Mathematics for Data Analysis
– Descriptive Statistics –

Matteo Gorgone

University of Messina, Department MIFT

email: mgorgone@unime.it

Statistics

Statistics (the name is historically related to the process of population census of states) is a brach of mathematics devoted to the collection, analysis, interpretation or explanation, and presentation of data.
When possible, statisticians collect data about the entire population (census).
When a census is not feasible, a chosen subset of the population (sample) is studied. Once a sample that is representative of the population is determined, data are collected for the sample members.

When data about the entire population, or a representative sample of the population have been collected, the obtained values are presented as a set of raw data. Unless the number of observations is small, raw data are unlikely to provide any information until they have been sorted in some way.
Then, the first step in the analysis falls into descriptive statistics: essentially, we focus on obtaining a small number of synthetic descriptors able to summarize the data.
Numerical descriptors include mean and standard deviation, but also median and mode, while frequencies and percentages are used in the description of categorical data.
Graphical representations of data are often used.

When we collect data of a sample, we must be aware of the elements of randomness in the choice of the sample. Therefore, to get meaningful conclusions about the entire population, methods of inferential statistics should be used, and probability results become essential.
The inferences may involve hypothesis testing, numerical estimates of the confidence of the descriptors of data, correlations between the different characters of the population (or of the sample).

Terminology

In applying statistics to a problem, we study what are called statistical units whose collection gives the statistical population.
For each statistical unit we measure and collect some characters (or variables). The variables of each statistical unit can be classified into two different types, depending on the type of values they take on:

1. numerical variable: if the values it assumes are numbers;
2. categorical variable: if the values it assumes are not numbers.

Example

Numerical variables: age, height, income, etc.
Categorical variables: eye colour, sex, breed, etc.
Some categorical data can be ordinal (like educational level), others cannot.

Classification of numerical variables

- A numeric variable is said to be discrete if the set of values that it can assume is finite or countable.
- A numeric variable is said to be continuous if the set of values that it can assume is the set $\mathbb{R}$ of real numbers or an interval of real numbers.

Example 1

By detecting with a measuring instrument the number of cosmic particles in 40 periods of one minute, the following data are obtained:
$\{0, 4, 2, 4, 1, 4, 4, 2, 3, 3, 1, 5, 2, 5, 3, 1, 8, 1, 2, 2, 5, 4, 2, 4, 1, 2, 3, 3, 3, 3, 1, 3, 3, 3, 2, 3, 2, 3, 5, 2\}$

We have $N = 40$ observations. The variable is discrete, because the number of observed particles is always a non-negative integer number, and the set of integers is infinite but countable.

Example 2

The following data are the result of 80 observations, in a given unit of measure, of the emission per day of a pollutant gas from an industrial plant:

$\{15.8, 22.7, 26.8, 19.1, 18.5, 14.4, 8.3, 25.9, 26.4, 9.8, 22.7, 15.2, 23.0, 29.6, 21.9, 10.5, 17.3, 6.2, 18.0, 22.9, 24.6, 19.4, 12.3, 15.9, 11.2, 14.7, 20.5, 26.6, 20.1, 17.0, 22.3, 27.5, 23.9, 17.5, 11.0, 20.4, 16.2, 20.8, 13.3, 18.1, 24.8, 26.1, 20.9, 21.4, 18.0, 24.3, 11.8, 17.9, 18.7, 12.8, 15.5, 19.2, 7.7, 22.5, 19.3, 9.4, 13.9, 28.6, 19.4, 21.6, 13.5, 24.6, 20.0, 24.1, 9.0, 17.6, 16.7, 16.9, 23.5, 18.4, 25.7, 20.1, 13.2, 23.7, 10.7, 19.0, 14.5, 18.1, 31.8, 28.5\}$

We have $N = 80$ observations. The variable is continuous, because the measurement of the quantity of gas emitted can be any positive real number.

Example 3

In a factory, the number of cases of malfunctions of a computer-controlled machine, and their causes, are recorded. The data for a certain month are as follows:

voltage fluctuations	6
instability of the control system	22
human error	13
instrument worn and not replaced	2
other causes	5
Total	48

We have $N = 48$ observations. The variable is categorical.

Frequency: absolute and relative

Suppose that for a population with N statistical units, we record a character (a variable) $\mathbf{x} = \{x_1, \ldots, x_N\}$ for each individual. Whatever the character is (numerical or categorical) we can compute the frequency of each value x_k of the character, that is the number of individuals having the value x_k. Suppose that we have M possible values.
The frequency of the k-th value can be absolute (denoted by f_k), or relative (the absolute frequency divided by N, denoted by $\widehat{f_k} = \dfrac{f_k}{N}$). The following relations are obvious:

$$\sum_{k=1}^{M} f_k = N, \qquad \sum_{k=1}^{M} \widehat{f_k} = 1.$$

Classes and frequency distribution

If the character is expressed by a number (for instance, the income), we can group the values in classes (for istance, each class contains the incomes in a given interval), and compute the frequency of each class.
Then, we construct a table representing the frequency distribution, *i.e.*, a table that collects the values according to the classes and the corresponding frequencies. The ordered data in the frequency distribution table are called grouped data.

About classes

The choice of how many classes and which intervals use are elements of arbitrariness. Several situations may arise:

- few classes, little information;
- too many classes, too many details;
- too many classes, not so much data.

Practical tips

- One can choose different number of classes, or classes with different extremes; in any case, the classes must not overlap and must contain all the data.
- The number of classes must depend on the number of data.
- Usually the classes all have the same size, but this characteristic is generally not mandatory and in some cases the data type can suggest the choice of different size classes.
- For a continuous numerical variable, it is necessary to specify whether the classes are closed to the right and/or to the left, that is if the data coinciding with the extremes of the class must be grouped in the class itself or in one of the adjacent classes.
- Once the data has been grouped, each exact value of the data is no longer used: all the data belonging to a certain class are represented by its midpoint, called the central value of the class.

Rules

- $$\text{classes number} = \sqrt{N};$$
- *Sturges*: $$\text{classes number} = 1 + \log_2 N;$$
- *Rice*: $$\text{classes number} = 2\sqrt[3]{N};$$
- *Freedman-Diaconis*: $$\text{classes size} = 2\frac{\text{IQR}}{\sqrt[3]{N}};$$
- $$\text{classes size} = \frac{\max(\{x_1, \dots, x_N\}) - \min(\{x_1, \dots, x_N\})}{\text{classes number}}.$$
- ...

Example 1: discrete numeric variables

In example 1, the observed variable is a discrete numeric variable, which can only assume integer values; since the assumed values are the integers $0, 1, 2, 3, 4, 5, 8$, it is natural to choose as classes the numbers $k = 0, 1, 2, 3, 4, 5, 6, 7, 8$ and count for each class the number of observations in which exactly k particles have been detected.

Class	f_k	$\widehat{f_k}$
0	1	0.025
1	6	0.15
2	10	0.25
3	12	0.3
4	6	0.15
5	4	0.1
6	0	0
7	0	0
8	1	0.025
Total	40	1

Example 1: discrete numerical variables

However, even for a discrete variable it is convenient use intervals as classes, instead of distinguishing all assumed values, especially when the data are numerous.

We can use classes comprising two possible values of the variable observed:

Class	f_k	$\widehat{f_k}$
$[0, 1]$	7	0.175
$[2, 3]$	22	0.55
$[4, 5]$	10	0.25
$[6, 7]$	0	0
$[8, 9]$	1	0.025
Total	40	1

Example 2: continuous numerical variables

In example 2, the observed variable is a continuous numeric variable. By using the Sturges rule

$$\text{classes number} = 1 + \log_2(80) \simeq 7$$

and

$$\text{classes size} = \frac{\max(\{x_1, \dots, x_N\}) - \min(\{x_1, \dots, x_N\})}{\text{classes number}} = \frac{31.8 - 6.2}{7} = 3.65714 \simeq 4,$$

we have

Class	f_k	$\widehat{f_k}$
$[5, 8.9]$	3	0.0375
$[9, 12.9]$	10	0.1250
$[13, 16.9]$	14	0.1750
$[17, 20.9]$	25	0.3125
$[21, 24.9]$	17	0.2125
$[25, 28.9]$	9	0.1125
$[29, 32.9]$	2	0.0250
Total	80	1

Classes have a "break" to avoid ambiguity!

Example 2: continuous numerical variables

It's better to choose classes closed on the left (open on the right) or classes closed on the right (open on the left):

Class	f_k	$\widehat{f_k}$
$[5,9[$	3	0.0375
$[9,13[$	10	0.1250
$[13,17[$	14	0.1750
$[17,21[$	25	0.3125
$[21,25[$	17	0.2125
$[25,29[$	9	0.1125
$[29,33[$	2	0.0250
Total	80	1

Class	f_k	$\widehat{f_k}$
$]5,9]$	4	0.05
$]9,13]$	9	0.1125
$]13,17]$	15	0.1875
$]17,21]$	24	0.3
$]21,25]$	17	0.2125
$]25,29]$	9	0.1125
$]29,33]$	2	0.0250
Total	80	1

Example 3: categorical variables

In Example 3, the "type of failure verified" variable is categorical and the data are already grouped in classes:

Class	f_k	$\widehat{f}_k$
voltage fluctuations	6	0.125
instability of the control system	22	0.458
human error	13	0.271
instrument worn and not replaced	2	0.042
other causes	5	0.104
Total	48	1

Cumulative frequency

Cumulative frequency is defined as a running total of frequencies, *i.e.*, the sum of all previous frequencies up to the current class. Cumulative frequencies are rapresented by a table of cumulative frequency distribution. We can consider absolute, relative and percentage cumulative frequencies.

Example 1: cumulative frequency distribution

Class	f_k
0	1
1	6
2	10
3	12
4	6
5	4
6	0
7	0
8	1
Total	40

Class	Absolute cum. freq.	Relative cum. freq.
$x \leq 0$	1	0.025
$x \leq 1$	7	0.175
$x \leq 2$	17	0.425
$x \leq 3$	29	0.725
$x \leq 4$	35	0.875
$x \leq 5$	39	0.975
$x \leq 6$	39	0.975
$x \leq 7$	39	0.975
$x \leq 8$	40	1

Example 1: cumulative frequency distribution

Class	f_k
$[0, 1]$	7
$[2, 3]$	22
$[4, 5]$	10
$[6, 7]$	0
$[8, 9]$	1
Total	40

Class	Absolute cum. freq.	Relative cum. freq.
$x \leq 1$	7	0.175
$x \leq 3$	29	0.725
$x \leq 5$	39	0.975
$x \leq 7$	39	0.975
$x \leq 9$	40	1

Example 2: cumulative frequency distribution

Class	f_k
$[5, 8.9]$	3
$[9, 12.9]$	10
$[13, 16.9]$	14
$[17, 20.9]$	25
$[21, 24.9]$	17
$[25, 28.9]$	9
$[29, 32.9]$	2
Total	80

Class	Absolute cum. freq.	Relative cum. freq.
$x \leq 4.9$	0	0
$x \leq 8.9$	3	0.725
$x \leq 12.9$	13	0.1625
$x \leq 16.9$	27	0.3375
$x \leq 20.9$	52	0.65
$x \leq 24.9$	69	0.8625
$x \leq 28.9$	78	0.975
$x \leq 32.9$	80	1

Example 2: cumulative frequency distribution

Class	f_k
$]5,9]$	4
$]9,13]$	9
$]13,17]$	15
$]17,21]$	24
$]21,25]$	17
$]25,29]$	9
$]29,33]$	2
Total	80

Class	Absolute cum. freq.	Relative cum. freq.
$x \leq 5$	0	0
$x \leq 9$	4	0.05
$x \leq 13$	13	0.1625
$x \leq 17$	28	0.35
$x \leq 21$	52	0.65
$x \leq 25$	69	0.8625
$x \leq 29$	78	0.975
$x \leq 33$	80	1

Example

Suppose we want to describe three different work organizations of 288 people, and the character we record for each individual is the number of completed operations in a day. Our data consist of three lists of 288 numbers.

Of course, it is quite difficult to compare the three organizations (in the following 3 slides we report the data) and select the best one. We need a way to extract some synthetic information or graphical representation of the data.

We can group the values in classes (this means that the interval between the minimum and maximum value of each list is divided in a fixed number of subintervals) and compute the frequencies of each class.

Graphical representation of data

The frequencies, as well as the distribution of the values, can be graphically presented in various ways!

Example: First Work Organization

{725, 724, 710, 724, 700, 724, 713, 692, 683, 712, 684, 707, 703, 691, 709, 702, 705, 715, 704, 705, 697, 725, 692, 719, 694, 717, 696, 707, 726, 703, 705, 712, 710, 697, 698, 694, 701, 715, 701, 707, 706, 701, 687, 708, 719, 713, 699, 702, 694, 708, 712, 704, 703, 687, 709, 693, 715, 707, 710, 700, 718, 702, 718, 705, 723, 718, 701, 698, 692, 684, 716, 710, 708, 707, 695, 726, 710, 709, 692, 707, 717, 709, 710, 718, 708, 720, 705, 714, 687, 707, 713, 708, 702, 686, 715, 696, 696, 711, 710, 715, 719, 717, 684, 705, 676, 695, 723, 707, 701, 692, 713, 700, 704, 726, 702, 706, 706, 700, 700, 687, 696, 694, 699, 709, 704, 704, 699, 706, 685, 713, 707, 690, 717, 721, 724, 704, 710, 697, 686, 713, 724, 688, 706, 715, 687, 702, 701, 708, 704, 705, 702, 701, 699, 699, 685, 712, 678, 706, 706, 695, 707, 718, 714, 698, 716, 714, 715, 702, 713, 710, 697, 711, 693, 697, 704, 714, 721, 703, 716, 706, 704, 717, 700, 692, 718, 699, 698, 690, 710, 703, 702, 719, 710, 725, 721, 713, 699, 703, 714, 707, 700, 716, 692, 719, 700, 709, 711, 702, 718, 712, 711, 691, 707, 714, 712, 698, 717, 714, 703, 709, 711, 704, 689, 712, 714, 711, 692, 720, 697, 698, 700, 689, 693, 707, 703, 712, 716, 713, 719, 712, 703, 705, 720, 704, 708, 712, 714, 713, 708, 696, 704, 699, 717, 695, 711, 697, 693, 701, 699, 697, 724, 713, 706, 705, 704, 707, 704, 719, 711, 700, 704, 706, 690, 703, 708, 694, 688, 703, 712, 722, 705, 700, 697, 697, 698, 705, 706, 694}

Example: Second Work Organization

{695, 686, 694, 690, 713, 704, 693, 697, 723, 694, 690, 721, 683, 701, 718, 715, 738, 694, 720, 680, 698, 691, 714, 699, 695, 709, 729, 717, 710, 714, 706, 711, 697, 728, 704, 692, 683, 696, 713, 674, 689, 683, 708, 704, 725, 695, 690, 696, 678, 725, 683, 700, 699, 705, 679, 710, 698, 686, 706, 731, 719, 693, 684, 684, 703, 691, 717, 681, 693, 709, 714, 688, 712, 688, 697, 729, 695, 697, 717, 679, 736, 671, 695, 739, 698, 696, 714, 711, 701, 720, 708, 706, 672, 713, 683, 695, 693, 670, 712, 677, 756, 693, 709, 688, 695, 722, 706, 686, 690, 685, 686, 681, 716, 709, 704, 679, 686, 676, 718, 683, 689, 696, 687, 736, 699, 685, 712, 723, 676, 693, 700, 745, 715, 743, 692, 718, 705, 708, 700, 713, 681, 723, 700, 698, 671, 714, 687, 687, 687, 683, 671, 677, 696, 696, 714, 713, 671, 688, 675, 671, 692, 725, 708, 699, 682, 686, 704, 714, 685, 711, 732, 688, 704, 720, 708, 733, 703, 693, 680, 690, 708, 704, 685, 685, 694, 702, 738, 702, 696, 709, 701, 687, 703, 701, 702, 693, 691, 701, 705, 685, 711, 693, 684, 670, 697, 732, 687, 737, 716, 716, 685, 741, 691, 705, 721, 735, 690, 705, 693, 698, 678, 704, 710, 686, 689, 686, 698, 684, 687, 696, 719, 679, 696, 701, 721, 681, 705, 714, 713, 678, 690, 721, 699, 725, 709, 718, 705, 744, 704, 686, 691, 712, 673, 698, 717, 711, 670, 726, 694, 723, 701, 683, 716, 671, 712, 704, 699, 705, 727, 719, 686, 699, 717, 688, 711, 695, 709, 699, 705, 718, 682, 697, 694, 670, 694, 708, 692, 702}

