. The data

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

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

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

Observe that

$$\min_{i \in \{1, \dots, n\}} x_i \leq \bar{x} \leq \max_{i \in \{1, \dots, n\}} x_i.$$

Exercise. Prove this property.

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

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

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

We have

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

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

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

We have

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

Exercise. Prove these two properties.

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

Example. The scores in the last 14 years of the seria A winners (tournaments with 20 teams) are

$$\mathbf{y} = (95, 91, 91, 87, 102, 87, 84, 82, 82, 84, 85, 97, 91, 86)$$

(from season 2017-2018 down to to season 2004-2005).

Since

$$\mathbf{y} = 87 + (8, 4, 4, 0, 15, 0, -3, -5, -5, -3, -2, 10, 4, -1)$$

we have

$$\bar{y} = 87 + \frac{26}{14} = 88 + \frac{6}{7}.$$

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

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

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

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

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

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

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

Prove that

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

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

$$\mathbb{R}^n \rightarrow \mathbb{R}, \quad \mathbf{x} \mapsto \bar{x},$$

is linear.

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

$$\bar{x} = \frac{\sum_{j=1}^l f_j v_j}{\sum_{j=1}^l f_j} = \sum_{j=1}^l \hat{f}_j v_j$$

where f_j and \hat{f}_j are the frequency and the relative frequency, respectively, of v_j in $\mathbf{x}, j \in \{1, \dots, l\}$.

In fact, we have

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

and then

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{j=1}^l f_j v_j}{\sum_{j=1}^l f_j} \quad \text{and} \quad \bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{j=1}^l f_j v_j}{n} = \sum_{j=1}^l \frac{f_j}{n} v_j = \sum_{j=1}^l \hat{f}_j v_j.$$

Example: consider the frequency table

v_j	f_j
3	2
4	1
5	3

We have

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

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 \mathbf{x} and \mathbf{y} : the first for drivers with helmet and the second for drivers without helmet. Describe \mathbf{x} and \mathbf{y} .

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

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

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

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

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

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

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

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

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

the quantity

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

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

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

$$\bar{x} = \sum_{j=1}^I \hat{f}_j v_j \quad \text{and} \quad \sum_{j=1}^I \hat{f}_j = 1.$$

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

$$x_i - \bar{x}, \quad i \in \{1, \dots, n\},$$

are called **deviations** from the mean.

We have

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

Exercise. Prove this property of the deviations by showing that, for $a \in \mathbb{R}$, we have

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

if and only if $a = \bar{x}$.

(1) shows in what sense the mean can be considered as the center of the data: it is placed in the middle so that the sum of the deviations from it is zero.

- Now we can give a physical interpretation of the mean \bar{x} as the center of the data \mathbf{x}

If n objects of the same mass m are placed in a rod at the abscissas x_i , $i \in \{1, \dots, n\}$, then \bar{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 be 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 $a = \bar{x}$.

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

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

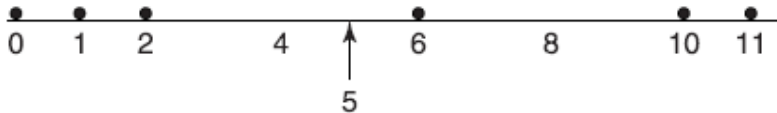
Example: consider

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

The mean

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

in this interpretation is shown below



Median

- The following data

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

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

The mean is $\bar{x} = 29.17$. 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 $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$, let

$$\mathbf{x}^{\text{ord}} := (x_1^{\text{ord}}, x_2^{\text{ord}}, \dots, x_n^{\text{ord}}).$$

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

As an example, for

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

we have

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

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

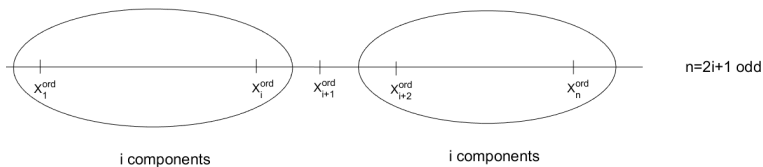
Formally, the median $m_{\mathbf{x}}$ of \mathbf{x} is defined as

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

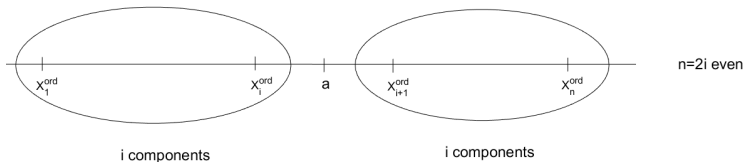
By this formal definition: if $n = 3$, the median is the second component of \mathbf{x}^{ord} ; if $n = 4$, it is the average of the second and the third components of \mathbf{x}^{ord} .

Exercise. For the data $\mathbf{x} = (0, -1, 0, 1, 2, 1, 0, -1)$ 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 \mathbf{x}^{ord} : the first i components of \mathbf{x}^{ord} are all $\leq x_{i+1}^{\text{ord}}$ and the last i components of \mathbf{x}^{ord} are all $\geq x_{i+1}^{\text{ord}}$.



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 \mathbf{x}^{ord} are all $\leq a$ and the last i components of \mathbf{x}^{ord} are all $\geq a$.

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

$$\mathbf{x}^{\text{ord}} = (23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 26, 89)$$

and then the median is

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

With respect to the mean $\bar{x} = 29.17$, 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 $\mathbf{x} = (0, -1, 0, 1, 2, 1, 0, -1)$ and $\mathbf{y} = (0, -1, 0, 1, 2000, 1, 0, -1)$. Are the means different? Are the medians different?

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

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

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

On the other hand, for $n \geq 3$, there is no change in the median since \mathbf{x}^{ord} and \mathbf{y}^{ord} differ only in the last component.

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

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

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

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

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

On the other hand, since $\mathbf{y}^{\text{ord}} = (\mathbf{x}^{\text{ord}}, a)$, as for the medians we have, for $n \geq 2$,

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

The change in the median is independent of a .

- *Example. Reconsider the scores of the winners of the seria A in the last 14 years*

$$\mathbf{y} = (95, 91, 91, 87, 102, 87, 84, 82, 82, 84, 85, 97, 91, 86),$$

where the mean is

$$\bar{y} = 88 + \frac{6}{7}.$$

On the other hand, the ordered n -tuple is

$$\mathbf{y}^{\text{ord}} = (82, 82, 84, 84, 85, 86, 87, 87, 91, 91, 91, 95, 97, 102)$$

and the median is

$$m_y = \frac{y_7^{\text{ord}} + y_8^{\text{ord}}}{2} = 87.$$

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

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

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

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

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

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

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

- 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 percentage 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 media 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.

HOME PAGE	SERIE A: 7ª giornata >		SI PARLA DI Serie A	
CALCIO				
PROBABILI FORMAZIONI				
CALCIOMERCATO				
CALCIO ESTERO				
FANTANEWS				
AUTOMOBILISMO				
MOTOCICLISMO				
PASSIONE MOTORI				
FESTIVAL DELLO SPORT				
CICLISMO				
BASKET				
NBA				
EUROLEGA				
FUORIGIOCO				
SPORT INVERNALI				
TENNIS LIVE				
RUGBY				
VOLLEY				
▼ ALTRI SPORT				
	SABATO 29 09 2018			
	Roma	Lazio	Fine	3 1
	Juventus	Napoli	Fine	3 1
	Inter	Cagliari	Fine	2 0
	DOMENICA 30 09 2018			
	Bologna	Udinese	Fine	2 1
	Chievo	Torino	Fine	0 1
	Fiorentina	Atalanta	Fine	2 0
	Frosinone	Genoa	Fine	1 2
	Parma	Empoli	Fine	1 0
	Sassuolo	Milan	Fine	1 4
	LUNEDÌ 01 10 2018			
	Sampdoria	SPAL	Fine	2 1
	SERIE A - 7ª GIORNATA			
	Juventus	21	Milan	9
	Napoli	15	Torino	9
	Fiorentina	13	SPAL	9
	Inter	13	Udinese	8
	Sassuolo	13	Bologna	7
	Genoa	12	Atalanta	6
	Lazio	12	Cagliari	6
	Sampdoria	11	Empoli	5
	Roma	11	Frosinone	1
	Parma	10	Chievo	-1
	Classifica		Calendario e risultati	

What is the appropriate statistical wording for saying that Milan is in the second column?

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

`median(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 ,

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

constructs a vector x with those frequencies and data values.

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

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

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 100 p th percentile, or the quantile p .

Informally, given a data $\mathbf{x} \in \mathbb{R}^n$, the 100 p th percentile of \mathbf{x} is defined as the value v that divides the n -tuple \mathbf{x}^{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 100 p th percentile of \mathbf{x} is defined as

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

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

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

Observe that:

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

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

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

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

- ▶ when np is not an integer, the $100p$ th percentile $v = x_{[np]}^{\text{ord}}$ has the property




























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

and the proportion of the components $x_1^{\text{ord}}, \dots, x_{[np]-1}^{\text{ord}}$ over all components is

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

(since $[np] - 1 < np < [np]$ and then $np - 1 < [np] - 1 < np$) and so this proportion is close to p .

Example: consider the 2017 Gross Domestic Product for countries in EU (in Billions of Euros)

 Germany	3,263.350	 Czech Republic	192.017
 United Kingdom	2,324.293	 Romania	187.868
 France	2,287.603	 Greece	177.735
 Italy	1,716.935	 Hungary	123.495
 Spain	1,163.662	 Slovakia	84.985
 Netherlands	731.168	 Luxembourg	55.378
 Sweden	477.858	 Bulgaria	50.430
 Poland	465.652	 Croatia	48.677
 Belgium	438.485	 Slovenia	43.278
 Austria	369.218	 Lithuania	41.857
 Ireland	296.152	 Latvia	26.857
 Denmark	288.374	 Estonia	23.002
 Finland	223.522	 Cyprus	19.214
 Portugal	193.049	 Malta	11.109

We determine the 90th percentile and the 20th 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

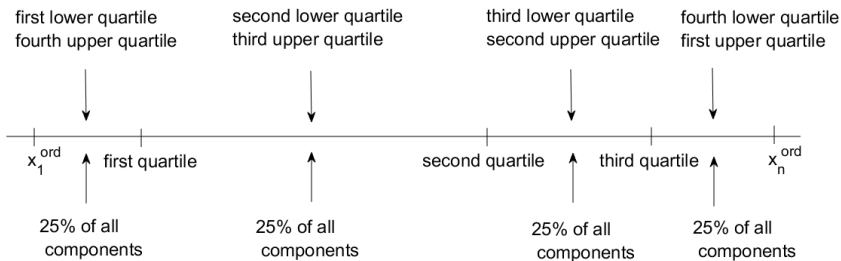
$$\lceil np \rceil = \begin{cases} 26 & \text{if } p = 0.9 \\ 6 & \text{if } p = 0.2. \end{cases}$$

The 90th percentile is 2,287.603 billions of euro (France's GDP) and the 20th percentile is 43.278 billions of euro (Slovenia's GDP).

- The **quartiles** of x are the 25th, 50th and 75th percentiles of 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 x^{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?

Exercise. Give a physical interpretation, similar to the physical interpretation of the mean, for the $100p$ th percentile, $p \in (0, 1)$, of a data $\mathbf{x} \in \mathbb{R}^n$.

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

`prctile(x, 100p)` or `quantile(x, p)`.

These two MATLAB functions `prctile` and `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

`percentile(x, 100p)`

for obtaining the right values of our definition.

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

Exercise. By using MATLAB, generate by

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

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

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

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

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

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

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

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

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

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

Mean and percentiles of a symmetric data

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

\mathbf{x}^{ord} has the form

$$\mathbf{x}^{\text{ord}} = \left(\underbrace{c - d_1, \dots, c - d_1}_{f_1 \text{ times}}, \dots, \underbrace{c - d_m, \dots, c - d_m}_{f_m \text{ times}}, \underbrace{c, \dots, c}_{f_{m+1} \text{ times}}, \right. \\ \left. \underbrace{c + d_m, \dots, c + d_m}_{f_m \text{ times}}, \dots, \underbrace{c + d_1, \dots, c + d_1}_{f_1 \text{ times}} \right)$$

For $i \in \{1, \dots, n\}$, we have $x_i^{\text{ord}} = c \mp d_{j_i}$, where $j_i \in \{1, \dots, m+1\}$ and $d_{m+1} = 0$ holds. Then

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

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

- The mean of \mathbf{x} is c .

In fact,

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

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

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

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

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

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

$$\lceil n - np \rceil = n + 1 - \lfloor np \rfloor .$$

In fact

$$\lfloor np \rfloor - 1 < np < \lfloor np \rfloor$$

and then

$$n - \lfloor np \rfloor < n - np < n - (\lfloor np \rfloor - 1) = n + 1 - \lfloor np \rfloor .$$

Thus

100(1 - p)-th percentile of \mathbf{x}

$$= x_{\lceil n - np \rceil}^{\text{ord}} = x_{n+1-\lfloor np \rfloor}^{\text{ord}}$$

$$= x_{\lfloor np \rfloor}^{\text{ord}} + 2 \left(c - x_{\lfloor np \rfloor}^{\text{ord}} \right)$$

$$= 100p\text{-th percentile of } \mathbf{x} + 2(c - 100p\text{-th percentile of } \mathbf{x}).$$

Exercise. Prove that the median of \mathbf{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

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

The mean is

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

The quartiles are

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

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

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

Mode

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

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

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

*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

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

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

$$\text{mode}(x).$$

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

$$I = \text{find}(f == \max(f)); v(I).$$

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

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

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

both of size $n = 5$. We have

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

but there is clearly more spread in \mathbf{b} than in \mathbf{a} .

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

$$x_i - \bar{x}, \quad i \in \{1, \dots, n\},$$

from the mean.

But, we cannot use

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

or the average

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

as a measure of the variability, since we have

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

We have to consider the absolute values of the deviations.

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

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

and

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

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 $x_i - \bar{x}$, $i \in \{1, \dots, n\}$: $n - 1$ deviations can be arbitrarily fixed and then the last one is determined by $\sum_{i=1}^n (x_i - \bar{x}) = 0$.

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

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

Example:

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

$$\bar{a} = 4$$

$$\mathbf{a} - \bar{a} = (-3, -2, 1, 2, 2)$$

$$s_a^2 = \frac{1}{4} (3^2 + 2^2 + 1^2 + 2^2 + 2^2) = \frac{22}{4} = 5.5;$$

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

$$\bar{b} = 4$$

$$\mathbf{b} - \bar{b} = (-44, -4, 1, 16, 31)$$

$$s_b^2 = \frac{1}{4} (44^2 + 4^2 + 1^2 + 16^2 + 31^2) = \frac{3170}{4} = 792.5.$$

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

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

is the moment of inertia of the system with respect to an axis perpendicular to the rod and passing through the center of mass \bar{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

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

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:

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

With this relation we have

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

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

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

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

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

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

Exercise. Prove these properties.

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

Example: let

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

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

By writing

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

we have

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

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

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

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

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 $\mathbf{x} \in \mathbb{R}^n$ is given by

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

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 mean because they have different dimensions.