Example: Third Work Organization

{698, 737, 727, 725, 704, 706, 691, 747, 726, 722, 710, 710, 726, 733, 732, 702, 737, 701, 717, 731, 711, 729, 707, 752, 709, 696, 742, 716, 690, 709, 715, 715, 675, 704, 724, 749, 748, 742, 720, 705, 714, 716, 728, 722, 734, 701, 707, 688, 727, 724, 723, 739, 743, 720, 702, 710, 749, 735, 736, 713, 706, 746, 723, 710, 731, 705, 704, 758, 744, 740, 716, 704, 714, 728, 721, 707, 727, 720, 708, 717, 708, 730, 744, 759, 729, 752, 716, 753, 697, 745, 712, 721, 722, 740, 702, 721, 705, 698, 729, 697, 723, 722, 714, 694, 749, 743, 715, 712, 718, 730, 721, 720, 724, 720, 698, 716, 681, 712, 750, 728, 715, 698, 731, 733, 722, 708, 729, 718, 693, 698, 721, 705, 707, 713, 709, 737, 696, 715, 717, 739, 715, 729, 706, 731, 706, 700, 696, 719, 699, 692, 735, 733, 734, 704, 704, 723, 723, 714, 723, 702, 713, 716, 739, 748, 702, 704, 704, 705, 731, 710, 724, 721, 693, 717, 735, 730, 714, 706, 713, 729, 706, 720, 734, 728, 726, 684, 735, 734, 730, 713, 698, 730, 707, 732, 752, 706, 690, 700, 749, 710, 686, 712, 739, 703, 738, 705, 725, 703, 721, 725, 700, 710, 708, 719, 706, 706, 716, 747, 717, 715, 688, 732, 724, 710, 744, 721, 713, 709, 711, 746, 715, 740, 710, 719, 712, 719, 730, 715, 735, 712, 748, 695, 710, 726, 708, 713, 709, 729, 731, 717, 735, 705, 739, 692, 722, 725, 724, 728, 734, 709, 718, 717, 731, 714, 729, 702, 734, 724, 719, 735, 709, 734, 724, 705, 726, 726, 730, 734, 712, 730, 733, 700, 718, 702, 744, 727, 740, 701}

Pie chart

A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.
In this graph, the frequencies percentages are represented by circular sectors having amplitudes proportional to the frequencies; by denoting with f the percentage frequency and with g the amplitude in degrees, we have

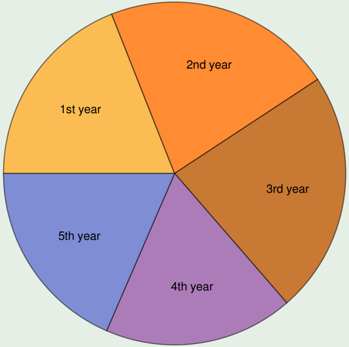
$$f : 100 = g : 360.$$

The pie chart is best suited for percentage frequencies and non-numerical variables.

Example: pie chart

Consider the number of students enrolled in the various years of a higher school (absolute frequencies) and the corresponding percentage frequencies.

Class	f_k	freq. %
1^{st} year	187	19.00%
2^{nd} year	214	21.75%
3^{rd} year	225	22.86%
4^{th} year	176	17.89%
5^{th} year	182	18.50 %
total	984	100.00%



Histograms

A histogram is a graph that provides a visual representation of the distribution of numerical data.

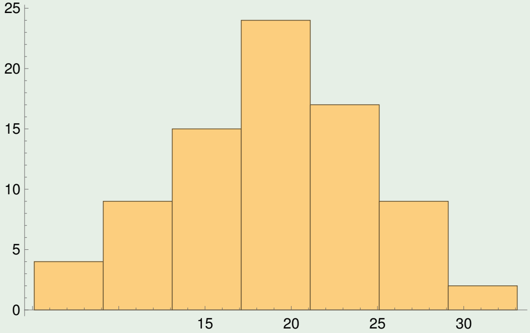
The histogram consists of adjacent rectangles, the bases of which are aligned on an oriented axis with a unit of measurement. The first step is divide the entire range of values into a series of intervals (group the values in classes) and then compute how many values fall into each interval (the absolute frequencies of each class). The classes must be adjacent (the adjacency shows the continuity of data) and are often (but not required to be) of equal size.

If the classes are of equal size, a rectangle is erected over the class with height proportional to the absolute frequency. A histogram may also be normalized to display relative frequencies. It then shows the proportion of cases that fall into each of several classes, with the sum of the heights equal to 1.

However, classes need not be of equal size; in that case, the erected rectangle has area proportional to the frequency of cases in the class. Thus, the vertical axis is not the frequency but the frequency density–number of cases per unit of the variable on the horizontal axis, *i.e.*, the ratio between the frequency of the class and the size of the class.

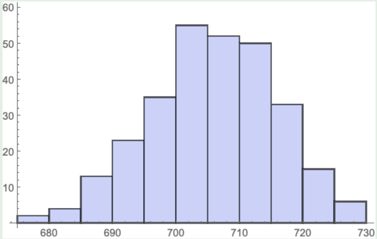
Example 2: histogram

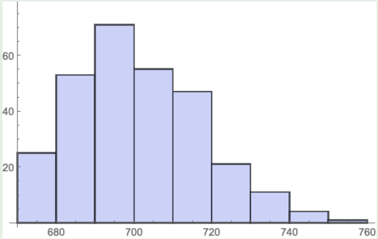
Class	f_k	$\widehat{f_k}$
$]5,9]$	4	0.05
$]9,13]$	9	0.1125
$]13,17]$	15	0.1875
$]17,21]$	24	0.3
$]21,25]$	17	0.2125
$]25,29]$	9	0.1125
$]29,33]$	2	0.0250
Total	80	1

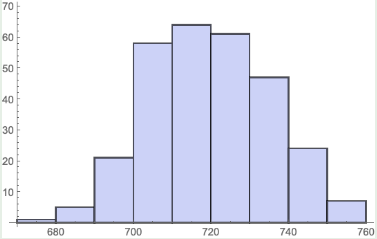


Example: histogram

For the three work organizations, we have the following histograms:





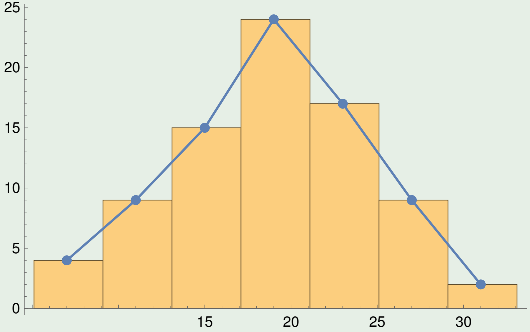


Frequency polygon

A frequency distribution can also be represented graphically with another type of graph: the frequency polygon. This polygon is obtained by joining together the points having as x–coordinate the middle value m_k of each class and as y–coordinate the corresponding frequency value.

Frequency polygon

Class	m_k	f_k
$]5, 9]$	7	4
$]9, 13]$	11	9
$]13, 17]$	15	15
$]17, 21]$	19	24
$]21, 25]$	23	17
$]25, 29]$	27	9
$]29, 33]$	31	2



In blue the frequency polygon.

Remark

Histograms are sometimes confused with bar charts. A histogram is used for continuous data, where the classes represent ranges of data, while a bar chart provides a plot of categorical variables.

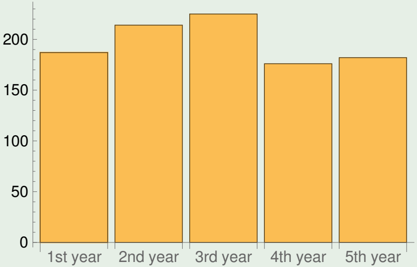
Bar chart

A bar chart is a graph that presents categorical or discrete data with rectangular bars with heights or width proportional to the values that they represent. The bars can be plotted vertically or horizontally. An axis of the graph shows the classes, and the other one represents the frequency of the classes. The size of the classes is constant; the rectangular bars are are usually not adjacent and are equidistant from each other. Some bar graphs present bars clustered in groups of more than one, showing the values of more than one measured variable. When there is no natural ordering of the categories being compared, bars on the chart may be arranged in any order.

Example: Bar chart

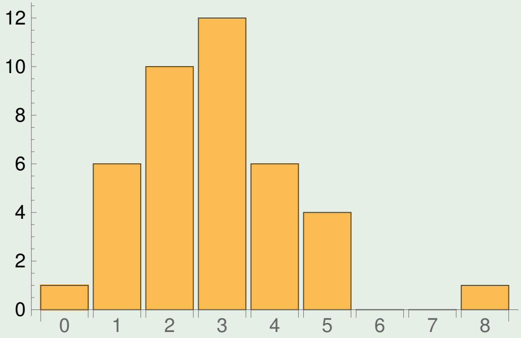
Consider the number of students enrolled in the various years of a higher school (absolute frequencies) and the corresponding percentage frequencies.

Class	f_k	freq. %
1^{st} year	187	19.00%
2^{nd} year	214	21.75%
3^{rd} year	225	22.86%
4^{th} year	176	17.89%
5^{th} year	182	18.50 %
total	984	100.00%



Example 1: Bar chart

Class	f_k
0	1
1	6
2	10
3	12
4	6
5	4
6	0
7	0
8	1
Total	40

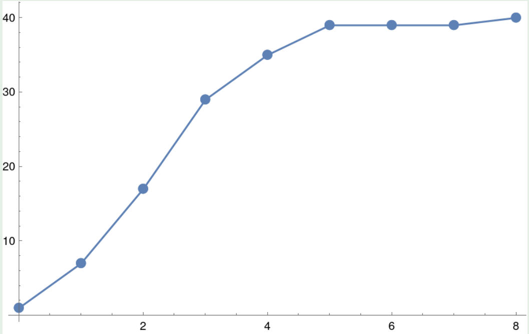


Cumulative polygon

An absolute cumulative frequency distribution is represented with a graph called cumulative polygon; the graph is obtained by plotting in the x-axis the upper limits of the classes and, for each of them, in y-axis the cumulative frequency of the corresponding class, and then joining the points obtained.

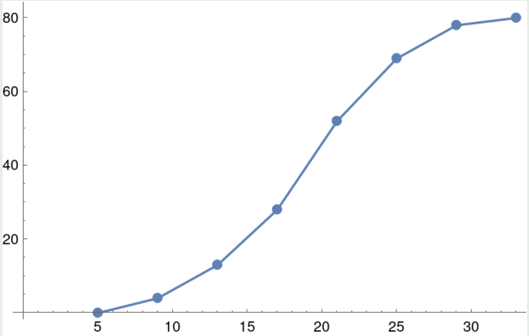
Example 1: cumulative polygon

Class	Absolute cum. freq.
$x \leq 0$	1
$x \leq 1$	7
$x \leq 2$	17
$x \leq 3$	29
$x \leq 4$	35
$x \leq 5$	39
$x \leq 6$	39
$x \leq 7$	39
$x \leq 8$	40



Example 2: cumulative polygon

Class	Absolute cum. freq.
$x \leq 5$	0
$x \leq 9$	4
$x \leq 13$	13
$x \leq 17$	28
$x \leq 21$	52
$x \leq 25$	69
$x \leq 29$	78
$x \leq 33$	80



Empirical distribution function

For relative cumulative frequency distributions, the graphical representation used (and that is important in probability) is given by the empirical distribution function (or empirical cumulative distribution function), say,

$$F(x) = \frac{\text{number of observations less or equal to } x}{\text{total number of observations}},$$

i.e., it is the ratio between the absolute cumulative frequencies related to a value x and the number of observations. It is a step function that jumps up by $1/N$ at each of the N data points.

We note that the cumulative relative frequencies coincide with the empirical distribution function evaluated in the upper limits of the classes.

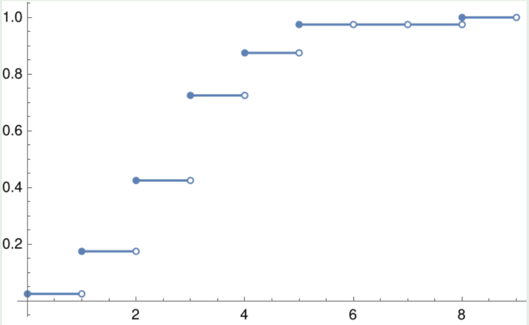
In other words, if $\{x_1, \dots, x_N\}$ are the ordered observations with corresponding relative frequencies $\widehat{f}_1, \dots, \widehat{f}_N$, the empirical distribution function che be written as

$$F(x) = \begin{cases} 0 & x < x_1, \\ F_i = \sum_{j=i}^{i} \widehat{f}_j & x_i \leq x < x_{i+1}, \\ 1 & x \geq x_N, \end{cases}$$

where F_i are the cumulative relative frequencies. In the plot, the x-axis represents the upper limits of the classes and and the y-axis the cumulative relative frequencies of the corresponding classes.

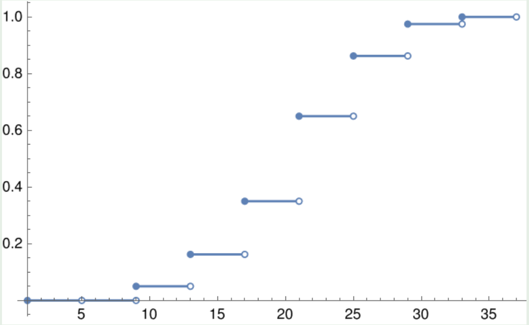
Example 1: empirical distribution function

Class	Relative cum. freq.
$x \leq 0$	0.025
$x \leq 1$	0.175
$x \leq 2$	0.425
$x \leq 3$	0.725
$x \leq 4$	0.875
$x \leq 5$	0.975
$x \leq 6$	0.975
$x \leq 7$	0.975
$x \leq 8$	1

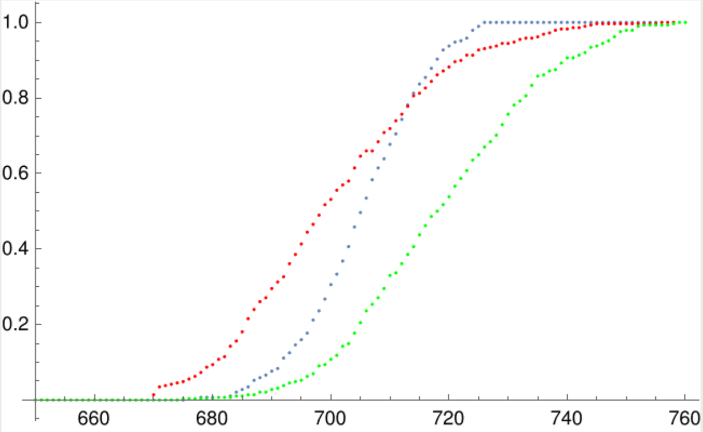


Example 2: empirical distribution function

Class	Relative cum. freq.
$x \leq 5$	0
$x \leq 9$	0.05
$x \leq 13$	0.1625
$x \leq 17$	0.35
$x \leq 21$	0.65
$x \leq 25$	0.8625
$x \leq 29$	0.975
$x \leq 33$	1



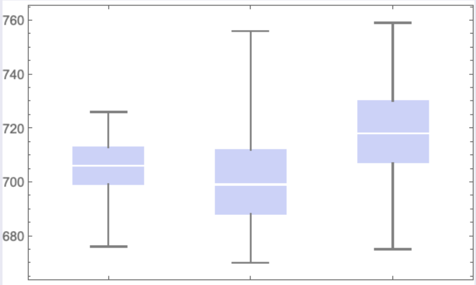
Example: empirical distribution function for the three work organizations



In blue, red, green, the first, second and third work organization, respectively.

Box and Whiskers plot

The box and whiskers plot is a standardized way of displaying the dataset based on the five-number summary: the minimum, the maximum, the median, the first and third quartiles.



The box and whiskers plot shows the minimum (bottom horizontal segment) and maximum (top horizontal segment) of data, as well as the first quartile (the bottom basis of gray rectangle), the median (the separator between the two gray rectangles) and the third quartile (the top basis of gray rectangle). The whiskers extending from the box indicate variability outside the first and third quartiles. The spacings in each subsection of the box and whiskers plot indicate the degree of dispersion (spread) and skewness of the data. Outliers that differ significantly from the rest of the dataset may be plotted as individual points beyond the whiskers on the box-plot.
The third work organization seems to be the best one.

Box and Whiskers Plot

It can be obtained if the data are numerical: in order to produce it we need to compute some synthetic descriptors of our data.

Synthetic descriptors of numerical data

1. Mean, mode, variance, standard deviation, coefficient of variation, skewness, kurtosis.
2. Minimum, maximum, median, quartiles, interquartile range.

Let us concentrate first on the descriptors in red that serve to produce the box and whiskers plot.

Minimum

Given a list of N numbers $\{x_1, x_2, \ldots, x_N\}$, the minimum is the lowest value x_k excluding any outliers.

Maximum

Given a list of N numbers $\{x_1, x_2, \ldots, x_N\}$, the maximum is the highest value x_k excluding any outliers.