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

$$s_{c+\mathbf{x}} = s_{\mathbf{x}} \quad \text{and} \quad s_{c\mathbf{x}} = |c| s_{\mathbf{x}}.$$

Exercise. Prove these properties of the standard deviation.

- For a data $\mathbf{x} \in \mathbb{R}^n$, observe that

$$s_x^2 = \frac{1}{n-1} \|\mathbf{x} - \bar{x}\|_2^2$$

$$s_x = \frac{1}{\sqrt{n-1}} \|\mathbf{x} - \bar{x}\|_2,$$

where $\mathbf{x} - \bar{x} = (x_1 - \bar{x}, \dots, x_n - \bar{x})$ and $\|\cdot\|_2$ is the euclidean norm.

Exercise. By using the triangle inequality

$$\|\mathbf{a} + \mathbf{b}\|_2 \leq \|\mathbf{a}\|_2 + \|\mathbf{b}\|_2$$

with $\mathbf{a} = \mathbf{x} - \bar{x}$ and $\mathbf{b} = \mathbf{y} - \bar{y}$, show that, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\mathbf{z} = \mathbf{x} + \mathbf{y}$, we have

$$s_z \leq s_x + s_y$$

and

$$s_z = s_x + s_y \iff \mathbf{x} = (c, \dots, c) \text{ for some } c \in \mathbb{R}$$

or $\mathbf{y} = \alpha \mathbf{x} + c$ for some $\alpha > 0$ and $c \in \mathbb{R}$.

Exercise. Show that, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have

$$|\mathbf{s}_y - \mathbf{s}_x| \leq \mathbf{s}_{y-x}$$

and

$$|\mathbf{s}_y - \mathbf{s}_x| = \mathbf{s}_{y-x} \Leftrightarrow \mathbf{x} = (c, \dots, c) \text{ for some } c \in \mathbb{R} \\ \text{or } \mathbf{y} = \alpha \mathbf{x} + \mathbf{c} \text{ for some } \alpha > 0 \text{ and } \mathbf{c} \in \mathbb{R}.$$

- Exercise. What is the relative error

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

of \hat{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

`std(x)` or `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 $\mathbf{x} \in \mathbb{R}^n$ is the **interquartile range** of \mathbf{x} given by

interquartile range of $\mathbf{x} :=$ third quartile of \mathbf{x} – first quartile of \mathbf{x} .

Roughly speaking, the interquartile range of \mathbf{x} is the length of the interval in which the middle half of the components of \mathbf{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

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

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

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

The first, second and third quartiles are

$$\text{for } p = 0.25 : y_{\lceil 2.5 \rceil}^{\text{ord}} = y_3^{\text{ord}} = 178$$

$$\text{for } p = 0.5 : \frac{1}{2} (y_5^{\text{ord}} + y_6^{\text{ord}}) = \frac{1}{2} (180 + 183) = 181.5$$

$$\text{for } p = 0.75 : y_{\lceil 7.5 \rceil}^{\text{ord}} = y_8^{\text{ord}} = 185.$$

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

The median 181.5 cm and the interquartile range 7 cm should be compared with the mean 181.6 cm and the standard deviation 4.48 cm.

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

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

Use

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

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, February, March and April 2018. Determine for each month the mean, the median, the standard deviation and the interquartile range.

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

Exercise. Let $\mathbf{x} \in \mathbb{R}^n$, let $c > 0$ and let $\mathbf{y} = c\mathbf{x}$. Prove that
interquartile range of $\mathbf{y} = c \cdot$ interquartile range of \mathbf{x} .

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

Exercise. Let $\mathbf{x} \in \mathbb{R}^n$, let $c \in \mathbb{R}$ and let $\mathbf{y} = c\mathbf{x}$. Prove that
interquartile range of $\mathbf{y} = |c| \cdot$ interquartile range of \mathbf{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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{x} = \mathbf{x}^{\text{ord}}$ and $\mathbf{y} = \mathbf{y}^{\text{ord}}$, and let $\mathbf{z} = \mathbf{x} + \mathbf{y}$. Prove that

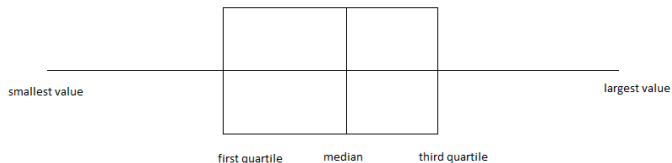
interquartile range of $\mathbf{z} =$ interquartile range of $\mathbf{x} +$ interquartile range of \mathbf{y} .

This property is not satisfied by the standard deviation.

Box plot

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

It is composed by a segment on the real line with extremes the smallest data value x_1^{ord} and the largest data value x_n^{ord} . 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

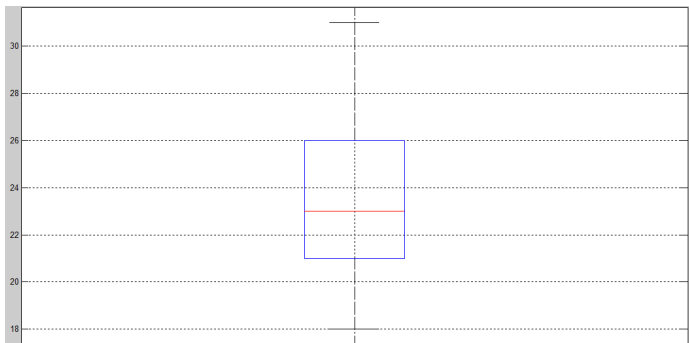
Since $n = 38$, the quartiles are

$$p = 0.25 : x_{[9.5]}^{\text{ord}} = x_{10}^{\text{ord}} = 21$$

$$p = 0.5 : \frac{x_{19}^{\text{ord}} + x_{20}^{\text{ord}}}{2} = 23$$

$$p = 0.75 : x_{[28.5]}^{\text{ord}} = x_{29}^{\text{ord}} = 26.$$

Box plot:



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

`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).

The histogrammed data \mathbf{x}^{hist}

- Consider a data $\mathbf{x} \in \mathbb{R}^n$ and a histogram for this data with K class intervals.

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

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

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 \mathbf{x}^{hist} ?

The data \mathbf{x}^{hist} contains the same information of the histogram: from the histogram we can construct \mathbf{x}^{hist} and, viceversa, from \mathbf{x}^{hist} we can obtain the histogram.

\mathbf{x}^{hist} is called the histogrammed version of \mathbf{x} .

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

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

The data \mathbf{x}^{hist} has frequency table

v_j	f
5	3
15	2
25	2
35	2

and so

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

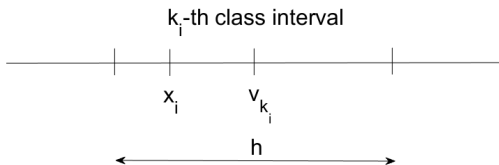
- In MATLAB the data \mathbf{x}^{hist} is obtained by

```
N = histcounts(x, boundaries);  
middlepoints = 1/2 * (boundaries(1 : end - 1) + boundaries(2 : end));  
xhist = construct(N, middlepoints);
```

for the data in the vector x and the class boundaries in the vector boundaries .

Exercise. By using MATLAB construct \mathbf{x}^{hist} of the previous example.

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



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

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

where h is the length of the class intervals. So we can write

$$\mathbf{x}^{\text{hist}} = \mathbf{x} + \boldsymbol{\varepsilon}$$

where $\mathbf{x} \in \mathbb{R}^n$ is such that $|\varepsilon_j| \leq \frac{h}{2}$, $j \in \{1, \dots, n\}$.

- When we pass from a data $\mathbf{x} \in \mathbb{R}^n$ to an histogram for it, the data \mathbf{x} is lost and we can only reconstruct its approximation \mathbf{x}^{hist} , which contains the same information as the histogram.

The distance between corresponding components of \mathbf{x}^{hist} and \mathbf{x} is not larger than $\frac{h}{2}$, where h is the length of the class intervals. Also the distance between means, standard deviations and percentiles of \mathbf{x}^{hist} and \mathbf{x} is not larger than $\frac{h}{2}$, as we now show.

- The mean of \mathbf{x}^{hist} is close to the mean of \mathbf{x} :

$$\left| \overline{\mathbf{x}^{\text{hist}}} - \bar{x} \right| \leq \frac{h}{2}.$$

In fact

$$\begin{aligned} \left| \overline{\mathbf{x}^{\text{hist}}} - \bar{x} \right| &= \left| \overline{\mathbf{x} + \boldsymbol{\varepsilon}} - \bar{x} \right| = \left| \bar{x} + \bar{\boldsymbol{\varepsilon}} - \bar{x} \right| = \left| \bar{\boldsymbol{\varepsilon}} \right| \\ &= \left| \frac{\sum_{i=1}^n \varepsilon_i}{n} \right| = \frac{\left| \sum_{i=1}^n \varepsilon_i \right|}{n} \\ &\leq \frac{\sum_{i=1}^n |\varepsilon_i|}{n} \leq \frac{\sum_{i=1}^n \frac{h}{2}}{n} = \frac{h}{2}. \end{aligned}$$

- The standard deviation of \mathbf{x}^{hist} is close to the standard deviation of \mathbf{x} :

$$|s_{\mathbf{x}^{\text{hist}}} - s_{\mathbf{x}}| \leq \sqrt{\frac{n}{n-1}} \frac{h}{2}.$$

In fact, since $\mathbf{x}^{\text{hist}} = \mathbf{x} + \varepsilon$ we have (see a previous exercise)

$$|s_{\mathbf{x}^{\text{hist}}} - s_{\mathbf{x}}| \leq s_{\varepsilon}$$

and

$$\begin{aligned} s_{\varepsilon} &= \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n \varepsilon_i^2 - n\bar{\varepsilon}^2 \right)} \leq \sqrt{\frac{1}{n-1} \sum_{i=1}^n \varepsilon_i^2} \\ &\leq \sqrt{\frac{1}{n-1} \sum_{i=1}^n \left(\frac{h}{2} \right)^2} = \sqrt{\frac{n}{n-1} \left(\frac{h}{2} \right)^2} \\ &= \sqrt{\frac{n}{n-1}} \frac{h}{2}. \end{aligned}$$

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

$$|100p\text{th percentile of } \mathbf{x}^{\text{hist}} - 100p\text{th percentile of } \mathbf{x}| \leq \frac{h}{2}.$$

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

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

if np is not an integer, then

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

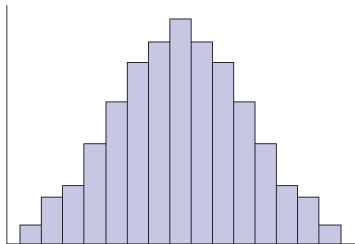
- Exercise. Consider as an histogram the following stem-and-leaf plot of a data \mathbf{x}

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							

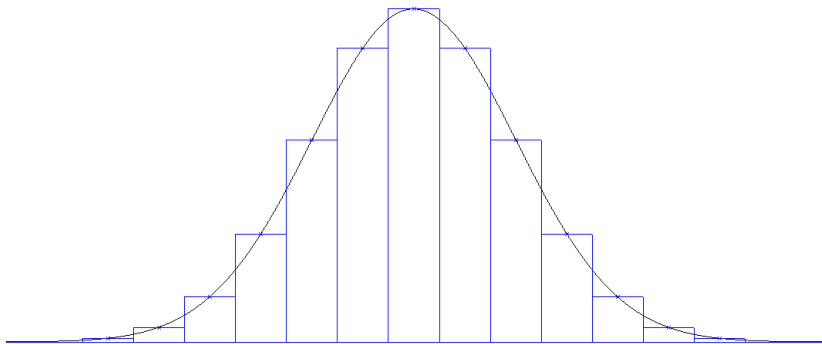
Compute the mean, the standard deviation, the quartiles and the modal values of \mathbf{x} and \mathbf{x}^{hist} .

Normal data

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



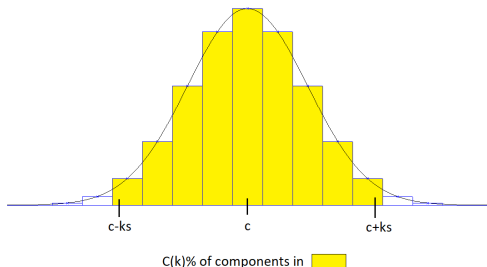
Qualitative description of the bell-shaped fashion decrease: in both sides, the curve interpolating the frequencies at the middle points of the intervals starts with an horizontal tangent at the middle point of the middle interval, then it is concave and, after a flex point, it becomes convex and goes 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, $C(k)\%$ of the components of \mathbf{x}^{hist} lie in the intervals between the interval containing $c - ks$ and the interval containing $c + ks$ (both included)

where c , the middle point of the middle interval, is the mean and the median of \mathbf{x}^{hist} and s is the standard deviation of \mathbf{x}^{hist} .



We express this by saying that $C(k)\%$ of the components of \mathbf{x}^{hist} are within k standard deviations from the mean.

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:

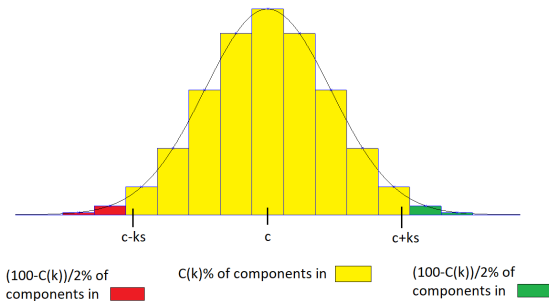
- 1) $C(1)\% = 68\%$, i.e. 68% of the components of \mathbf{x}^{hist} are within one standard deviation from the mean;
- 2) $C(2)\% = 95\%$, i.e. 95% of the components of \mathbf{x}^{hist} are within two standard deviations from the mean;
- 3) $C(3)\% = 99.7\%$, i.e. 99.7% of the components of \mathbf{x}^{hist} are within three standard deviations from the mean.

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

- The empirical rule for the bell fashion decrease in a normal data \mathbf{x} can be restated in terms of percentiles.

Since the histogram is symmetric with respect to the middle interval, the percentage of components of \mathbf{x}^{hist} in the intervals before the interval containing $c - ks$ (excluded) and the percentage of components of \mathbf{x}^{hist} in the intervals after the interval containing $c + ks$ (excluded) are equal and they are equal to

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



Let v_1 be the middle point of the interval containing $c - ks$ and let v_2 be the the middle point of the interval containing $c + ks$.

We have that

$$\frac{100 - C(k)}{2} \% = 100p\%, \quad \text{with } p = \frac{1 - C(k)\%}{2}$$

of the components of \mathbf{x}^{hist} are $< v_1$ and the rest is $\geq v_1$ and that

$$\left(100 - \frac{100 - C(k)}{2}\right) \% = 100(1 - p)\%$$

of the components of \mathbf{x}^{hist} are $\leq v_2$ and the rest is $> v_2$.