Median

Given a list of N numbers $\{x_1, x_2, \ldots, x_N\}$, the median is the value x_k such that half numbers are less than x_k and half numbers are greater than x_k, *i.e.*, the *middle* value, when those numbers are listed in order from smallest to greatest.

Convention for median

Consider a list of N numbers $\{x_1, x_2, \ldots, x_N\}$ ordered from smallest to greatest. Then,

- if N is odd:
$$\text{median}(\{x_1, x_2, \ldots, x_N\}) = x_{\frac{N+1}{2}};$$
- if N is even:
$$\text{median}(\{x_1, x_2, \ldots, x_N\}) = \frac{x_{N/2} + x_{N/2+1}}{2}.$$

Remark

We note that the median is a value such that

$$F(\text{median}) = \frac{1}{2},$$

where $F(x)$ is the empirical distribution function.

Example

Consider the list $\{1, 4, 7, 12, 15\}$; the median is 7. Moreover, there is no number that leaves exactly 50% of the observations to its left and to its right, in fact:

$$F(7) = \frac{3}{5} = 0.6.$$

Example

Consider the list $\{1, 4, 7, 9, 12, 15\}$; in this case, every value between 7 and 9 is the median: it is usual in such cases to take as median the arithmetic mean between 7 and 9, that is

$$\text{median}(\{1, 4, 7, 9, 12, 15\}) = \frac{7+9}{2} = 8.$$

The median leaves exactly 50% of the observations to its left and to its right, in fact:

$$F(8) = \frac{3}{6} = 0.5.$$

Properties of the median

- if $x_1 = \cdots = x_N = a$, then
$$\text{median}(\{x_1, \ldots, x_N\}) = a;$$
- $\min(\{x_1, \ldots, x_N\}) \leq \text{median}(\{x_1, \ldots, x_N\}) \leq \max(\{x_1, \ldots, x_N\})$;
- given a list of N numbers $\{x_1, \ldots, x_N\}$ ordered from smallest to greatest, and let $m = \text{median}(\{x_1, \ldots, x_N\})$ be its median. Then:
$$\sum_{k=1}^{N} |x_k - m| \leq \sum_{k=1}^{N} |x_N - a|, \qquad a \quad \text{constant},$$
i.e., the median is the number that minimizes the sum of the absolute values of the differences between the data and a costant;
- the median is a robust descriptor, since it is not very sensitive to the presence of outliers;
- the median of a transformation that preserves the ordering of the data, that is an monotonically increasing transformation, coincides with the transformation of the median; if we define such a transformation $f(\cdot)$, with
$$y_k = f(x_k), \qquad k = 1, \ldots, N,$$
then
$$\text{median}(\{y_1, \ldots, y_N\}) = f\left(\text{median}(\{x_1, \ldots, x_N\})\right);$$

Grouped data: approximation of the median

Suppose we do not know the values of the statistical units but only the corresponding absolute frequencies f_k associated to the classes $]a_{k-1}, a_k]$, where $k = 1, \dots, M$.

The median cannot be computed exactly!

In these situations, a suitable choice is using a linear approximation:

$$\text{median} \approx a_{j-1} + (a_j - a_{j-1}) \frac{1/2 - F(a_{j-1})}{F(a_j) - F(a_{j-1})},$$

where median $\in\,]a_{j-1}, a_j]$, and F is the empirical distribution function.

Example: grouped data

Consider the classes $]a_{k-1}, a_k]$, where $k = 1, \ldots, 5$, and the corresponding absolute frequencies f_k as follows:

Class	f_k
$]0,1]$	1
$]1,2]$	4
$]2,3]$	4
$]3,4]$	2
$]4,5]$	1

We have $N = 12$. The median should be chosen between the $6-$th and $7-$th observation. Therefore, the median belongs to the class $]2,3]$, *i.e.*, $a_{j-1} = 2$ and $a_j = 3$.
Moreover, we have

$$F(2) = \frac{5}{12} = 0.42, \qquad F(3) = \frac{9}{12} = 0.75.$$

Then, by using the linear approximation for the median:

$$\text{median} \approx a_{j-1} + (a_j - a_{j-1}) \frac{1/2 - F(a_{j-1})}{F(a_j) - F(a_{j-1})} = 2 + (3-2)\frac{0.5 - 0.42}{0.75 - 0.42} = 2.25.$$

Example: grouped data

Consider the classes $]a_{k-1}, a_k]$, where $k = 1, \dots, 5$, and the corresponding absolute frequencies f_k as follows:

Class	f_k
$]0, 1]$	1
$]1, 2]$	4
$]2, 3]$	4
$]3, 4]$	2
$]4, 5]$	1

We have $N = 12$. The median should be chosen between the $6-$th and $7-$th observation.
Therefore, the median belongs to the class $]2, 3]$.
Alternatively, we can suppose that the four data belonging to the class $]2, 3]$ are equally distributed and, for example, are equal to 2.25, 2.5, 2.75, 3.
Under this assumption, the $6-$th and $7-$th observation are equal to 2.25 and 2.5, respectively.
Then, we can compute the median as

$$\text{median} = \frac{2.25 + 2.5}{2} = 2.375.$$

Quantiles

In statistics, quantiles are cut points dividing the observations in a sample in the same way.

A p–quantile, where $p \in [0, 1]$, is number that is larger $100 \cdot p\%$ of the observed data and smaller of the remaining $100 \cdot (1 - p)\%$.

By using the empirical distribution function F, the p-quantile can be defined as

$$Q_p = \inf\{x : F(x) \geq p\}.$$

First Quartile

Given a list of N numbers $\{x_1, x_2, \ldots, x_N\}$, the first quartile is the value x_k such that 25% of numbers are less than x_k and 75% of numbers are greater than x_k, *i.e.*, the middle number between the smallest number (minimum) and the median. It is the median of the lower half of the dataset!

Third Quartile

Given a list of N numbers $\{x_1, x_2, \ldots, x_N\}$, the third quartile is the value x_k such that 75% of numbers are less than x_k and 25% of numbers are greater than x_k, *i.e.*, the middle value between the median and the highest number (maximum). It is the median of the upper half of the dataset!

Remark

The median is just the Second Quartile!

Interquartile Range

The Interquartile Range (IQR) is the difference between the third and first quartile. It represents a measure of how the data are concentrated around the median.

Quartiles and outliers

There are methods by which to check for outliers in the discipline of statistics and statistical analysis. The basic idea of descriptive statistics, when encountering an outlier, is that we have to explain by further analysis the cause or origin of the outlier. In the case of quartiles, the Interquartile Range (IQR) may be used to characterize the data when there may be extremities that skew the data. The various quartiles are robust descriptors since they are scarcely affected by adding extremal values to the data (data greater than the maximum or less than the minimum).

There is also a mathematical method to determine “fences”, *i.e.*, upper and lower bounds from which to check for outliers. After determining the first and third quartiles, and the interquartile range, then fences are computed by using the following formula:

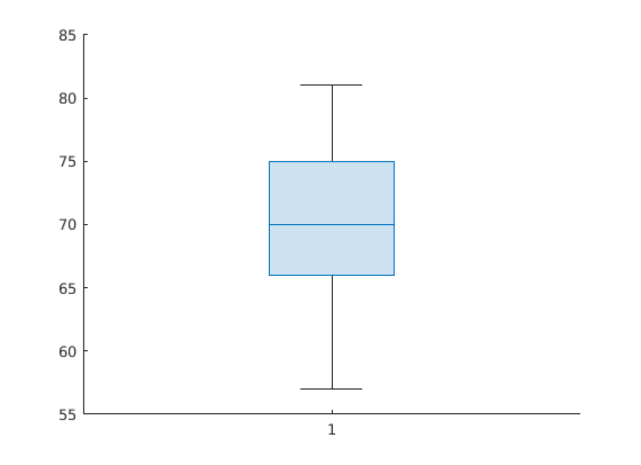
$$\text{Lower fence} = Q_1 - 1.5 \cdot IQR,$$

$$\text{Upper fence} = Q_3 + 1.5 \cdot IQR,$$

where Q_1 and Q_3 are the first and third quartiles, respectively. The lower fence is the lower limit and the upper fence is the upper limit of data, and any data lying outside these defined bounds can be considered an outlier.

Example without outliers: box and whiskers plot

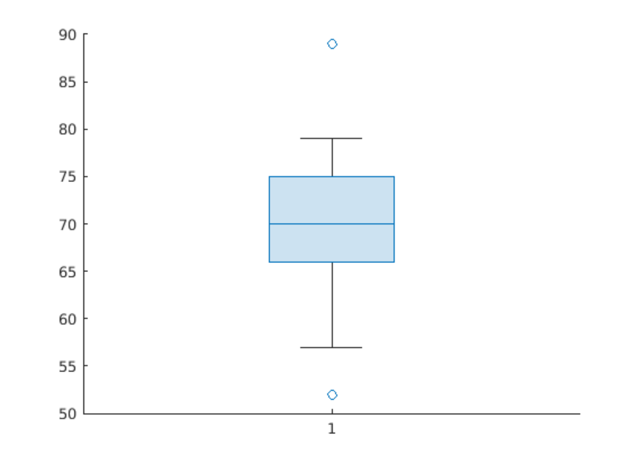
Let us consider the list
$\{57, 57, 57, 58, 63, 66, 66, 67, 67, 68, 69, 70, 70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 81\}$.



$$\text{minimum} = 57; \qquad Q_2 = \text{median} = 70; \qquad \text{maximum} = 81;$$
$$Q_1 = 66; \qquad Q_3 = 75; \qquad IQR = Q_3 - Q_1 = 9;$$
$$\text{Lower fence} = Q_1 - 1.5 \cdot IQR = 66 - 13.5 = 52.5;$$
$$\text{Upper fence} = Q_3 + 1.5 \cdot IQR = 75 + 13.5 = 88.5.$$

Example with outliers: box and whiskers plot

Let us consider the list
$\{52, 57, 57, 58, 63, 66, 66, 67, 67, 68, 69, 70, 70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 89\}$.



$$\text{minimum} = 52; \qquad Q_2 = \text{median} = 70; \qquad \text{maximum} = 89;$$
$$Q_1 = 66; \qquad Q_3 = 75; \qquad IQR = Q_3 - Q_1 = 9;$$
$$\text{Lower fence} = Q_1 - 1.5 \cdot IQR = 66 - 13.5 = 52.5;$$
$$\text{Upper fence} = Q_3 + 1.5 \cdot IQR = 75 + 13.5 = 88.5.$$

Computing methods for quartiles

Given a list of N numbers $\{x_1, x_2, \ldots, x_N\}$, there is no universal agreement on selecting the quartile values.

Method 1

1. Use the median to divide the ordered list into two-halves. Then:
   - if there is an odd number of elements in the original ordered list, do not include the median in either half;
   - if there is an even number of elements in the original ordered list, split this numbers set exactly in half;
2. the first quartile is the median of the lower half of the numbers, whereas the third quartile is the median of the upper half of the numbers.

Computing methods for quartiles

Example: Method 1 – odd number of elements

Consider the list $\{6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49\}$. We have $N = 11$.
The median

$$Q_2 = \text{median}(\{6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49\}) = 40,$$

must not be included in the two halves, which are

$$\{6, 7, 15, 36, 39\}, \qquad \{41, 42, 43, 47, 49\}.$$

The first quartile is the median of the lower half:

$$Q_1 = \text{median}(\{6, 7, 15, 36, 39\}) = 15.$$

The third quartile is the median of the upper half:

$$Q_3 = \text{median}(\{41, 42, 43, 47, 49\}) = 43.$$

Computing methods for quartiles

Example: Method 1 – even number of elements

Consider the list $\{7, 15, 36, 39, 40, 41\}$. We have $N = 6$.
The median

$$Q_2 = \text{median}(\{7, 15, 36, 39, 40, 41\}) = \frac{36 + 39}{2} = 37.5.$$

The list must be divided in two halves, which are

$$\{7, 15, 36\}, \qquad \{39, 40, 41\}.$$

The first quartile is the median of the lower half:

$$Q_1 = \text{median}(\{7, 15, 36\}) = 15.$$

The third quartile is the median of the upper half:

$$Q_3 = \text{median}(\{39, 40, 41\}) = 40.$$

Computing methods for quartiles

Method 2

1. Use the median to divide the ordered list into two-halves. Then:
   - if there is an odd number of elements in the original ordered list, include the median in both halves;
   - if there is an even number of elements in the original ordered list, split this data set exactly in half;
2. the first quartile is the median of the lower half of the numbers, whereas the third quartile is the median of the upper half of the numbers.

Computing methods for quartiles

Example: Method 2 – odd number of elements

Consider the list $\{6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49\}$. We have $N = 11$.
The median

$$Q_2 = \text{median}(\{6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49\}) = 40,$$

must be included in the two halves, which are

$$\{6, 7, 15, 36, 39, 40\}, \qquad \{40, 41, 42, 43, 47, 49\}.$$

The first quartile is the median of the lower half:

$$Q_1 = \text{median}(\{6, 7, 15, 36, 39, 40\}) = \frac{15 + 36}{2} = 25.5.$$

The third quartile is the median of the upper half:

$$Q_3 = \text{median}(\{40, 41, 42, 43, 47, 49\}) = \frac{42 + 43}{2} = 42.5.$$

Example: Method 2 – even number of elements

In the case of list with an even number of elements, Method 1 and Method 2 provide the same results.

Computing methods for quartiles

Method 3

1. Compute the median;
   - if there is an odd number of elements in the original ordered list, go to step 2 or 3;
   - if there is an even number of elements in the original ordered list, include the median as new number in the original list;
2. if there are $N = (4n+1)$ numbers, then the first quartile is 25% of the n-th number plus 75% of the $(n+1)$-th number, whereas the third quartile is 75% of the $(3n+1)$-th number plus 25% of the $(3n+2)$-th number, *i.e.*,

$$Q_1 = 0.25 \cdot x_n + 0.75 \cdot x_{n+1}, \qquad Q_3 = 0.75 \cdot x_{3n+1} + 0.25 \cdot x_{3n+2};$$

3. if there are $N = (4n+3)$ numbers, then the first quartile is 75% of the $(n+1)$-th number plus 25% of the $(n+2)$-th number, whereas the third quartile is 25% of the $(3n+2)$-th number plus 75% of the $(3n+3)$-th number, *i.e.*,

$$Q_1 = 0.75 \cdot x_{n+1} + 0.25 \cdot x_{n+2}, \qquad Q_3 = 0.25 \cdot x_{3n+2} + 0.75 \cdot x_{3n+3}.$$

Computing methods for quartiles

Example: Method 3 – odd number of elements

Consider the list $\{6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49\}$. We have $N = 11$.
The median is

$$Q_2 = \text{median}(\{6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49\}) = 40.$$

There are $N = (4n + 3)$ numbers, with $n = 2$. In fact, $11 = 4 \cdot 2 + 3$.
The first quartile is

$$Q_1 = 0.75 \cdot x_{n+1} + 0.25 \cdot x_{n+2} = 0.75 \cdot x_3 + 0.25 \cdot x_4 = 0.75 \cdot 15 + 0.25 \cdot 36 = 20.25.$$

The third quartile is

$$Q_3 = 0.25 \cdot x_{3n+2} + 0.75 \cdot x_{3n+3} = 0.25 \cdot x_8 + 0.75 \cdot x_9 = 0.25 \cdot 42 + 0.75 \cdot 43 = 42.75.$$

Computing methods for quartiles

Example: Method 3 – even number of elements

Consider the list $\{7, 15, 36, 39, 40, 41\}$. We have $N = 6$.
The median

$$Q_2 = \text{median}(\{7, 15, 36, 39, 40, 41\}) = \frac{36+39}{2} = 37.5.$$

and we include it as a new number in the list, *i.e.*, we consider the list

$$\{7, 15, 36, 37.5, 39, 40, 41\}.$$

There are $N = (4n + 3)$ numbers, with $n = 1$. In fact, $7 = 4 \cdot 1 + 3$.
The first quartiles is

$$Q_1 = 0.75 \cdot x_{n+1} + 0.25 \cdot x_{n+2} = 0.75 \cdot x_2 + 0.25 \cdot x_3 = 0.75 \cdot 15 + 0.25 \cdot 36 = 20.25.$$

The third quartile is

$$Q_3 = 0.25 \cdot x_{3n+2} + 0.75 \cdot x_{3n+3} = 0.25 \cdot x_5 + 0.75 \cdot x_6 = 0.25 \cdot 39 + 0.75 \cdot 40 = 39.75.$$

Computing methods for quartiles

Method 4: large datasets. . .

If we have an ordered large datasets $\{x_1, \ldots, x_N\}$, we can use the empirical quantile function to compute the p–th empirical quantile:

$$q(p) = x_k + \alpha\,(x_{k+1} - x_k)\,,$$

where $k = [p(N+1)]$, $[\cdot]$ denotes the integer part of a number, and $\alpha = p(N+1) - k$.