So

$C(k)\%$ of the components of \mathbf{x}^{hist} lie in the intervals between the interval containing $c - ks$ and the interval containing $c + ks$ (both included)

is equivalent to

v_1 is the $100p$ th of percentiles of \mathbf{x}^{hist} and v_2 is the $100(1 - p)$ th percentile of \mathbf{x}^{hist} .

By this equivalence, since

$$\frac{100 - C(1)}{2} = \frac{100 - 68}{2} = 16, \quad \frac{100 - C(2)}{2} = \frac{100 - 95}{2} = 2.5$$

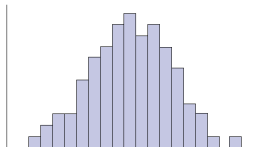
and $\frac{100 - C(3)}{2} = \frac{100 - 99.7}{2} = 0.15$

we can restate the points 1), 2) and 3) of the empirical rule as:

- 1bis) middle point of the interval containing $c - s = 16$ th percentile of \mathbf{x}^{hist}
middle point of the interval containing $c + s = 84$ th percentile of \mathbf{x}^{hist} .
- 2bis) middle point of the interval containing $c - 2s = 2.5$ th percentile of \mathbf{x}^{hist}
middle point of the interval containing $c + 2s = 97.5$ th percentile of \mathbf{x}^{hist} .
- 3bis) middle point of the interval containing $c - 3s = 0.15$ th percentile of \mathbf{x}^{hist}
middle point of the interval containing $c + 3s = 99.85$ th percentile of \mathbf{x}^{hist} .

Exercise. Show that, in our definition of percentiles, a percentile of \mathbf{x}^{hist} is a middle point of an interval or the middle point between two consecutive middle points of intervals (an interval boundary).

- A data \mathbf{x} is said **approximately normal** if it has a histogram with the following characteristics:
 - ▶ the histogram is approximately symmetric with respect to the mean interval: the mean interval is the interval containing the mean of \mathbf{x}^{hist} ;
 - ▶ the highest frequency is approximately in the mean interval;



- ▶ For any $k > 0$, the percentage of the components of \mathbf{x}^{hist} in the intervals between the interval containing $c - ks$ and the interval containing $c + ks$ (both included), with c middle point of the mean interval and s standard deviation of \mathbf{x}^{hist} , is approximately $C(k)\%$.

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: since $\overline{x^{\text{hist}}} = 74.6$, the mean interval is the interval with middle point $c = 75$.

A confirmation that the histogram satisfies the first characteristic is the closeness of mean and median of x^{hist} : $m_{x^{\text{hist}}} = x_{13}^{\text{hist}} = 75$.

We use the empirical rule to check the bell-shaped fashion decrease. We have $c = 75$ and $s = 12.74$.

Since $c - s = 62.23$ and $c + s = 87.74$, we should find approximately 68% of the components in the intervals of middle points 65, 75 and 85 : $\frac{18}{25} = 72\%$ of components are in these intervals.

About the percentiles we have

middle point of the interval containing $c - s = 65$

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

middle point of the interval containing $c + s = 85$

$$84\text{th percentile} = \frac{1}{2} (x_{21}^{\text{hist}} + x_{22}^{\text{hist}}) = 85 \quad (25 \cdot 0.84 = 21 \text{ is an integer}),$$

Since $c - 2s = 49.52$ and $c + 2s = 100.48$, we should find approximately 95% of the components in the intervals of middle points from 45 to 105 (both included): 100% of components are in these intervals.

Exercise. Check whether the maximum temperatures in Pordenone during January, collected in the years from 1999 to 2018 and found in

<http://www.ilmeteo.it/portale/archivio-meteo>,

are approximately normal.

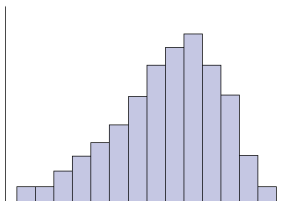
Exercise. Find on the web site

<https://www.testbusters.it/graduatorie-medicina-2018/>

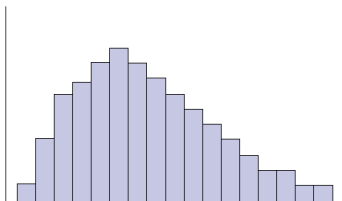
the results of the admission test to the Italian Doctor of Medicine degree for the academic year 2018-2019 and check whether they are approximately normal.

Other types of data

- A data $\mathbf{x} \in \mathbb{R}^n$ is said **skewed** if it has a histogram with the following characteristics:
 - ▶ there is an unique interval with the highest frequency, that is shifted with respect to the median interval (the interval containing the median of \mathbf{x}^{hist}) ;
 - ▶ from the interval with the highest frequency, the histogram decreases in the longest side (the side with the median interval) more slowly than the other side, with a long tail. The longest side is called the skewed side.



skewed to the left.



skewed to the right.

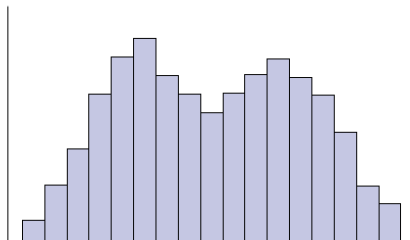
An example of a skewed data, indeed an example of a right-skewed data, is given by the pro capita personal incomes in the states of US.

Exercise. For a general histogram, characterize the position of the median interval (the interval containing the median of \mathbf{x}^{hist}) in terms of areas of the rectangles with bases the intervals. Also characterize the position of the mean interval (the interval containing the mean of \mathbf{x}^{hist}) in terms of an equilibrium.

Exercise. In a skewed data, for definition, the median interval is shifted with respect to the interval with the highest frequency in the skewed side. What about the position of the mean interval with respect to median interval?

Exercise. Find in www.statista.com the number of deaths in Finland in 2016 by age. Is this data skewed?

- A data $\mathbf{x} \in \mathbb{R}^n$ 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 \mathbf{x}^{hist} .

Exercise. What is the definition of m -modal data, where $m \in \{1, 2, 3, \dots\}$? Is a normal data or a skewed data m -modal for some $m \in \{1, 2, 3, \dots\}$?

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

Example : the following stem-and-leaf plot gives the weights in pounds (1pound = 0.45kg) 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, 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, 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, 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, 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 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, 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, 7, 7, 7, 7, 8, 8, 9, 9
11		0, 0, 1, 1, 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 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, 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 \mathbf{x} and \mathbf{y} are approximately normal.

For \mathbf{x}^{hist} , we have $\overline{\mathbf{x}^{\text{hist}}} = 126.44$ and so the mean interval is the interval with middle point $c = 125$. Mean and median are close: $m_{\mathbf{x}^{\text{hist}}} = x_{49}^{\text{hist}} = 125$. We have $s = 15.62$.

Since $c - s = 109.39$ and $c + s = 140.62$, we should find approximately 68% of the components in the intervals of middle points from 105 to 145 : $\frac{86}{97} = 88.7\%$ of components are in these intervals.

About the percentiles we have

middle point of the interval containing $c - s = 105$

16th percentile = $x_{[97 \cdot 0.16]}^{\text{hist}} = x_{16}^{\text{hist}} = 115$

middle point of the interval containing $c + s = 145$

84th percentile = $x_{[97 \cdot 0.84]}^{\text{hist}} = x_{82}^{\text{hist}} = 145$.

Since $c - 2s = 93.78$ and $c + 2s = 156.22$, we should find approximately 95% of the components in the intervals of middle points from 95 to 155 : $\frac{94}{97} = 96.9\%$ of components are in these intervals.

About the percentiles we have

middle point of the interval containing $c - 2s = 95$

$$2.5\text{th percentile} = x_{[97 \cdot 0.025]}^{\text{hist}} = x_3^{\text{hist}} = 95$$

middle point of the interval containing $c + 2s = 155$

$$97.5\text{th percentile} = x_{[97 \cdot 0.975]}^{\text{hist}} = x_{95}^{\text{hist}} = 155.$$

Since $c - 3s = 78.16$ and $c + 3s = 171.84$, we should find approximately 99.7% of the components in the intervals of middle points from 75 to 175 : 100% of components are in these intervals.

For y^{hist} , we have $\overline{y^{\text{hist}}} = 175.19$ and so the mean interval is the interval with middle point $c = 175$, it is not the interval with the maximum frequency. Mean and median are close:

$$m_{y^{\text{hist}}} = y_{52}^{\text{hist}} = 175. \quad \text{We have } s = 21.14.$$

Since $c - s = 153.86$ and $c + s = 196.14$, we should find approximately 68% of the components in the intervals of middle points from 155 to 195 : $\frac{80}{103} = 77.7\%$ of components are in these intervals.

About the percentiles we have

middle point of the interval containing $c - s = 155$

$$16\text{th percentile} = y_{\lceil 103 \cdot 0.16 \rceil}^{\text{hist}} = y_{17}^{\text{hist}} = 155$$

middle point of the interval containing $c + s = 195$

$$84\text{th percentile} = y_{\lceil 103 \cdot 0.84 \rceil}^{\text{hist}} = y_{87}^{\text{hist}} = 195.$$

Since $c - 2s = 132.71$ and $c + 2s = 217.29$, we should find approximately 95% of the components in the intervals of middle points from 135 to 215 : $\frac{101}{103} = 98.1\%$ of components are in these intervals.

About the percentiles we have

middle point of the interval containing $c - 2s = 135$

2.5th percentile = $y_{\lceil 103 \cdot 0.025 \rceil}^{\text{hist}} = y_3^{\text{hist}} = 135$

middle point of the interval containing $c + 2s = 215$

97.5th percentile = $y_{\lceil 103 \cdot 0.975 \rceil}^{\text{hist}} = y_{101}^{\text{hist}} = 215$.

Since $c - 3s = 111.57$ and $c + 3s = 238.43$, we should find approximately 99.7% of the components in the intervals of middle points from 105 to 235 : $\frac{102}{103} = 99.0\%$ of components are in these intervals.

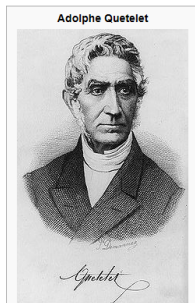
Some considerations about normal data

- It is a fact that:
 - ▶ if the data $\mathbf{x} \in \mathbb{R}^n$ 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 $n \rightarrow \infty$ and $h \rightarrow 0$, h being the length of the 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")



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 $1 \text{ inch} = 2.54 \text{ cm}$ and he found that, with the exception of the class intervals around $62 \text{ inch} = 151.9 \text{ cm}$, 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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. We assume that both \mathbf{x} and \mathbf{y} have not all the components equal.

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

We say that:

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

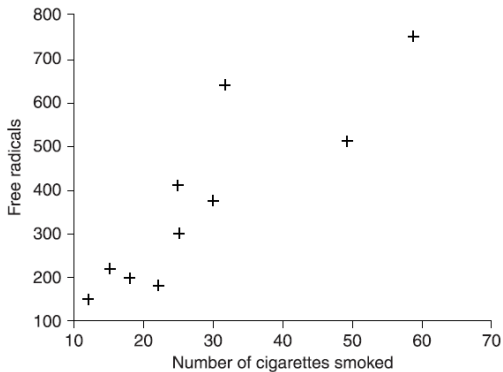
Example: consider the average daily number of cigarettes smoked (x) and the number of free radicals (y), in a suitable unit, found in the lungs of $n = 10$ smokers.

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:



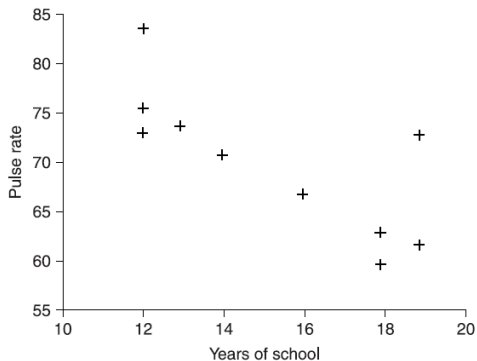
There is a positive correlation between x and y .

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 \mathbf{x} and \mathbf{y} , we observe that:
 - ▶ in case of a positive correlation, it is expected that, for many indices $i \in \{1, 2, \dots, n\}$, the deviations from the mean $x_i - \bar{x}$ and $y_i - \bar{y}$ have the same sign, i.e. $(x_i - \bar{x})(y_i - \bar{y})$ is positive;
 - ▶ in case of a negative correlation, it is expected that, for many indices $i \in \{1, 2, \dots, n\}$, $x_i - \bar{x}$ and $y_i - \bar{y}$ have different sign, i.e. $(x_i - \bar{x})(y_i - \bar{y})$ is negative.

Therefore, we can consider

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

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

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

$$\frac{x_i - \bar{x}}{\|\mathbf{x} - \bar{x}\|_2} = \frac{x_i - \bar{x}}{\sqrt{\sum_{k=1}^n (x_k - \bar{x})^2}}$$

$$\frac{y_i - \bar{y}}{\|\mathbf{y} - \bar{y}\|_2} = \frac{y_i - \bar{y}}{\sqrt{\sum_{k=1}^n (y_k - \bar{y})^2}}$$

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

Since we are assuming that both \mathbf{x} and \mathbf{y} have not all the components equal, both $\mathbf{x} - \bar{x}$ and $\mathbf{y} - \bar{y}$ are different from $\mathbf{0}$.

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 \mathbf{x} and \mathbf{y} , the quantity

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

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

- Previously, we have given an informal and qualitative definition of positive and negative correlations between the data \mathbf{x} and \mathbf{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 \mathbf{x} and \mathbf{y} if $r_{x,y} > 0$;
- ▶ there is a **negative correlation** between \mathbf{x} and \mathbf{y} if $r_{x,y} < 0$.

- The correlation coefficient can be written as

$$r_{x,y} = \frac{\langle \mathbf{x} - \bar{\mathbf{x}}, \mathbf{y} - \bar{\mathbf{y}} \rangle}{\|\mathbf{x} - \bar{\mathbf{x}}\|_2 \cdot \|\mathbf{y} - \bar{\mathbf{y}}\|_2}, \quad (2)$$

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

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

and

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

To better understand this, observe that the right-hand side of (2) is $\cos \theta$, where $\theta \in [0, \pi]$ is the angle between the vectors $\mathbf{x} - \bar{\mathbf{x}}$ and $\mathbf{y} - \bar{\mathbf{y}}$: so $|r_{x,y}| = |\cos \theta| \leq 1$, $r_{x,y} = \cos \theta = 1 \Leftrightarrow \theta = 0$ (i.e. $\mathbf{y} - \bar{\mathbf{y}}$ is a positive multiple of $\mathbf{x} - \bar{\mathbf{x}}$) and $r_{x,y} = \cos \theta = -1 \Leftrightarrow \theta = \pi$ (i.e. $\mathbf{y} - \bar{\mathbf{y}}$ is a negative multiple of $\mathbf{x} - \bar{\mathbf{x}}$).

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

$$\mathbf{y} - \bar{y} = a(\mathbf{x} - \bar{x})$$

for some $a \neq 0$ and this means

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

i.e.