Example: Method 4 – Empirical quantiles

Let us consider the list
$\{52, 57, 57, 58, 63, 66, 66, 67, 67, 68, 69, 70, 70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 89\}$.

We have $N = 24$ data points. Then, compute median, first and third quartiles.

Median: $Q_2 = q(0.5) = x_{12} + (0.5 \cdot 25 - 12) \cdot (x_{13} - x_{12}) = 70 + (0.5 \cdot 25 - 12)(70 - 70) = 70.$

First quartile: $Q_1 = q(0.25) = x_6 + (0.25 \cdot 25 - 6) \cdot (x_7 - x_6) = 66 + (0.25 \cdot 25 - 6)(66 - 66) = 66.$

Third quartile: $Q_3 = q(0.75) = x_{18} + (0.75 \cdot 25 - 18) \cdot (x_{19} - x_{18}) = 75 + (0.75 \cdot 25 - 18)(75 - 75) = 75.$

Computing methods for quartiles

Example: Method 4 – odd number of elements

Consider the list $\{6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49\}$. We have $N = 11$.

For $p = 0.5$, we have $k = [p(N+1)] = [0.5 \cdot 12] = [6] = 6$ and $\alpha = p(N+1) - k = 0.5 \cdot 12 - 6 = 0$, then the median is

$$Q_2 = q(0.5) = x_6 + \alpha \cdot (x_7 - x_6) = 40 + 0 \cdot (41 - 40) = 40.$$

For $p = 0.25$, we have $k = [p(N+1)] = [0.25 \cdot 12] = [3] = 3$ and $\alpha = p(N+1) - k = 0.25 \cdot 12 - 3 = 0$, then the first quartile is

$$Q_1 = q(0.25) = x_3 + \alpha \cdot (x_4 - x_3) = 15 + 0 \cdot (36 - 15) = 15.$$

For $p = 0.75$, we have $k = [p(N+1)] = [0.75 \cdot 12] = [9] = 9$ and $\alpha = p(N+1) - k = 0.75 \cdot 12 - 9 = 0$, then the third quartile is

$$Q_3 = q(0.75) = x_9 + \alpha \cdot (x_{10} - x_9) = 43 + 0 \cdot (47 - 43) = 43.$$

Computing methods for quartiles

Example: Method 4 – even number of elements

Consider the list $\{7, 15, 36, 39, 40, 41\}$. We have $N = 6$.

For $p = 0.5$, we have $k = [p(N+1)] = [0.5 \cdot 7] = [3.5] = 3$ and $\alpha = p(N+1) - k = 0.5 \cdot 7 - 3 = 0.5$, then the median is

$$Q_2 = q(0.5) = x_3 + \alpha \cdot (x_4 - x_3) = 36 + 0.5 \cdot (39 - 36) = 37.5.$$

For $p = 0.25$, we have $k = [p(N+1)] = [0.25 \cdot 7] = [1.75] = 1$ and $\alpha = p(N+1) - k = 0.25 \cdot 7 - 1 = 0.75$, then the first quartile is

$$Q_1 = q(0.25) = x_1 + \alpha \cdot (x_2 - x_1) = 7 + 0.75 \cdot (15 - 7) = 13.$$

For $p = 0.75$, we have $k = [p(N+1)] = [0.75 \cdot 7] = [5.25] = 5$ and $\alpha = p(N+1) - k = 0.75 \cdot 7 - 5 = 0.25$, then the third quartile is

$$Q_3 = q(0.75) = x_5 + \alpha \cdot (x_6 - x_5) = 40 + 0.25 \cdot (41 - 40) = 40.25.$$

Computing methods for quartiles

Example: summary of computing methods for quartiles – odd number of elements

Consider the list $\{6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49\}$.

	Method 1	Method 2	Method 3	Method 4
Q_1	15	25.5	20.25	15
Q_2	40	40	40	40
Q_3	43	42.5	42.75	43

Example: summary of computing methods for quartiles – even number of elements

Consider the list $\{7, 15, 36, 39, 40, 41\}$.

	Method 1	Method 2	Method 3	Method 4
Q_1	15	15	20.25	13
Q_2	37.5	37.5	37.5	37.5
Q_3	40	40	39.75	40.25

Grouped data: approximation of the quantiles

Suppose we do not know the values of the statistical units but only the corresponding absolute frequencies f_k associated to the classes $]a_{k-1}, a_k]$, where $k = 1, \dots, M$.
For p–quantiles, a suitable choice is using a linear approximation:

$$Q_p \approx a_{j-1} + (a_j - a_{j-1}) \frac{p - F(a_{j-1})}{F(a_j) - F(a_{j-1})},$$

where the p–quantiles are in the class $]a_{j-1}, a_j]$, and F is the empirical distribution function.

Remark

- A general strategy is to assume that the data belonging to the class $]a_{j-1}, a_j]$ are equal to the central value $m_j = \dfrac{a_{j-1} + a_j}{2}$, and then proceed as usual.
- Another strategy is to divide the class $]a_{j-1}, a_j]$ in sub-classes as many as are the corresponding frequencies, and the proceed as usual.
- Contexts in which data are already grouped into classes are very rare.
- If the original data are available, the indices can be exactly computed, without needing approximations.

Example: grouped data – approximation of the quantiles

Consider the classes $]a_{k-1}, a_k]$, where $k = 1, \dots, 5$, and the corresponding absolute frequencies f_k as follows:

Class	f_k
$]0, 1]$	1
$]1, 2]$	4
$]2, 3]$	4
$]3, 4]$	2
$]4, 5]$	1

We have $N = 12$. The first quartile should be chosen between the $3-$th and $4-$th observation. Therefore, the first quartile belongs to the class $]1, 2]$, *i.e.*, $a_{j-1} = 1$ and $a_j = 2$.
Moreover, we have

$$F(1) = \frac{1}{12} = 0.09, \qquad F(2) = \frac{5}{12} = 0.42.$$

Then:

$$Q_{0.25} \approx a_{j-1} + (a_j - a_{j-1})\frac{0.25 - F(a_{j-1})}{F(a_j) - F(a_{j-1})} = 1 + (2-1)\frac{0.25 - 0.09}{0.42 - 0.09} = 1.5.$$

. . . Example: grouped data – approximation of the quantiles

Consider the classes $]a_{k-1}, a_k]$, where $k = 1, \ldots, 5$, and the corresponding absolute frequencies f_k as follows:

Class	f_k
$]0, 1]$	1
$]1, 2]$	4
$]2, 3]$	4
$]3, 4]$	2
$]4, 5]$	1

We have $N = 12$. The third quartile should be chosen between the $9-$th and $10-$th observation. Therefore, the third quartile belongs to the class $]3, 4]$, *i.e.*, $a_{j-1} = 3$ and $a_j = 4$.

Moreover, we have

$$F(3) = \frac{9}{12} = 0.75, \qquad F(4) = \frac{11}{12} = 0.92.$$

Then:

$$Q_{0.75} \approx a_{j-1} + (a_j - a_{j-1}) \frac{0.75 - F(a_{j-1})}{F(a_j) - F(a_{j-1})} = 3 + (4-3)\frac{0.75 - 0.75}{0.92 - 0.75} = 3.$$

Mean

For numerical data we can compute the arithmetic mean:

$$\mu = \mathrm{mean}(\{x_1, \dots, x_N\}) = \frac{1}{N}\sum_{k=1}^{N} x_k,$$

that is a measure of the central tendency of data.

Properties of the mean

- if $x_1 = \dots = x_N = a$, then
$$\mu = \frac{1}{N}\sum_{k=1}^{N} x_k = \frac{Na}{N} = a;$$
- $\min(\{x_1, \dots, x_N\}) \leq \mu \leq \max(\{x_1, \dots, x_N\})$;
- the sum of the differences between the data from the mean (the so-called deviations) is equal to zero:
$$\sum_{k=1}^{N}(x_k - \mu) = 0;$$
the arithmetic mean constitutes the center of mass of the distribution of frequency!

Properties of the mean

- $\forall \lambda \in \mathbb{R}$, we have
$$\sum_{k=1}^{N}(x_k-\lambda)^2=\sum_{k=1}^{N}(x_k-\mu)^2+N(\mu-\lambda)^2;$$
in fact,
$$\begin{aligned}\sum_{k=1}^{N}(x_k-\lambda)^2 &= \sum_{k=1}^{N}(x_k-\lambda+\mu-\mu)^2=\sum_{k=1}^{N}((x_k-\mu)+(\mu-\lambda))^2=\\ &=\sum_{k=1}^{N}((x_k-\mu)^2+(\mu-\lambda)^2+2(x_k-\mu)(\mu-\lambda))=\\ &=\sum_{k=1}^{N}(x_k-\mu)^2+\sum_{k=1}^{N}(\mu-\lambda)^2+2(\mu-\lambda)\sum_{k=1}^{N}(x_k-\mu)=\\ &=\sum_{k=1}^{N}(x_k-\mu)^2+N(\mu-\lambda)^2+2(\mu-\lambda)\cdot 0;\end{aligned}$$
- the sum of squares of the deviations from a constant is minimal if and only if the constant is set equal to the mean; in fact,
$$\sum_{k=1}^{N}(x_k-\lambda)^2>\sum_{k=1}^{N}(x_k-\mu)^2, \qquad \text{if } \lambda\neq\mu;$$

Properties of the mean

- the mean remains unchanged if a subset of data is replaced with theirs partial mean, *i.e.*,

$$\text{mean}(\{x_1,\dots,x_k,x_{k+1},\dots,x_N\}) = \text{mean}(\{m,\dots,m,x_{k+1},\dots,x_N\}),$$

where

$$m = \text{mean}(\{x_1,\dots,x_k\}) = \frac{1}{k}\sum_{i=1}^{k} x_i.$$

In fact,

$$\text{mean}(\{m,\dots,m,x_{k+1},\dots,x_N\}) = \frac{1}{N}\left(\underbrace{\frac{1}{k}\sum_{i=1}^{k} x_i + \cdots + \frac{1}{k}\sum_{i=1}^{k} x_i}_{k\text{ times}} + \sum_{i=k+1}^{N} x_i\right) =$$

$$= \frac{1}{N}\left(k\cdot\frac{1}{k}\sum_{i=1}^{k} x_i + \sum_{i=k+1}^{N} x_i\right) = \frac{1}{N}\left(\sum_{i=1}^{k} x_i + \sum_{i=k+1}^{N} x_i\right) = \frac{1}{N}\sum_{i=1}^{N} x_i;$$

Properties of the mean

- if we apply to the data a linear transformation, say if we define

  $$y_k = ax_k + b, \qquad a, b \text{ constants}, \quad k = 1, \ldots, N,$$

  then

  $$\overline{\mu} = \text{mean}(\{y_1, \ldots, y_N\}) = \frac{1}{N}\sum_{k=1}^{N} y_k = a\mu + b,$$

  i.e., the arithmetic mean of a linear transformation applied to data is equal to the linear transformation applied to the arithmetic mean of data;
- the mean of a non-linear transformation of data is, in general, not equal to the non-linear transformation applied to the mean; if we define $f(\cdot)$ a non-linear transformation such that

  $$y_k = f(x_k), \qquad k = 1, \ldots, N,$$

  then

  $$\text{mean}(\{y_1, \ldots, y_N\}) = \frac{1}{N}\sum_{k=1}^{N} f(x_k) \neq f\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right) = f\left(\text{mean}(\{x_1, \ldots, x_N\})\right);$$

Properties of the mean: recursive formula

Given the list $\{x_1, \ldots, x_N\}$, and let

$$\mu_k = \text{mean}(\{x_1, \ldots, x_k\}) = \frac{1}{k}\sum_{i=1}^{k} x_i$$

be the arithmetic mean of $\{x_1, \ldots, x_k\}$ with $1 \leq k < N$.
Then,

$$\mu_{k+1} = \text{mean}(\{x_1, \ldots, x_k, x_{k+1}\}) = \frac{k}{k+1}\mu_k + \frac{1}{k+1}x_{k+1}.$$

Proof.

$$\mu_{k+1} = \frac{1}{k+1}\sum_{i=1}^{k+1} x_i = \frac{k}{k+1}\frac{1}{k}\left(\sum_{i=1}^{k} x_i + x_{k+1}\right) = \frac{k}{k+1}\left(\frac{1}{k}\sum_{i=1}^{k} x_i\right) + \frac{1}{k+1}x_{k+1}.$$

Example

Let $\mu_9 = 26$ and $x_{10} = 30$. Compute μ_{10}.

$$\mu_{10} = \frac{9}{10}26 + \frac{30}{10} = 26.4$$

Grouped data: mean of a frequency distribution

Given a list of N numbers $\{x_1, \dots, x_N\}$, suppose we collect $M \leq N$ distinct values x_k $(k = 1, \dots, M)$ with the corresponding k-th absolute frequency f_k. Then, the mean is

$$\mu = \text{mean}(\{x_1, \dots, x_N\}) = \frac{\displaystyle\sum_{k=1}^{M} x_k f_k}{\displaystyle\sum_{k=1}^{M} f_k} = \frac{1}{N} \sum_{k=1}^{M} x_k f_k.$$

If we consider the relative frequencies $\widehat{f_k} = \dfrac{f_k}{N}$, we have

$$\mu = \text{mean}(\{x_1, \dots, x_N\}) = \frac{\displaystyle\sum_{k=1}^{M} x_k f_k}{\displaystyle\sum_{k=1}^{M} f_k} = \frac{1}{N} \sum_{k=1}^{M} x_k f_k = \sum_{k=1}^{M} x_k \widehat{f_k}.$$

Example 1: grouped data – mean of a frequency distribution

Consider the collected values x_k and the corresponding absolute frequencies f_k as follows:

x_k	f_k
1	2
2	7
3	1
4	10

The mean is

$$\mu = \frac{1}{N}\sum_{k=1}^{M} x_k f_k = \frac{1}{20}(1\cdot 2 + 2\cdot 7 + 3\cdot 1 + 4\cdot 10) = \frac{59}{20} = 2.95$$

Example 2: grouped data – mean of a frequency distribution

Consider the collected values x_k and the corresponding relative frequencies $\widehat{f}_k$ as follows:

x_k	$\widehat{f}_k$
-1	0.1
0	0.2
1	0.5
2	0.2

The mean is

$$\mu = \sum_{k=1}^{M} x_k \widehat{f}_k = -1 \cdot 0.1 + 0 \cdot 0.2 + 1 \cdot 0.5 + 2 \cdot 0.2 = 0.8$$

Grouped data: approximation of the mean

Suppose we do not know the values x_k associated to the statistical units but we have only the corresponding absolute frequencies f_k associated to the classes $]a_{k-1}, a_k]$, where $k = 1, \dots, M$.

The mean cannot be computed exactly!

In these situations, an approximation often used is

$$\mu = \text{mean}(\{x_1, \dots, x_N\}) \approx \frac{\displaystyle\sum_{k=1}^{M} m_k f_k}{\displaystyle\sum_{k=1}^{M} f_k} = \frac{1}{N}\sum_{k=1}^{M} m_k f_k,$$

where m_k is the middle value of the k–th class, *i.e.*, $m_k = \dfrac{a_{k-1} + a_k}{2}$.

If we consider the relative frequencies $\widehat{f_k} = \frac{f_k}{N}$, we have

$$\mu = \text{mean}(\{x_1, \dots, x_N\}) \approx \frac{1}{N}\sum_{k=1}^{M} m_k f_k = \sum_{k=1}^{M} m_k \widehat{f_k}.$$

Example: grouped data – approximation of the mean

Consider the classes $]a_{k-1}, a_k]$, where $k = 1, \dots, 4$, with the corresponding middle values m_k and absolute frequencies f_k as follows:

Class	m_k	f_k
$]0,1]$	0.5	1
$]1,2]$	1.5	4
$]2,3]$	2.5	4
$]3,4]$	3.5	2

then, the mean can be approximated as

$$\mu \approx \frac{1}{11}\sum_{k=1}^{4} m_k f_k = \frac{0.5 \cdot 1 + 1.5 \cdot 4 + 2.5 \cdot 4 + 3.5 \cdot 2}{11} \simeq 2.14.$$

Example: grouped data – approximation of the mean

Consider the classes $]a_{k-1}, a_k]$, where $k = 1, \dots, 4$, with the corresponding middle values m_k and relative frequencies $\widehat{f}_k$ as follows:

Class	m_k	$\widehat{f}_k$
$]-1, 1]$	0	0.1
$]1, 4]$	2.5	0.4
$]4, 7]$	5.5	0.4
$]7, 9]$	8	0.1

then, the mean can be approximated as

$$\mu \approx \sum_{k=1}^{4} m_k \widehat{f}_k = 0 \cdot 0.1 + 2.5 \cdot 0.4 + 5.5 \cdot 0.4 + 8 \cdot 0.1 = 4.$$

Weighted mean

Given a list of N numbers $\{x_1, x_2, \ldots, x_N\}$ with the corresponding non-negative weights $\{m_1, m_2, \ldots, m_N\}$, we can compute the weighted mean:

$$\mu(\{x_1, \ldots, x_N\}) = \frac{\displaystyle\sum_{i=1}^{N} m_i x_i}{\displaystyle\sum_{i=1}^{N} m_i}.$$

Therefore, data elements with a high weight contribute more to the weighted mean than do elements with a low weight. The weights cannot be negative.