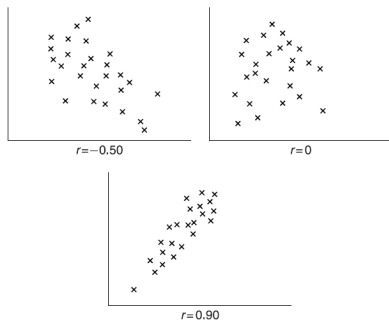
$$y_i = \underbrace{a}_{=m} x_i + \underbrace{\bar{y} - a\bar{x}}_{=q}, \quad i \in \{1, \dots, n\}.$$

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 \mathbf{x} and \mathbf{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 the correlation between \mathbf{x} and \mathbf{y} .

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



We say that \mathbf{x} and \mathbf{y} are **uncorrelated** if $r_{x,y} = 0$, **correlated** if $r_{x,y} \neq 0$, **strongly correlated** if $|r_{x,y}| \geq 0.7$ and **weakly correlated** if $|r_{x,y}| \leq 0.3$. 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 \mathbf{p} and \mathbf{q} are given by

$$\mathbf{p} = \alpha \mathbf{x} + \beta$$

$$\mathbf{q} = \gamma \mathbf{y} + \delta,$$

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

In fact

$$\begin{aligned} r_{p,q} &= \frac{\sum_{i=1}^n (\alpha x_i + \beta - \overline{\alpha \mathbf{x} + \beta}) \cdot (\gamma y_i + \delta - \overline{\gamma \mathbf{y} + \delta})}{(n-1) \cdot \mathbf{s}_{\alpha \mathbf{x} + \beta} \cdot \mathbf{s}_{\gamma \mathbf{y} + \delta}} \\ &= \frac{\sum_{i=1}^n (\alpha x_i + \beta - \alpha \bar{x} - \beta) \cdot (\gamma y_i + \delta - \gamma \bar{y} - \delta)}{(n-1) \cdot |\alpha| \mathbf{s}_x \cdot |\gamma| \mathbf{s}_y} \\ &= \frac{\sum_{i=1}^n \alpha (x_i - \bar{x}) \cdot \gamma (y_i - \bar{y})}{(n-1) \cdot |\alpha| \mathbf{s}_x \cdot |\gamma| \mathbf{s}_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \mathbf{s}_x \mathbf{s}_y} = r_{x,y}. \end{aligned}$$

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 \mathbf{x} and \mathbf{y} are data of temperatures measured in F° . Does the correlation coefficient between \mathbf{x} and \mathbf{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

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

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

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

Exercise. Prove this new formula.

Example: consider the paired data

<i>Year</i>	<i>x (whole milk)</i>	<i>y (low-fat milk)</i>
1980	17.1	10.6
1984	14.7	11.5
1988	12.8	13.2

giving the US pro-capita consumption (in gallons: 1 gallon = 3.785 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 \mathbf{x} and \mathbf{y} , we use

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

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

We have

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

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

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

$$\begin{aligned} r_{\mathbf{x},\mathbf{y}} = r_{\mathbf{p},\mathbf{q}} &= \frac{\sum_{k=1}^3 p_k q_k - 3 \cdot \bar{p} \cdot \bar{q}}{\sqrt{\left(\sum_{k=1}^3 p_k^2 - 3\bar{p}^2\right) \left(\sum_{k=1}^3 q_k^2 - 3\bar{q}^2\right)}} \\ &= \frac{1.71 - 3 \cdot 2.0667 \cdot 1.1667}{\sqrt{(22.10 - 3 \cdot 2.0667^2)(7.57 - 3 \cdot 1.1667^2)}} = -0.97. \end{aligned}$$

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

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

$\text{corrcoef}(x,y)$ 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

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

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 $\mathbf{x} = (0, 0.1, 0.2, \dots, 1)$ and \mathbf{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. $y = \text{rand}(11, 1)$ in MATLAB.

- Exercise. Collect on Wikipedia for the twenty teams of Seria A in season 2017-2018, the number of goals scored (data \mathbf{x}), the number of goals conceded (data \mathbf{y}) and the number of points obtained (data \mathbf{z}). Study the correlation between \mathbf{x} and \mathbf{z} , \mathbf{y} and \mathbf{z} , and \mathbf{x} and \mathbf{y} .

The regression line

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

When $|r_{x,y}| < 1$, there is not a straight line passing through all the points (x_i, y_i) , $i \in \{1, \dots, n\}$, but the trend is given by the **regression line** of the paired data \mathbf{x} and \mathbf{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 \mathbf{x} tend to go with smaller values of \mathbf{y} and larger values of \mathbf{x} values tend to go with larger values of \mathbf{y} ;
- ▶ negative sign: the regression line heads downward and smaller values of \mathbf{x} tend to go with larger values of \mathbf{y} and larger values of \mathbf{x} tend to go with smaller values of \mathbf{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

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

Moreover, we have

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

for the regression line.

Proof.

We use the deviations

$$\Delta_i = y_i - \bar{y} \text{ and } \delta_i = x_i - \bar{x}, \quad i \in \{1, \dots, n\}$$

and the vectors of the deviations

$$\mathbf{\Delta} = \mathbf{y} - \bar{y} = (\Delta_1, \dots, \Delta_n), \quad \mathbf{\delta} = \mathbf{x} - \bar{x} = (\delta_1, \dots, \delta_n) \in \mathbb{R}^n.$$

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

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

where

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

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 $\mathbf{z} = (m, p) \in \mathbb{R}^2$. We have that

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

has components

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

So the regression line is the line $y = m_0x + q_0$, with $q_0 = p_0 + (\bar{y} - m\bar{x})$, such that $\mathbf{z}_0 = (m_0, p_0)$ minimizes among all $\mathbf{z} = (m, p) \in \mathbb{R}^2$ the quantity

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

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 $\mathbf{c} \in \mathbb{R}^k$, find $\mathbf{u}_0 \in \mathbb{R}^\ell$ that minimizes among all $\mathbf{u} \in \mathbb{R}^\ell$ the quantity

$$\|\mathbf{c} - B\mathbf{u}\|_2^2.$$

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

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

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

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

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

Now, we prove this.

Proof.

We look for elements

$$\mathbf{b}_0 \in \text{ran}(B) = \{B\mathbf{u} : \mathbf{u} \in \mathbb{R}^\ell\}$$

minimizing $\|\mathbf{c} - \mathbf{b}\|_2^2$ among all $\mathbf{b} \in \text{ran}(B)$.

Once we have found such minimizing elements \mathbf{b}_0 , our minimums \mathbf{u}_0 are the vectors $\mathbf{u}_0 \in \mathbb{R}^\ell$ such that $B\mathbf{u}_0 = \mathbf{b}_0$.

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

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

It is a minimum of $\|\mathbf{c} - \mathbf{b}\|_2^2$ among all $\mathbf{b} \in \text{ran}(B)$ and it is the sole minimum. The minimum value $\|\mathbf{c} - \mathbf{b}_0\|_2^2$ satisfies

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

Proof.

In fact, we have, for $\mathbf{b} \in \text{ran}(B)$,

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

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

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

\mathbf{b}_0 is a minimum of $\|\mathbf{c} - \mathbf{b}\|_2^2$ among all $\mathbf{b} \in \text{ran}(B)$. Moreover, if $\mathbf{b} \in \text{ran}(B)$ is another minimum, then

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

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

Proof.