Example: weighted mean

Compute the grade mean for university exams associated with a variable number of credits (weights).

Grade	CFU
24	8
25	6
28	8
26	12

The weighted mean is:

$$\mu(\{x_1, x_2, x_3, x_4\}) = \frac{\displaystyle\sum_{i=1}^{4} m_i x_i}{\displaystyle\sum_{i=1}^{4} m_i} = \frac{1}{34}(8 \cdot 24 + 6 \cdot 25 + 8 \cdot 28 + 12 \cdot 26) = 25.82,$$

that is different from the arithmetic mean $\mu = \dfrac{1}{4}\displaystyle\sum_{i=1}^{4} x_i = \dfrac{24 + 25 + 28 + 26}{4} = 25.75!$

Geometric mean

Given a list of N numbers $\{x_1, x_2, \ldots, x_N\}$, with $x_k > 0 \ \ \forall k = 1, \ldots, N$, we can compute the geometric mean $\mathbb{G}$:

$$\mathbb{G}(\{x_1, x_2, \ldots, x_N\}) = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N} = \left(\Pi_{k=1}^{N} x_k\right)^{\frac{1}{N}},$$

which indicates the central tendency or typical value of a set of numbers by using the product of their values.

Geometric mean: properties

- $\min(\{x_1, \ldots, x_N\}) \leq \mathbb{G} \leq \max(\{x_1, \ldots, x_N\})$.
- $\mathbb{G} \leq \mu$, where μ is the arithmetic mean.
- $\mathbb{G}(\{\lambda x_1, \ldots, \lambda x_N\}) = \lambda \mathbb{G}(\{x_1, \ldots, x_N\}) \quad \forall \lambda > 0$.
- $\mathbb{G}(\{x_1, \ldots, x_N\}) = \left(\Pi_{k=1}^{N} x_k\right)^{\frac{1}{N}} = \exp\left(\frac{1}{N} \sum_{k=1}^{N} \ln(x_k)\right)$.
- The geometric mean is mostly used where the values considered are multiplied together and not added. Typical examples are growth rates, such as interest rates or inflation rates.
- Small values (with respect to the arithmetic mean) are more influential in the geometric mean than large values.

Example: geometric mean

Suppose we measure inflation on an annual scale, and that over three years subsequent, we have respectively inflation rates of 2.5%, 2%, 1.5%, respectively. It is correct to say that the mean inflation over these three years was 2% (the arithmetic mean of the three data)?
The answer is no.
In fact, in this case, the price of a good whose initial price was p would be, after three years,

$$p \cdot (1.02) \cdot (1.02) \cdot (1.02) = p \cdot 1.061208.$$

Actually, after a year the price becomes $p_1 = p \cdot (1.025)$; at the end of the second year, the price increased by 2%, *i.e.*, $p_2 = p_1 \cdot (1.02)$. Similarly, at the end of the third year the price $p_3 = p_2 \cdot (1.015)$. Then, after three years, the price becomes

$$p \cdot (1.025) \cdot (1.02) \cdot (1.015) = p \cdot 1.0611825.$$

This result can be recovered by using the geometric mean, *i.e.*,

$$\mathbb{G}(\{x_1, x_2, x_3\}) = \sqrt[3]{1.025 \cdot 1.02 \cdot 1.015} = 1.01999183.$$

Then, the final price is given by

$$p \cdot \mathbb{G} \cdot \mathbb{G} \cdot \mathbb{G} = p \cdot 1.01999183^3 = p \cdot 1.0611825.$$

Harmonic mean

Given a list of N numbers $\{x_1, x_2, \ldots, x_N\}$, we can compute the harmonic mean $\mathbb{H}$:

$$\mathbb{H}(\{x_1, x_2, \ldots, x_N\}) = \frac{N}{\dfrac{1}{x_1} + \dfrac{1}{x_2} + \cdots + \dfrac{1}{x_N}} = \left(\frac{1}{N} \sum_{k=1}^{N} \frac{1}{x_k} \right)^{-1},$$

that is the reciprocal of the arithmetic mean of the reciprocals of the given set of observations.

Harmonic mean: properties

- $\min(\{x_1, \ldots, x_N\}) \leq \mathbb{H} \leq \max(\{x_1, \ldots, x_N\})$.
- If $x_k > 0 \quad \forall k = 1, \ldots, N$, then

  $$\mathbb{H} \leq \mathbb{G} \leq \mu,$$

  where $\mathbb{G}$ and μ are the geometric and arithmetic mean, respectively.
- $\mathbb{H}(\{\lambda x_1, \ldots, \lambda x_N\}) = \lambda \mathbb{H}(\{x_1, \ldots, x_N\}) \quad \forall \lambda \in \mathbb{R}$.
- The harmonic mean is sometimes appropriate for situations when an average rate is desired.
- Since the harmonic mean of a list of numbers tends strongly toward the least elements of the list, it tends (compared to the arithmetic mean) to mitigate the impact of large outliers and aggravate the impact of small ones.

Example: harmonic mean

Suppose we have two cars that drive for 10 *km* and 20 *km*, respectively, with one liter of fuel. It may seems that the mean distance is 15 Km/L (the arithmetic mean), but does it make sense? Suppose we have to travel 100 *km* with both cars. By relying on the mean distance as has been defined, we need

$$2 \cdot \frac{\text{number of Kilometers}}{\text{mean distance}} = 2 \cdot \frac{100}{15}\,\text{Liters} = 13.333\,\text{Liters}.$$

Obviously wrong answer, since we need 10 L for the first car and 5 L for the second, with a total of 15 L. So the only one knowledge of the mean distance defined by the arithmetic mean, leads us to make mistakes. Now, consider the mean distance by means of the harmonic mean:

$$\mathbb{H}(\{x_1, x_2\}) = \left(\frac{1}{2}\sum_{k=1}^{2}\frac{1}{x_k}\right)^{-1} = \left(\frac{\frac{1}{10}+\frac{1}{20}}{2}\right)^{-1} = 13.333\ Km/L.$$

Therefore, in order to travel 100 Km with the two cars, we need

$$2 \cdot \frac{\text{number of Kilometers}}{\text{mean distance}} = 2 \cdot \frac{100}{13.333}\,\text{Liters} = 15\,\text{Liters}.$$

Chisini mean

Given a list of N numbers $\{x_1, \ldots, x_N\}$ and a function $f(x_1, \ldots, x_N)$ depending on the N variables entering the list, the Chisini mean, or the mean of numbers x_i with respect to f, can be defined as the unique $\mathbb{M}$, if there exists, such that

$$f(x_1, \ldots, x_N) = f(\mathbb{M}, \ldots, \mathbb{M}).$$

Properties

- The Chisini mean is that value that does not alter the value of the function f when substituting the constant value $\mathbb{M}$ to the variables x_k $(k = 1, \ldots, N)$.
- The arithmetic, geometric and harmonic means are all Chisini means.

Example: the arithmetic mean is a Chisini mean

If we define the function f as

$$f(x_1, x_2, \dots, x_N) = x_1 + x_2 + \cdots + x_N,$$

the mean of numbers x_i with respect to f will be the arithmetic mean!
In fact, from

$$f(x_1, \dots, x_N) = f(\mathbb{M}, \dots, \mathbb{M}),$$

we have

$$x_1 + x_2 + \cdots + x_N = \underbrace{\mathbb{M} + \mathbb{M} + \cdots + \mathbb{M}}_{N \text{ times}} = N\mathbb{M},$$

and it follows that

$$\mathbb{M} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

Example: the geometric mean is a Chisini mean

If we define the function f as

$$f(x_1, x_2, \dots, x_N) = x_1 \cdot x_2 \cdot \dots \cdot x_N,$$

with $x_k > 0 \ \forall k = 1, \dots, N$, the mean of numbers x_i with respect to f will be the geometric mean! In fact, from

$$f(x_1, \dots, x_N) = f(\mathbb{M}, \dots, \mathbb{M}),$$

we have

$$x_1 \cdot x_2 \cdot \dots \cdot x_N = \underbrace{\mathbb{M} \cdot \mathbb{M} \cdot \dots \cdot \mathbb{M}}_{N \text{ times}} = \mathbb{M}^N,$$

and it follows that

$$\mathbb{M} = \sqrt[N]{x_1 \cdot x_2 \cdot \dots \cdot x_N}.$$

Example: the harmonic mean is a Chisini mean

If we define the function f as

$$f(x_1, x_2, \ldots, x_N) = \frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N} = \sum_{k=1}^{N} \frac{1}{x_k},$$

the mean of numbers x_i with respect to f will be the harmonic mean!
In fact, from

$$f(x_1, \ldots, x_N) = f(\mathbb{M}, \ldots, \mathbb{M}),$$

we have

$$\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N} = \underbrace{\frac{1}{\mathbb{M}} + \frac{1}{\mathbb{M}} + \cdots + \frac{1}{\mathbb{M}}}_{N \text{ times}} = \frac{N}{\mathbb{M}},$$

and it follows that

$$\mathbb{M} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}} = \left(\frac{1}{N} \sum_{k=1}^{N} \frac{1}{x_k} \right)^{-1}.$$

The mean is not a robust indicator

Consider the data $\{4, 4.9, 5, 5.3, 5.7, 6\}$; their mean is $\mu = 5.15$.
If we add another value to the data, for instance 3.2, or 8, the mean changes:

$$\begin{aligned} \text{mean}(\{3.2, 4, 4.9, 5, 5.3, 5.7, 6\}) &= 4.871, \\ \text{mean}(\{4, 4.9, 5, 5.3, 5.7, 6, 8\}) &= 5.557. \end{aligned}$$

The mean alone is not sufficient!

The mean of a set of data is a useful descriptor but cannot be used as the only synthetic descriptor. In fact, if you consider the following two sets of data

$$\{3, 3, 3, 3, 3, 3, 3\}, \qquad \text{and} \qquad \{2.6, 2.7, 2.8, 3, 3.2, 3.3, 3.4\},$$

they share the same mean,

$$\begin{aligned} \text{mean}(\{3, 3, 3, 3, 3, 3, 3\}) &= 3, \\ \text{mean}(\{2.6, 2.7, 2.8, 3, 3.2, 3.3, 3.4\}) &= 3; \end{aligned}$$

each element of the first set of data is equal to the mean, whereas the elements of the second set of data are distributed around the mean.

Dispersion of data around the mean

In addition to the mean we need a descriptor for the dispersion of the data around the mean.

Variance

Given a list of N numbers $\{x_1, x_2, \dots x_N\}$ with mean μ, their variance, denoted by σ^2, is

$$\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{k=1}^{N}(x_k-\mu)^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k^2+\mu^2-2\mu x_k) = \\
&= \frac{1}{N}\sum_{k=1}^{N}x_k^2 + \frac{1}{N}\sum_{k=1}^{N}\mu^2 - \frac{2\mu}{N}\sum_{k=1}^{N}x_k = \\
&= \frac{1}{N}\sum_{k=1}^{N}x_k^2 + \frac{N\mu^2}{N} - \frac{2\mu}{N}\sum_{k=1}^{N}x_k = \frac{1}{N}\sum_{k=1}^{N}x_k^2 + \mu^2 - 2\mu^2 = \\
&= \frac{1}{N}\sum_{k=1}^{N}x_k^2 - \mu^2.
\end{aligned}$$

Therefore, the variance is the difference between the arithmetic mean of the squared data and the square of the arithmetic mean.

Example: variance

Let us consider the list $\{1,2,3,5\}$. Compute the variance.
We have $N=4$, and

$$\sigma^2 = \text{variance}(\{1,2,3,5\}) = \frac{1}{4}\left(\sum_{k=1}^{4} x_k^2\right) - \mu^2.$$

The arithmetic mean is

$$\mu = \text{mean}(\{1,2,3,5\}) = \frac{1}{4}\sum_{k=1}^{4} x_k = \frac{1+2+3+5}{4} = 2.75.$$

The arithmetic mean of the squared data is

$$\frac{1}{4}\sum_{k=1}^{4} x_k^2 = \frac{1+4+9+25}{4} = 9.75.$$

Then, the variance is

$$\sigma^2 = 9.75 - 2.75^2 = 9.75 - 7.56 = 2.19.$$

Unbiased sample variance

Using probability, a more correct definition of variance when the data refer to a sample is

$$\widetilde{\sigma^2} = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - \mu)^2,$$

that is called unbiased sample variance, and the use of the term $N - 1$ is called Bessel's correction.

Rough justification

Since we use in the definition of the variance the mean, the N data are not independent (that is, the data have $N - 1$ degrees of freedom).
However, there is a rigorous justification based on the necessity of inferring the variance of a population from the variation of a random sample extracted from the population. In fact, in many practical situations, the true variance of a population is not known a priori and must be computed somehow. When dealing with extremely large populations, it is not possible to count every object in the population, so the computation must be performed on a sample of the population. Sample variance can also be applied to the estimation of the variance of a continuous distribution from a sample of that distribution.

Correct estimation of variance

$$\begin{aligned}
\widetilde{\sigma^2} &= \frac{1}{N-1}\sum_{k=1}^{N}(x_k-\mu)^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_k^2+\mu^2-2\mu x_k) = \\
&= \frac{1}{N-1}\sum_{k=1}^{N}x_k^2 + \frac{1}{N-1}\sum_{k=1}^{N}\mu^2 - \frac{2\mu}{N-1}\sum_{k=1}^{N}x_k = \\
&= \frac{1}{N-1}\sum_{k=1}^{N}x_k^2 + \frac{N}{N-1}\mu^2 - \frac{2N}{N-1}\mu^2 = \\
&= \frac{N}{N-1}\left(\frac{1}{N}\sum_{k=1}^{N}x_k^2\right) - \frac{N}{N-1}\mu^2 = \\
&= \frac{N}{N-1}\left(\frac{1}{N}\sum_{k=1}^{N}x_k^2-\mu^2\right) = \frac{N}{N-1}\sigma^2.
\end{aligned}$$

σ^2 gives an estimate of the population variance that is biased by a factor of $\dfrac{N-1}{N}$. For this reason, σ^2 is referred to as the biased sample variance.

Properties of the variance

Given a list of N numbers $\{x_1, \ldots, x_N\}$, then:

- variance$(\{x_1, \ldots, x_N\}) \geq 0$;
- variance$(\{x_1, \ldots, x_N\}) = 0$ if and only if $x_1 = \ldots = x_N = \mu$.

Example: variance

Variance is useful for comparing two data sets having the same mean.
For instance, for the two sets of data having the same mean 3, it is:

$$\text{variance}(\{3, 3, 3, 3, 3, 3, 3\}) = 0,$$
$$\text{variance}(\{2.6, 2.7, 2.8, 3, 3.2, 3.3, 3.4\}) = 0.097.$$

Remark

The more the data deviate from the mean, the greater the variance!

Properties of the variance

If we apply to the data a linear transformation, say if we define

$$y_k = ax_k + b, \qquad a, b \text{ constants}, \; k = 1, \dots, N,$$

then

$$\overline{\sigma^2} = \text{variance}(\{y_1, \dots, y_N\}) = a^2 \text{variance}(\{x_1, \dots, x_N\}) = a^2\sigma^2.$$

In fact, we know that

$$\overline{\mu} = \text{mean}(\{y_1, \dots, y_N\}) = a\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right) + b = a\mu + b,$$

whence,

$$\overline{\sigma^2} = \text{variance}(\{y_1, \dots, y_N\}) = \frac{1}{N}\sum_{k=1}^{N}(y_k - \overline{\mu})^2 =$$

$$= \frac{1}{N}\sum_{k=1}^{N}(ax_k + b - a\mu - b)^2 = \frac{1}{N}\sum_{k=1}^{N}(a(x_k - \mu))^2 =$$

$$= \frac{1}{N}\sum_{k=1}^{N} a^2(x_k - \mu)^2 = a^2\left(\frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2\right) = a^2\sigma^2.$$

Example: variance

Given the list $\{1,2,3,5\}$, the variance is

$$\sigma^2 = \text{variance}(\{1,2,3,5\}) = \frac{1}{4}\left(\sum_{k=1}^{4} x_k^2\right) - \mu^2 = \frac{1}{4}\left(\sum_{k=1}^{4} x_k^2\right) - \left(\frac{1}{4}\sum_{k=1}^{4} x_k\right)^2 =$$

$$= \frac{1+4+9+25}{4} - \left(\frac{1+2+3+5}{4}\right)^2 = 9.75 - 2.75^2 = 9.75 - 7.56 = 2.19.$$

If we apply to the list $\{1,2,3,5\}$ the linear transformation

$$y_k = 2x_k + 1, \qquad k = 1, \dots, 4,$$

we have

$$\overline{\sigma^2} = \text{variance}(\{3,5,7,11\}) = 2^2 \text{variance}(\{1,2,3,5\}) = 2^2 \cdot \sigma^2 = 2^2 \cdot 2.19 = 8.75.$$

The correct estimation of the variance is

$$\widetilde{\sigma^2} = \frac{4}{3}\overline{\sigma^2} = \frac{4}{3}8.75 = 11.67.$$

Properties of the variance: recursive formula

Suppose we know the arithmetic mean μ_N and the variance σ_N^2 of a list of N numbers $\{x_1, \dots, x_N\}$. Adding a new number x_{N+1} to the list, the variance σ_{N+1}^2 of $N+1$ numbers can be computed as

$$\sigma_{N+1}^2 = \text{variance}(\{x_1, \dots, x_N, x_{N+1}\}) = \frac{N}{N+1}(\sigma_N^2 + \mu_N^2) + \frac{1}{N+1}x_{N+1}^2 - \mu_{N+1}^2.$$

Firstly, recall that we are able to compute the arithmetic mean μ_{N+1} of $N+1$ numbers by means of the recursive formula

$$\mu_{N+1} = \text{mean}(\{x_1, \dots, x_N, x_{N+1}\}) = \frac{N}{N+1}\mu_N + \frac{1}{N+1}x_{N+1}.$$

Then, from

$$\sigma_N^2 = \frac{1}{N}\sum_{k=1}^{N} x_k^2 - \mu_N^2,$$

we have

$$\sum_{k=1}^{N} x_k^2 = N(\sigma_N^2 + \mu_N^2) \Longrightarrow \sum_{k=1}^{N+1} x_k^2 = N(\sigma_N^2 + \mu_N^2) + x_{N+1}^2.$$

Finally,

$$\sigma_{N+1}^2 = \frac{1}{N+1}\sum_{k=1}^{N+1} x_k^2 - \mu_{N+1}^2 = \frac{1}{N+1}\left(N(\sigma_N^2 + \mu_N^2) + x_{N+1}^2\right) - \mu_{N+1}^2.$$

Example: recursive formula for variance

Let us consider the list $\{1,2,3,5\}$. We have that

$$\mu_4 = \text{mean}(\{1,2,3,5\}) = \frac{1}{4}\sum_{k=1}^{4} x_k = \frac{1+2+3+5}{4} = 2.75.$$

and

$$\sigma_4^2 = 9.75 - 2.75^2 = 9.75 - 7.56 = 2.19.$$

Add the number 7, so we have the new list $\{1,2,3,5,7\}$. We have

$$\mu_5 = \text{mean}(\{1,2,3,5,7\}) = \frac{4}{5}\mu_4 + \frac{1}{5}x_5 = \frac{4}{5}2.75 + \frac{7}{5} = 3.6.$$

and

$$\sigma_5^2 = \text{variance}(\{1,2,3,5,7\}) = \frac{4}{5}(\sigma_4^2 + \mu_4^2) + \frac{1}{5}x_5^2 - \mu_5^2 = \frac{4}{5}(2.19 + 2.75^2) + \frac{7^2}{5} - 3.6^2 = 4.642$$

The correct estimation of the variance is

$$\widetilde{\sigma^2}_5 = \frac{5}{4}\sigma_5^2 = 5.8.$$

More on variance

The variance is related to the mean value of the squared differences between all possible couples of data. In fact, it is:

$$
\begin{aligned}
&\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)^2 = \\
&\quad = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - \mu + \mu - x_j)^2 = \\
&\quad = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - \mu)^2 + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_j - \mu)^2 - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - \mu)(x_j - \mu) = \\
&\quad = \frac{N}{N^2}\sum_{i=1}^{N}(x_i - \mu)^2 + \frac{N}{N^2}\sum_{j=1}^{N}(x_j - \mu)^2 - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - \mu)(x_j - \mu) = \\
&\quad = \frac{2N}{N^2}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{2}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = 2\sigma^2.
\end{aligned}
$$

Grouped data: approximation of the variance

Suppose we do not know the values x_k associated to the statistical units but we have only the corresponding absolute frequencies f_k associated to the classes $]a_{k-1}, a_k]$, where $k = 1, \dots, M$.
In these situations, an approximation often used is

$$\sigma^2 = \text{variance}(\{x_1, \dots, x_N\}) \approx \frac{1}{N} \sum_{k=1}^{M} (m_k - \mu)^2 f_k,$$

where μ is the approximation of the mean

$$\mu \approx \frac{1}{N} \sum_{k=1}^{M} m_k f_k,$$

and m_k is the middle value of the k–th class, *i.e.*, $m_k = \dfrac{a_{k-1} + a_k}{2}$.
If we consider the relative frequencies $\widehat{f_k} = \frac{f_k}{N}$, we have

$$\sigma^2 = \text{variance}(\{x_1, \dots, x_N\}) \approx \sum_{k=1}^{M} (m_k - \mu)^2 \widehat{f_k}.$$

Example: grouped data – approximation of the variance

Consider the classes $]a_{k-1}, a_k]$, where $k = 1, \dots, 4$, with the corresponding middle values m_k and absolute frequencies f_k as follows:

Class	m_k	f_k
$]0,1]$	0.5	1
$]1,2]$	1.5	4
$]2,3]$	2.5	4
$]3,4]$	3.5	2

We have recovered the approximated mean as $\mu \approx \frac{1}{11}\sum_{k=1}^{4} m_k f_k = \frac{0.5\cdot 1+1.5\cdot 4+2.5\cdot 4+3.5\cdot 2}{11} \simeq 2.14$, and the approximation of variance is

$$\sigma^2 \approx \frac{1}{11}\sum_{k=1}^{4}(m_k-\mu)^2 f_k =$$
$$= \frac{(0.5-2.14)^2\cdot 1+(1.5-2.14)^2\cdot 4+(2.5-2.14)^2\cdot 4+(3.5-2.14)^2\cdot 2}{11} \simeq 0.78.$$

Example: grouped data – approximation of the variance

Consider the classes $]a_{k-1}, a_k]$, where $k = 1, \ldots, 4$, with the corresponding middle values m_k and relative frequencies $\widehat{f}_k$ as follows:

Class	m_k	$\widehat{f}_k$
$]-1, 1]$	0	0.1
$]1, 4]$	2.5	0.4
$]4, 7]$	5.5	0.4
$]7, 9]$	8	0.1

We have recovered the approximated mean as
$\mu \approx \sum_{k=1}^{4} m_k \widehat{f}_k = 0 \cdot 0.1 + 2.5 \cdot 0.4 + 5.5 \cdot 0.4 + 8 \cdot 0.1 = 4$, and the approximation of variance is

$$\sigma^2 \approx \sum_{k=1}^{4} (m_k - \mu)^2 \widehat{f}_k =$$
$$= (0-4)^2 \cdot 0.1 + (2.5-4)^2 \cdot 0.4 + (5.5-4)^2 \cdot 0.4 + (8-4)^2 \cdot 0.1 \simeq 5.$$

Standard deviation

The square root of variance is the standard deviation, denoted by σ.

Variance and standard deviation

Variance and standard deviation are called indexes of dispersion or indexes of variability, because they measure the dispersion of the data around the mean.

The variance and standard deviation values, because they measure the absolute variation in a data set, depend on the unit of measurement of the data.

In particular:

- the unit of measurement of variance is equal to the square of the unit of measurement of the data;
- the unit of measurement of the standard deviation is equal to the unit of measurement of the data.

Other synthetic descriptors: mode

Given a list of N data $\{x_1, x_2, \ldots, x_N\}$, the mode is the value x_k, if any, having the greatest absolute frequency. When a set of data has a mode, we call unimodal the distribution of data. When more modes exist, we have bimodal or in general multimodal distributions.

About mean, median and mode...

- Mode is mostly used with qualitative data, for which it is not possible to compute mean and median.
- Mode is not useful when the data are many and mostly different from each other; in such cases mode may not exist or be away from the center of the data set. For this reason, this descriptor is rarely used.
- Mean, median and mode are called indexes of position or indexes of central tendency, because they describe around which value the dataset is centered.
- The median is preferable to the mean when you want to eliminate the effects of extreme values very different from the other data: the reason is that the median does not use all the data, but only the central data or the two central data; however, using only the central data makes the median insensitive to all other data values and this may be a limitation of this index.

Example 1: mode

The dataset

$$\{3, 3, 5, 4, 7, 7, 7, 9, 2, 1\}$$

has mode $x = 7$.

Example 2: mode

The dataset

$$\{3, 3, 3, 5, 4, 7, 7, 7, 9, 2, 1\}$$

has two modes: $x = 3$ and $x = 7$.

Example 3: mode

The dataset

$$\{3, 5, 4, 7, 8, 6, 9, 2, 1\}$$

has no mode, because each data occurs only once.

Other synthetic descriptors

1. Coefficient of variation (CV):

$$\mathrm{CV} = \frac{\sigma}{|\mu|},$$

that is a measure of variability (the differences) between the observed data (measured also with different units of measurement). It shows the percentage of the standard deviation in relation to the arithmetic mean, and it is a dimensionless parameter, since the mean and standard deviation are expressed in the same unit of measurement.
The variability may depend on the level of the phenomena considered!

2. Range of variation:

$$\mathrm{range} = \max(\{x_1, \ldots, x_N\}) - \min(\{x_1, \ldots, x_N\});$$

it is quick to calculate but too much sensitive to possible outliers.

3. MAD (Median Absolute Deviations):

$$\mathrm{MAD} = \mathrm{median}(\{|x_1 - \widehat{x}|, \ldots, |x_N - \widehat{x}|\}),$$

where $\widehat{x}$ is the median of $\{x_1, \ldots, x_N\}$.
This descriptor is robust against the presence of anomalous values.

Example: coefficient of variation

Given a sample of 200 parcels whose weight and volume are known. By calculating the mean and the standard deviation of the two measurements, the following values are obtained:

$$\begin{aligned} &\text{mean weight:}\quad \overline{x}_w = 9\,Kg, && \text{standard deviation weight:}\quad \sigma_w = 1.5\,Kg,\\ &\text{mean volume:}\quad \overline{x}_v = 2.7\,m^3, && \text{standard deviation volume:}\quad \sigma_v = 0.6\,m^3. \end{aligned}$$

Let us compare the variability of weight and volume.
Since weight and volume are expressed in different units of measurement, it is necessary to take into account the relative variability of the observations by computing the coefficient of variation.
For the weight, the coefficient of variation is

$$\mathrm{CV} = \frac{\sigma}{|\mu|} = \frac{1.5}{9} = 0.1667.$$

For the volume, the coefficient of variation is

$$\mathrm{CV} = \frac{\sigma}{|\mu|} = \frac{0.6}{2.7} = 0.2222.$$

Therefore, with respect to the mean, the volume of parcels has more variability than the weight.

Standardization of data

The properties of mean and variance with respect to a linear transformation of data are useful in the process of standardization of data. In this way, data coming from different contexts (and measured with different scales) can be compared.

Let $\{x_1, \ldots, x_N\}$ a set of N numbers, with mean μ and variance σ^2.
The standardization of the data is obtained by computing the data $\{z_1, \ldots, z_N\}$ so defined:

$$z_k = \frac{x_k - \mu}{\sigma}, \qquad k = 1, \ldots, N.$$

It is immediate to verify that

$$\overline{\mu} = \text{mean}(\{z_1, \ldots, z_N\}) = \frac{1}{N}\sum_{k=1}^{N} z_k = 0,$$

and

$$\overline{\sigma^2} = \text{variance}(\{z_1, \ldots, z_N\}) = \frac{1}{N-1}\sum_{k=1}^{N}(z_k - \overline{\mu})^2 = 1.$$

Example: standardization of data

Let us consider the list $\mathbf{x} = \{1, 4, 6, 10, 11\}$. We have:

$$\mu = \text{mean}(\{1, 4, 6, 10, 11\}) = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1+4+6+10+11}{5} = 6.4,$$

$$\sigma = \sqrt{\frac{1}{4}\sum_{i=1}^{5}(x_i-\mu)^2} =$$

$$= \sqrt{\frac{1}{4}\left((1-6.4)^2 + (4-6.4)^2 + (6-6.4)^2 + (10-6.4)^2 + (11-6.4)^2\right)} = 4.15933.$$

The new standardized variables $\mathbf{z} = \{z_1, \ldots, z_5\}$ such that

$$z_k = \frac{x_k - \mu}{\sigma}, \qquad k = 1, \ldots, 5$$

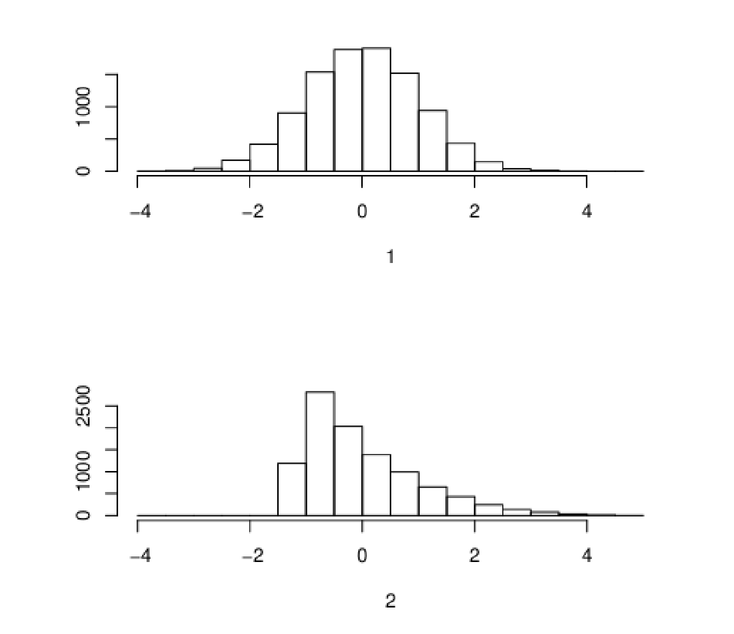
are

$$\mathbf{z} = \left\{\frac{1-6.4}{4.15933}, \frac{4-6.4}{4.15933}, \frac{6-6.4}{4.15933}, \frac{10-6.4}{4.15933}, \frac{11-6.4}{4.15933}\right\},$$

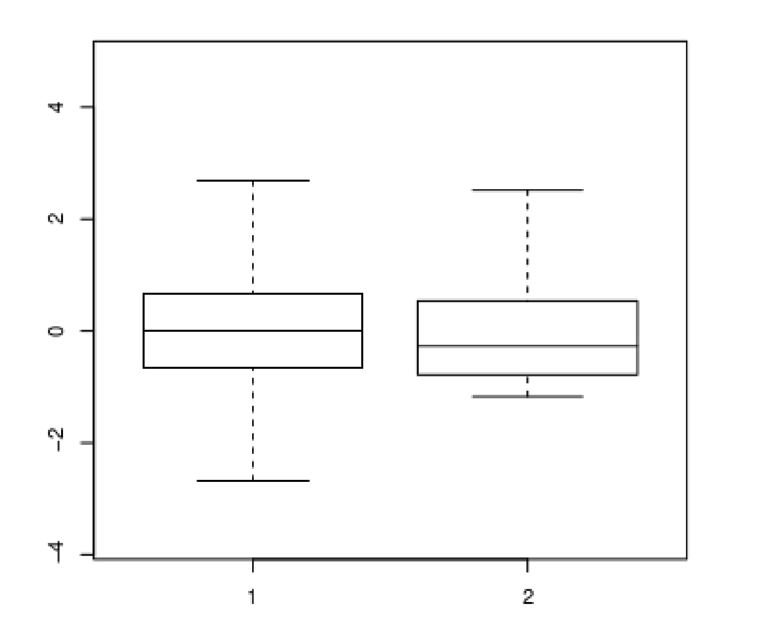
i.e.,

$$\mathbf{z} = \{-1.29829, -0.577016, -0.0961694, 0.865525, 1.10595\}.$$

Standardized datasets: histograms



Standardized datasets: box and whiskers plot



Symmetry of distribution

Another characteristic of the data we consider is the symmetry of their distribution. If the tail towards high values is much more pronounced than the tail towards low values (tail on the right), the distribution is said to be positive skewness. In the opposite case (left tail more pronounced, or tail on the left) would be called negative skewness.