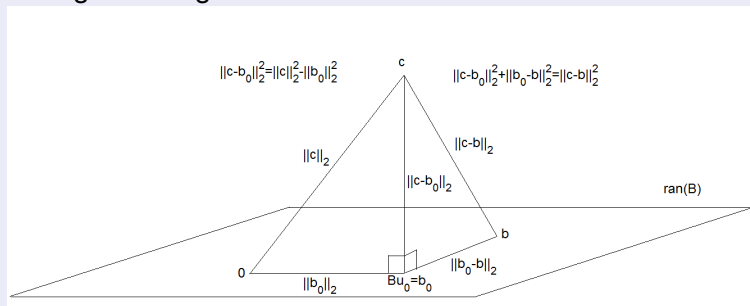
Finally, by taking $\mathbf{b} = \mathbf{0}$ in

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

which holds for all $\mathbf{b} \in \text{ran}(B)$, we obtain

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

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



Proof.

Now, the condition

$$\langle \mathbf{c} - \mathbf{b}_0, \mathbf{v} \rangle = 0 \text{ for any } \mathbf{v} \in \text{ran}(B) \quad (3)$$

is equivalent to

$$\langle \mathbf{c} - \mathbf{b}_0, \mathbf{b}^{(i)} \rangle = 0 \text{ for any } i \in \{1, \dots, \ell\}, \quad (4)$$

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

In fact: the columns of B are particular elements of $\text{ran}(B)$ and so (3) \Rightarrow (4); any $\mathbf{v} \in \text{ran}(B)$ is a linear combination

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

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

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

and so (4) \Rightarrow (3).

Proof.

Finally, the condition

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

can be expressed as

$$B^T (\mathbf{c} - \mathbf{b}_0) = \mathbf{0} \tag{5}$$

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

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

We conclude that a solution of the least squares problem is an element $\mathbf{u}_0 \in \mathbb{R}^\ell$ such that $B\mathbf{u}_0 = \mathbf{b}_0$, where \mathbf{b}_0 satisfies (5). So, we have to find the solutions \mathbf{u}_0 of the system

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

Moreover, we have

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

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

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

and

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

Proof.

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

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

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

Moreover, since for the minimum value *rss* we have

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

we obtain

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

- The regression line is the line $y = mx + q$ minimizing the euclidean norm

$$\|\mathbf{y} - m\mathbf{x} - q\|_2 = \sqrt{rss}$$

of

$$\mathbf{y} - m\mathbf{x} - q = (y_1 - mx_1 - q, \dots, y_n - mx_n - q)$$

and such minimal norm is

$$\begin{aligned} \sqrt{rss} &= \sqrt{(n-1) s_y^2 (1 - r_{x,y}^2)} = \sqrt{(n-1) s_y^2} \cdot \sqrt{1 - r_{x,y}^2} \\ &= \|\mathbf{y} - \bar{y}\|_2 \cdot \sqrt{1 - r_{x,y}^2} \end{aligned}$$

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

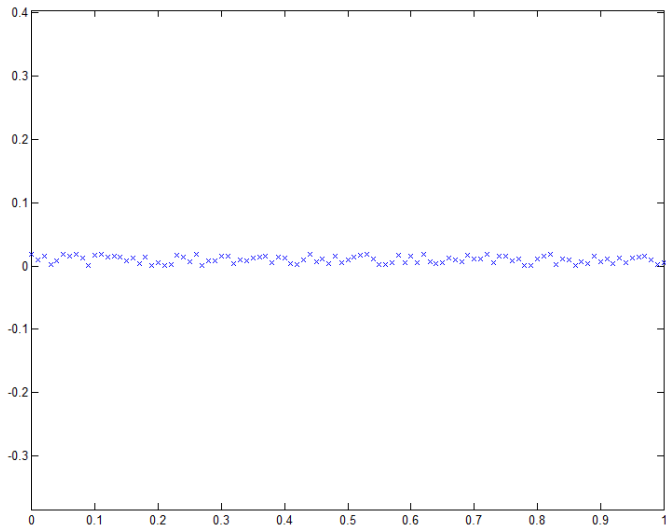
Question: is

$$\sqrt{rss} = \|\mathbf{y} - \bar{y}\|_2 \cdot \sqrt{1 - r_{x,y}^2}$$

a good measure of the error with respect to a perfect linear correlation?

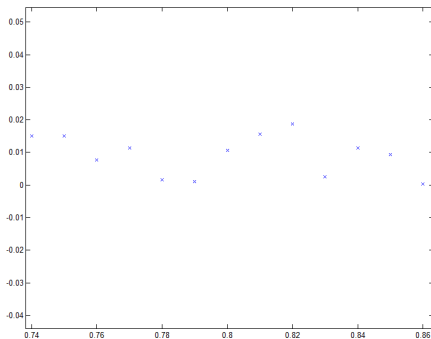
Observe that \sqrt{rss} is small when $\|\mathbf{y} - \bar{y}\|_2$ is small, i.e. the components of \mathbf{y} are close to the mean, or when $\sqrt{1 - r_{x,y}^2}$ is small, i.e. $r_{x,y}$ is close to one, i.e. we are close to have a perfect linear correlation between \mathbf{x} and \mathbf{y} .

Consider the scatter diagram



In this case \sqrt{rss} is small because $\|\mathbf{y} - \bar{y}\|_2$ 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

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

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

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

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 $\|\mathbf{y} - \bar{y}\|_2$ 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 $\|\mathbf{y} - \bar{y}\|_2$ is large.

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

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

or

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

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

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

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 minimum of f_1 or f_∞ .

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

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

Exercise. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and let $\mathbf{p} = \alpha\mathbf{x} + \beta$ and $\mathbf{q} = \gamma\mathbf{y} + \delta$, 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 \mathbf{x}, \mathbf{y} and the slope of the regression line for the paired data \mathbf{p}, \mathbf{q} .

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

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

- The regression line is computed in MATLAB by the function `regline`: for paired data in the vectors x and y

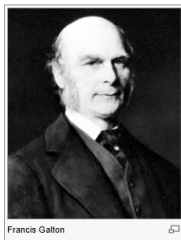
$$[m, q] = \text{regline}(x, y)$$

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)



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 $i \in \{1, \dots, n\}$,

- ▶ 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

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

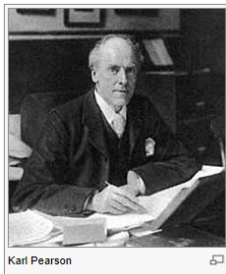
has $|m| = |r_{x,y}| \leq 1$ and so at a point (x, y) on the regression line we have

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

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)



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

Causation and association

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

Can we say that the smaller values of \mathbf{x} are the cause of the smaller (larger) values of \mathbf{y} and the larger values of \mathbf{x} are the cause of the larger (smaller) values of \mathbf{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.

- Observe that in the two previous examples, the unexpressed factor UF appears in two different position with respect to the correlated quantities A and B : for A = "years of schooling" and B = "pulse of rate", we have

$$A \rightarrow UF \rightarrow B$$

and for A = "reaction of enemy air force" and B = "precision in the bombings", we have

$$UF \rightarrow A \text{ and } UF \rightarrow B.$$

In the first situation we can say the A is an **indirect cause** of B . In the second situation that UF is the **common cause** of A and B .

Exercise. For A = "years of schooling" and B = "pulse of rate", try to think about some unexpressed factor UF with which we have

$$B \rightarrow UF \rightarrow A$$

or

$$UF \rightarrow A \text{ and } UF \rightarrow B.$$

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.

Exercise. In the following situations there is a causation (direct or indirect). Say what is the cause.

- ▶ The positive correlation between rotation speed of windmills and the speed of wind.
- ▶ The negative correlation between Body Mass Index and time cycling.
- ▶ The negative correlation in the years between debt and growth for some countries (for example Italy and Japan).
- ▶ The positive correlation for children between violence episodes and TV watching.

Look at Wikipedia "Correlation does not imply causation" for the causes.