Other synthetic descriptors: Skewness

A measure of the symmetry (or the lack of symmetry) of the distribution of data around the mean is given by the skewness:

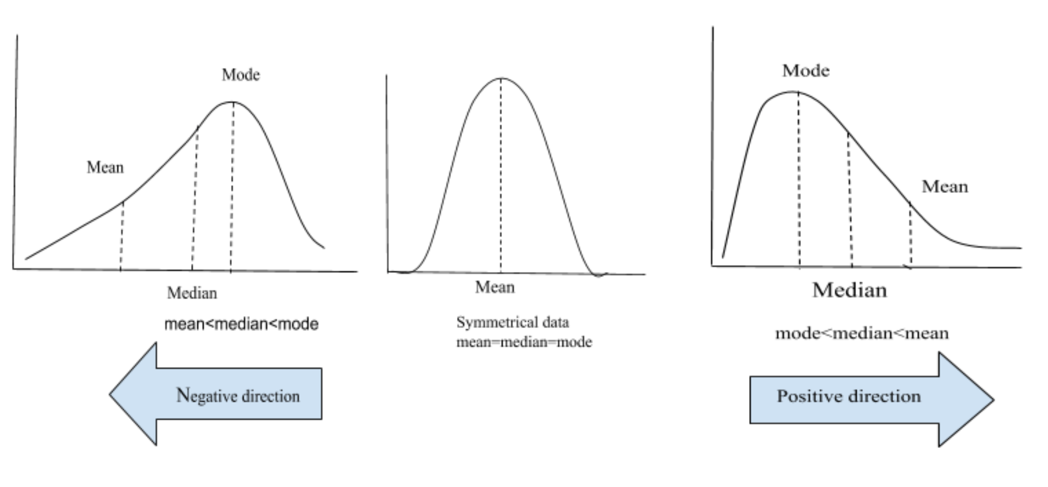
$$\text{skewness} = \frac{1}{N\sigma^3}\sum_{k=1}^{N}(x_k - \mu)^3.$$

We have a positive skewness if the data greater than the mean prevail against data less than the mean, negative skewness in the opposite case. If the data are distributed symmetrically around the mean, the positive terms and negatives in the summation will compensate each other and therefore we will have skewness equal to zero.

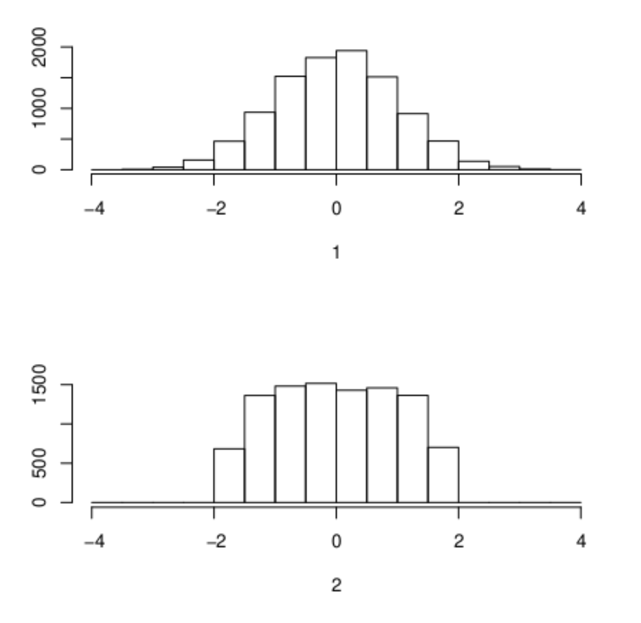
Skewness is detectable when the median does not coincide with the mean; in fact, if the mean is greater than the median, the distribution has positive skewness, whereas if the mean is smaller than the median, the distribution has negative skewness.

Skewness is invariant with respect to linear transformations of data!

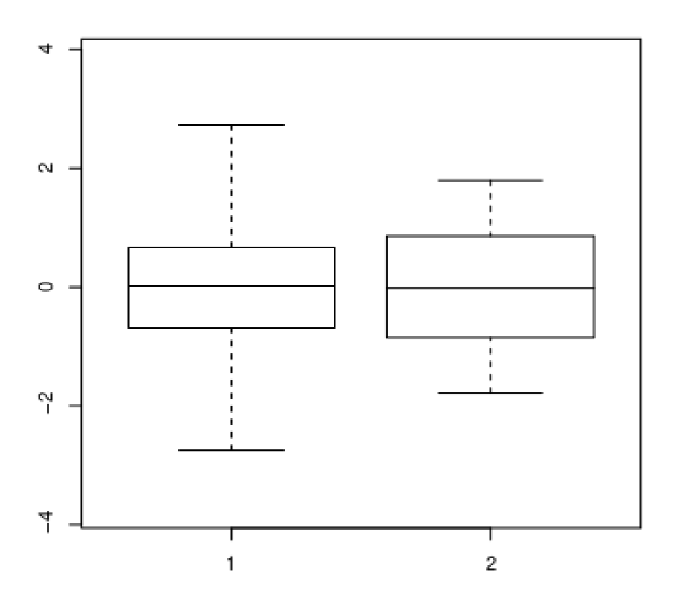
Other synthetic descriptors: Skewness



Standardized datasets: histograms



Standardized datasets: box and whiskers plot



Other synthetic descriptors: Kurtosis

The degree of “flattening” of a distribution with the same variability (skewness equal to zero) is measured by the kurtosis:

$$\text{kurtosis} = \frac{1}{N\sigma^4}\sum_{k=1}^{N}(x_k-\mu)^4.$$

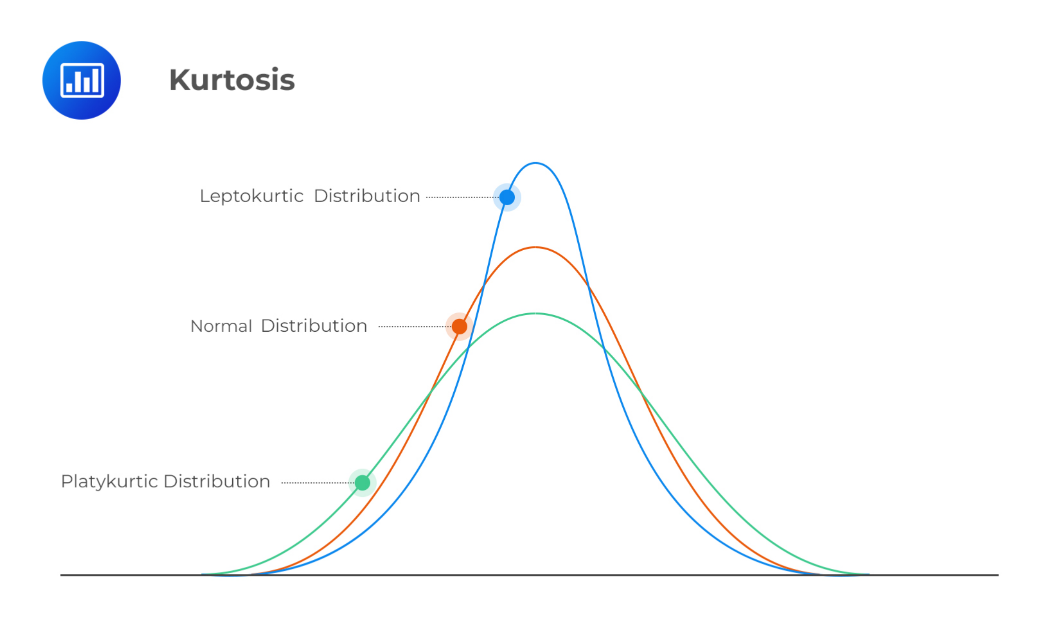
It is a measure of whether the data are heavy-tailed or light-tailed with respect to the normal distribution of data around the mean. There are three types of kurtosis:

- mesokurtic: distributions that are moderate in breadth and curves with a medium peaked height;
- leptokurtic: more values in the distribution tails and more values close to the mean (*i.e.*, sharply peaked with heavy tails);
- platykurtic: fewer values in the tails and fewer values close to the mean (*i.e.*, the curve has a flat peak and has more dispersed scores with lighter tails).

Remark

- kurtosis ≥ 0 and kurtosis $= 0$ only if the data are constant.
- Kurtosis can be seen as a relationship between two indices of variability. The index at numerator is chosen to be more sensitive to the presence of heavy tails with respect to the index at denominator.
- For convention, kurtosis $= 3$ denotes mesokurtic distributions.
- Kurtosis is invariant with respect to linear transformations of data.

Other synthetic descriptors: Kurtosis



Kurtosis

Sinthetic descriptors

1st Work Organization

$$\begin{aligned}
&\text{1st Quartile} = 699,\\
&\text{median} = 706,\\
&\text{3rd Quartile} = 713,\\
&\text{mean} = 705.472,\\
&\text{variance} = 99.8668,\\
&\text{skewness} = -0.193553,\\
&\text{kurtosis} = 2.69996.
\end{aligned}$$

2nd Work Organization

$$\begin{aligned}
&\text{1st Quartile} = 688,\\
&\text{median} = 699,\\
&\text{3rd Quartile} = 712,\\
&\text{mean} = 700.781,\\
&\text{variance} = 276.025,\\
&\text{skewness} = 0.452882,\\
&\text{kurtosis} = 2.95124.
\end{aligned}$$

3rd Work Organization

$$\begin{aligned}
&\text{1st Quartile} = 707,\\
&\text{median} = 718.5,\\
&\text{3rd Quartile} = 730,\\
&\text{mean} = 719.177,\\
&\text{variance} = 247.54,\\
&\text{skewness} = 0.100962,\\
&\text{kurtosis} = 2.60925.
\end{aligned}$$

Covariance

Given two lists (or variables) of N numbers, say $\mathbf{x} = \{x_1, \ldots, x_N\}$ and $\mathbf{y} = \{y_1, \ldots, y_N\}$, with the corresponding arithmetic mean $\overline{x} = \text{mean}(\{x_1, \ldots, x_N\})$ and $\overline{y} = \text{mean}(\{y_1, \ldots, y_N\})$, their covariance is

$$\begin{aligned} \text{cov}(\mathbf{x}, \mathbf{y}) &= \frac{1}{N}\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y}) = \frac{1}{N}\sum_{k=1}^{N} x_k(y_k - \overline{y}) - \frac{\overline{x}}{N}\sum_{k=1}^{N}(y_k - \overline{y}) = \\ &= \frac{1}{N}\sum_{k=1}^{N} x_k y_k - \frac{\overline{y}}{N}\sum_{k=1}^{N} x_k = \frac{1}{N}\sum_{k=1}^{N} x_k y_k - \overline{x}\,\overline{y} = \overline{xy} - \overline{x}\,\overline{y}, \end{aligned}$$

being $\overline{xy} = \text{mean}(\{x_1 y_1, \ldots, x_N y_N\})$, that measures how much the two variables vary together.
The sign of the covariance shows the tendency in the linear relationship between the variables:

- if $\text{cov}(\mathbf{x}, \mathbf{y}) > 0 \Rightarrow$ the data $\mathbf{y}$ increases when the data $\mathbf{x}$ increases, *i.e.*, there is a positive linear relationship among the data;
- if $\text{cov}(\mathbf{x}, \mathbf{y}) < 0 \Rightarrow$ the data $\mathbf{y}$ decreases when the data $\mathbf{x}$ increases, *i.e.*, there is a negative linear relationship among the data;
- if $\text{cov}(\mathbf{x}, \mathbf{y}) = 0 \Rightarrow$ there is no linear relationship between the the data.

Covariance: properties

- $\mathrm{cov}(\mathbf{x},\mathbf{y}) = \mathrm{cov}(\mathbf{y},\mathbf{x})$.
  In fact:
  $$\mathrm{cov}(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{k=1}^{N}(x_k-\overline{x})(y_k-\overline{y}) = \frac{1}{N}\sum_{k=1}^{N}(y_k-\overline{y})(x_k-\overline{x}) = \mathrm{cov}(\mathbf{y},\mathbf{x}).$$
- $\mathrm{cov}(\mathbf{x},\mathbf{x}) = \sigma^2(\mathbf{x}) \geq 0$.
  In fact:
  $$\mathrm{cov}(\mathbf{x},\mathbf{x}) = \frac{1}{N}\sum_{k=1}^{N}(x_k-\overline{x})(x_k-\overline{x}) = \frac{1}{N}\sum_{k=1}^{N}(x_k-\overline{x})^2 = \sigma^2(\mathbf{x}) \geq 0.$$
- $\mathrm{cov}(\mathbf{x},-\mathbf{x}) = -\sigma^2(\mathbf{x}) \leq 0$.
  In fact:
  $$\mathrm{cov}(\mathbf{x},-\mathbf{x}) = \frac{1}{N}\sum_{k=1}^{N}(x_k-\overline{x})(-x_k+\overline{x}) = -\frac{1}{N}\sum_{k=1}^{N}(x_k-\overline{x})^2 = -\sigma^2(\mathbf{x}) \leq 0.$$

Covariance: properties

If we apply a linear transformation to the variables $\mathbf{x} = \{x_1, \ldots, x_N\}$ and $\mathbf{y} = \{y_1, \ldots, y_N\}$, say

$$v_k = ax_k + b, \qquad w_k = cy_k + d \qquad a, b, c, d \in \mathbb{R}, \quad k = 1, \ldots, N,$$

then

$$\mathrm{cov}(\mathbf{v}, \mathbf{w}) = ac\,\mathrm{cov}(\mathbf{x}, \mathbf{y}),$$

where $\mathbf{v} = \{v_1, \ldots, v_N\}$ and $\mathbf{w} = \{w_1, \ldots, w_N\}$. In fact, we know that

$$\overline{v} = \mathrm{mean}(\{v_1, \ldots, v_N\}) = a\left(\frac{1}{N}\sum_{k=1}^{N} x_k\right) + b = a\overline{x} + b,$$

$$\overline{w} = \mathrm{mean}(\{w_1, \ldots, w_N\}) = c\left(\frac{1}{N}\sum_{k=1}^{N} y_k\right) + d = c\overline{y} + d,$$

whence,

$$\mathrm{cov}(\mathbf{v}, \mathbf{w}) = \frac{1}{N}\sum_{k=1}^{N}(v_k - \overline{v})(w_k - \overline{w}) = \frac{1}{N}\sum_{k=1}^{N}(ax_k + b - a\overline{x} - b)(cy_k + d - c\overline{y} - d) =$$

$$= ac\frac{1}{N}\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y}) = ac\,\mathrm{cov}(\mathbf{x}, \mathbf{y}).$$

Covariance: properties

If we apply a transformation to the variables $\mathbf{x} = \{x_1, \ldots, x_N\}$, $\mathbf{y} = \{y_1, \ldots, y_N\}$, $\mathbf{v} = \{v_1, \ldots, v_N\}$ and $\mathbf{w} = \{w_1, \ldots, w_N\}$ such that

$$p_k = ax_k + by_k, \qquad q_k = cv_k + dw_k, \qquad a, b, c, d \in \mathbb{R}, \quad k = 1, \ldots, N,$$

then

$$\operatorname{cov}(\mathbf{p}, \mathbf{q}) = ac \operatorname{cov}(\mathbf{x}, \mathbf{v}) + ad \operatorname{cov}(\mathbf{x}, \mathbf{w}) + bc \operatorname{cov}(\mathbf{y}, \mathbf{v}) + bd \operatorname{cov}(\mathbf{y}, \mathbf{w}),$$

where $\mathbf{p} = \{p_1, \ldots, p_N\}$ and $\mathbf{q} = \{q_1, \ldots, q_N\}$. In fact:

$$\begin{aligned}
\operatorname{cov}(\mathbf{p}, \mathbf{q}) &= \frac{1}{N}\sum_{k=1}^{N}(p_k - \overline{p})(q_k - \overline{q}) = \frac{1}{N}\sum_{k=1}^{N}(ax_k + by_k - a\overline{x} - b\overline{y})(cv_k + dw_k - c\overline{v} - d\overline{w}) = \\
&= \frac{1}{N}\sum_{k=1}^{N}(a(x_k - \overline{x}) + b(y_k - \overline{y}))(c(v_k - \overline{v}) + d(w_k - \overline{w})) = \\
&= ac\frac{1}{N}\sum_{k=1}^{N}(x_k - \overline{x})(v_k - \overline{v}) + ad\frac{1}{N}\sum_{k=1}^{N}(x_k - \overline{x})(w_k - \overline{w}) + \\
&+ bc\frac{1}{N}\sum_{k=1}^{N}(y_k - \overline{y})(v_k - \overline{v}) + bd\frac{1}{N}\sum_{k=1}^{N}(y_k - \overline{y})(w_k - \overline{w}) = \\
&= ac \operatorname{cov}(\mathbf{x}, \mathbf{v}) + ad \operatorname{cov}(\mathbf{x}, \mathbf{w}) + bc \operatorname{cov}(\mathbf{y}, \mathbf{v}) + bd \operatorname{cov}(\mathbf{y}, \mathbf{w}).
\end{aligned}$$

Covariance: properties

If we apply a transformation to the variables $\mathbf{x} = \{x_1, \dots, x_N\}$ and $\mathbf{y} = \{y_1, \dots, y_N\}$ such that

$$w_k = ax_k + by_k, \qquad a, b, \in \mathbb{R}, \qquad k = 1, \dots, N,$$

then

$$\sigma^2(\mathbf{w}) = a^2\sigma^2(\mathbf{x}) + b^2\sigma^2(\mathbf{y}) + 2ab\,\mathrm{cov}(\mathbf{x}, \mathbf{y}),$$

where $\mathbf{w} = \{w_1, \dots, w_N\}$. In fact:

$$\begin{aligned}\sigma^2(\mathbf{w}) &= \frac{1}{N}\sum_{k=1}^{N}(w_k - \overline{w})^2 = \frac{1}{N}\sum_{k=1}^{N}(ax_k + by_k - a\overline{x} - b\overline{y})^2 = \\ &= \frac{1}{N}\sum_{k=1}^{N}(a(x_k - \overline{x}) + b(y_k - \overline{y}))^2 = \\ &= \frac{a^2}{N}\sum_{k=1}^{N}(x_k - \overline{x})^2 + \frac{b^2}{N}\sum_{k=1}^{N}(y_k - \overline{y})^2 + \frac{2ab}{N}\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y}) = \\ &= a^2\sigma^2(\mathbf{x}) + b^2\sigma^2(\mathbf{y}) + 2ab\,\mathrm{cov}(\mathbf{x}, \mathbf{y}).\end{aligned}$$

Unbiased sample covariance

When data refer to a sample, a correct definition of covariance is given by means of the Bessel's correction:

$$\mathrm{cov}(\mathbf{x},\mathbf{y}) = \frac{1}{N-1}\sum_{k=1}^{N}(x_k-\overline{x})(y_k-\overline{y}) = \frac{N}{N-1}(\overline{xy}-\overline{x}\,\overline{y}).$$

Cauchy-Schwarz inequality

Given two lists $\mathbf{x} = \{x_1, \ldots, x_N\}$ and $\mathbf{y} = \{y_1, \ldots, y_N\}$, it is

$$\mathrm{cov}^2(\mathbf{x},\mathbf{y}) \leq \sigma^2(\mathbf{x})\sigma^2(\mathbf{y})$$

i.e.,

$$-\sigma(\mathbf{x})\sigma(\mathbf{y}) \leq \mathrm{cov}(\mathbf{x},\mathbf{y}) \leq \sigma(\mathbf{x})\sigma(\mathbf{y}).$$

Geometric interpretation of covariance

- The covariance is maximal, *i.e.*, $\mathrm{cov}(\mathbf{x},\mathbf{y}) = \sigma(\mathbf{x})\sigma(\mathbf{y})$, when the points are aligned along an increasing straight line.
- The covariance is minimal, *i.e.*, $\mathrm{cov}(\mathbf{x},\mathbf{y}) = -\sigma(\mathbf{x})\sigma(\mathbf{y})$, when the points are aligned along a decreasing straight line.
- The covariance is zero, *i.e.*, $\mathrm{cov}(\mathbf{x},\mathbf{y}) = 0$, when the points are scattered.

Correlation coefficient

We may consider the correlation coefficient, denoted by ρ, between the two lists (or variables) **x** and **y**, say

$$\rho = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sqrt{\sigma^2(\mathbf{x})\sigma^2(\mathbf{y})}}.$$

that measures how much the data are linearly correlated, *i.e.*, as much as the absolute value of the correlation coefficient approaches to 1.

This coefficient is independent on the scale used to measure the data.

Due to the Cauchy-Schwarz inequality, it is

$$-1 \leq \rho \leq 1.$$

Correlation coefficient: property

The correlation coefficient is equal to the covariance of the standardized data.
In fact:

$$\rho = \frac{\operatorname{cov}(\mathbf{x},\mathbf{y})}{\sqrt{\sigma^2(\mathbf{x})\sigma^2(\mathbf{y})}} = \frac{\operatorname{cov}(\mathbf{x},\mathbf{y})}{\sigma(\mathbf{x})\sigma(\mathbf{y})} =$$

$$= \frac{1}{\sigma(\mathbf{x})\sigma(\mathbf{y})}\frac{1}{N}\sum_{k=1}^{N}(x_k-\overline{x})(y_k-\overline{y}) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\overline{x}}{\sigma(\mathbf{x})}\right)\left(\frac{y_k-\overline{y}}{\sigma(\mathbf{y})}\right) = \operatorname{cov}(\mathbf{z}_x,\mathbf{z}_y).$$

Geometric interpretation of correlation coefficient

- If $\rho > 0$, as one variable increases, the other one also increases. If $\rho = 1$, the points are aligned along an increasing straight line.
- If $\rho < 0$, as one variable increases, the other one decreases. If $\rho = -1$, the points are aligned along a decreasing straight line.
- If $\rho = 0$ does not exist a linear relationship among the variables.

Example: computation of covariance and correlation coefficient

Let $\mathbf{x} = \{1, 4, 6, 10, 11\}$ and $\mathbf{y} = \{0, 1, 2, 3, 4\}$ be two lists.

$$\overline{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 6.4, \qquad \overline{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = 2,$$

$$\sigma^2(\mathbf{x}) = \frac{1}{4}\sum_{i=1}^{5}(x_i - \overline{x})^2 = \frac{1}{4}\left((1-6.4)^2 + (4-6.4)^2 + (6-6.4)^2 + (10-6.4)^2 + (11-6.4)^2\right) = 17.3,$$

$$\sigma^2(\mathbf{y}) = \frac{1}{4}\sum_{i=1}^{5}(y_i - \overline{y})^2 = \frac{1}{4}\left((0-2)^2 + (1-2)^2 + (2-2)^2 + (3-2)^2 + (4-2)^2\right) = 2.5,$$

$$\begin{aligned}\mathrm{cov}(\mathbf{x},\mathbf{y}) = \frac{1}{4}\sum_{i=1}^{5}(x_i - \overline{x})(y_i - \overline{y}) &= \frac{1}{4}\big((1-6.4)(0-2) + (4-6.4)(1-2) + (6-6.4)(2-2) \\ &\quad + (10-6.4)(3-2) + (11-6.4)(4-2)\big) = 6.5.\end{aligned}$$

Furthermore, the correlation coefficient is positive and very close to 1! In fact:

$$\rho = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sqrt{\sigma^2(\mathbf{x})\sigma^2(\mathbf{y})}} = \frac{6.5}{\sqrt{17.3 \cdot 2.5}} = 0.988372.$$

Limitations of Correlation

- If the covariance is zero, then there is no linear relationship among the data. But it is possible there exists another relation (non-linear) between the data!
- Correlation does not imply causation: if two variables are highly correlated, it does not mean one variable causes the other to change.
- Correlation is sensitive to outliers: extreme values can distort the correlation, making the relationship appear stronger or weaker than it is.

Example

Consider the lists $\mathbf{x} = \{-3, -2, -1, 0, 1, 2, 3\}$ and $\mathbf{y} = \{9, 4, 1, 0, 1, 4, 9\}$. We have:

$$\overline{x} = \frac{1}{7}\sum_{k=1}^{7} x_k = \frac{-3-2-1+0+1+2+3}{7} = 0, \quad \overline{y} = \frac{1}{7}\sum_{k=1}^{7} y_k = \frac{9+4+1+0+1+4+9}{7} = 4;$$

then, the covariance is:

$$\begin{aligned}\operatorname{cov}(\mathbf{x}, \mathbf{y}) &= \frac{1}{6}\sum_{k=1}^{7}(x_k - \overline{x})(y_k - \overline{y}) = \\ &= \frac{-3(9-4) - 2(4-4) - 1(1-4) + 0(0-4) + 1(1-4) + 2(4-4) + 3(9-4)}{6} = 0.\end{aligned}$$

Then, there is not a linear relationship between $\mathbf{x}$ and $\mathbf{y}$. However, there is a quadratic relation!
In fact:

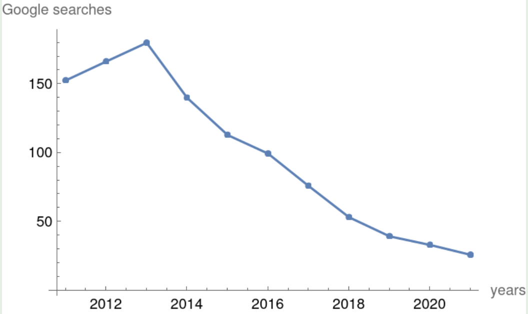
$$y_i = x_i^2, \qquad i = 1, \ldots, 7.$$

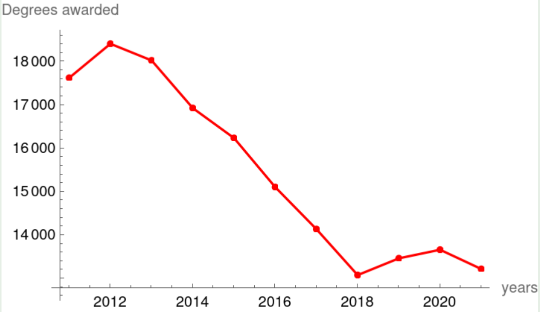
Example: spurious correlations

Let us consider the numerical data

$$\mathbf{x}_1 = \{152.583, 166.333, 179.917, 139.917, 112.917, 99.4167, 75.9167, 53.1667, 39.3333, 33.0833, 25.9167\},$$
$$\mathbf{x}_2 = \{17624, 18409, 18025, 16926, 16236, 15105, 14136, 13078, 13464, 13662, 13218\},$$

that are the collections of the relative volume of Google searches for 'facebook' (Worldwide, without quotes) and the associates degrees conferred by postsecondary institutions with a field of study in 'Visual and performing arts', respectively. The data have been collected in the range interval of years $[2011, 2021]$. We can see the following trends:



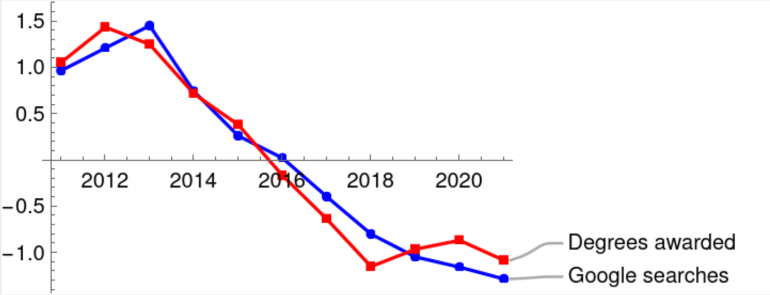


Example: spurious correlations

The data are expressed in different scales of misure. Then, let us standardize the two numerical lists, *i.e.*, construct the vectors $\mathbf{z}_i = \dfrac{\mathbf{x}_i - \overline{\mathbf{x}}_i}{\sigma(\mathbf{x}_i)}$ $(i = 1, 2)$:

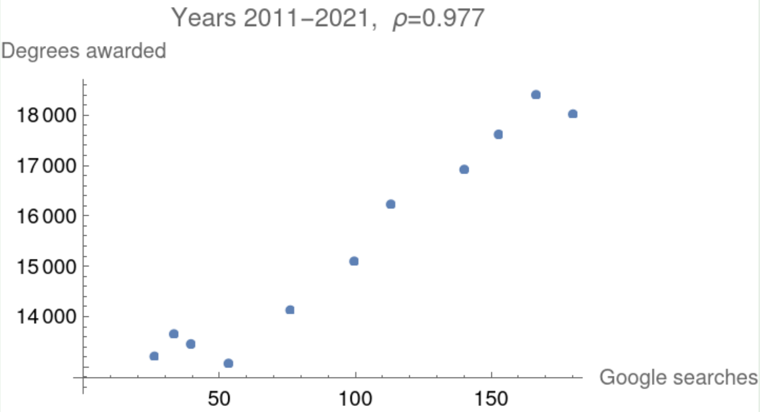
$$\begin{aligned}\mathbf{z}_1 =& \{0.968796, 1.21305, 1.45435, 0.743799, 0.264175, 0.0243579, -0.393092,\\ & -0.79722, -1.04295, -1.15398, -1.28128\},\\ \mathbf{z}_2 =& \{1.05746, 1.43822, 1.25196, 0.71889, 0.384205, -0.164388, -0.634403,\\ & -1.14759, -0.960357, -0.864317, -1.07968\}.\end{aligned}$$

We can see the corresponding trends together:



Example: spurious correlations

The two vectors looks highly correlated! In fact, the correlation coefficient between $\mathbf{x}_1$ and $\mathbf{x}_2$ is 0.977.



The question is: does all this make sense? Of course, NO! Correlation does not imply causation!

Example: spurious correlations

Let us consider the numerical data

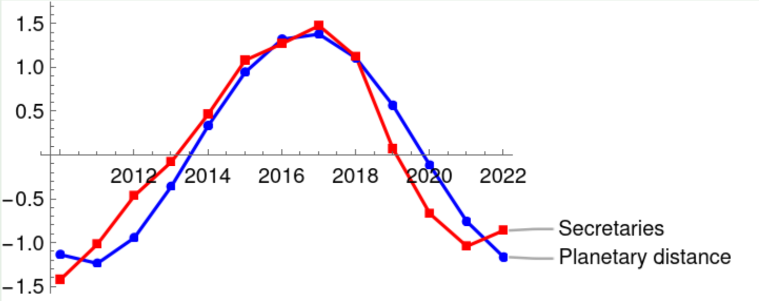
$$\begin{aligned}\mathbf{x}_1 = \{&4.97305, 4.95401, 5.00953, 5.12053, 5.25217, 5.36786, 5.43929,\\ &5.45028, 5.39873, 5.29605, 5.16633, 5.04464, 4.96744\},\\ \mathbf{x}_2 = \{&3510, 4020, 4720, 5210, 5900, 6680, 6920, 7180, 6730, 5400, 4460, 3990, 4220\},\end{aligned}$$

that are the collections of the the average distance between Jupiter and the Sun as measured on the first day of each month and the the number of secretaries and administrative assistants, except legal, medical, and executive in Alaska, respectively. The data have been collected in the range interval of years $[2010, 2022]$. In order to make the data comparable, let us standardize them:

$$\begin{aligned}\mathbf{z}_1 = \{&-1.13187, -1.23227, -0.939492, -0.35414, 0.340056, 0.95014, 1.32682,\\ &1.38478, 1.11293, 0.571454, -0.112616, -0.754341, -1.16145\},\\ \mathbf{z}_2 = \{&-1.4156, -1.01296, -0.460327, -0.0734823, 0.471258, 1.08705, 1.27653,\\ &1.48179, 1.12653, 0.0765187, -0.665591, -1.03665, -0.855066\}.\end{aligned}$$

Example: spurious correlations

The following plot shows the trends:



The two vectors looks highly correlated! In fact, the correlation coefficient is $\rho = 0.95$.

Correlation is sensitive to outliers!

Let us consider the couples of numerical data

$$\begin{aligned}\mathbf{x}_1 &= \{4.97305, 4.95401, 5.00953, 5.12053, 5.25217, 5.36786, 5.43929,\\ &\quad 5.45028, 5.39873, 5.29605, 5.16633, 5.04464, 4.96744\},\\ \mathbf{x}_2 &= \{3510, 4020, 4720, 5210, 5900, 6680, 6920, 7180, 6730, 5400, 4460, 3990, 4220\},\end{aligned}$$

and

$$\begin{aligned}\mathbf{y}_1 &= \{4.97305, 4.95401, 5.00953, 5.12053, 5.25217, 5.36786, 5.43929,\\ &\quad 5.45028, 5.39873, 5.29605, 5.16633, 5.04464, 4.96744, 4.9\},\\ \mathbf{y}_2 &= \{3510, 4020, 4720, 5210, 5900, 6680, 6920, 7180, 6730, 5400, 4460, 3990, 4220, 7150\},\end{aligned}$$

where, in the second one, we have just added the outlier $(4.9, 7150)$. We have that

$$\mathrm{corr}(\mathbf{x}_1, \mathbf{x}_2) = 0.95$$

and

$$\mathrm{corr}(\mathbf{y}_1, \mathbf{y}_2) = 0.66$$

Covariance matrix

If we have p lists (or variables, representing p features) $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p \in \mathbb{R}^N$, we may construct the $p \times p$ covariance matrix whose entries are the covariances associated with all possible pairs of variables $\mathbf{x}_i$, *i.e.*,

$$\operatorname{cov}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{N}\sum_{k=1}^{N}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j), \qquad i, j = 1, \dots, p.$$

Since the covariance of a list with itself is its variance ($\operatorname{cov}(\mathbf{x}_i, \mathbf{x}_i) = \sigma^2(\mathbf{x}_i)$), in the main diagonal we have the variances of each variable. The covariance matrix is symmetric, and positive semi-definite; then it can be diagonalized. By looking for its eigenvalues and eigenvectors, and using an orthogonal basis, the distribution of data can be characterized: this is the object of Principal Components Analysis which can be seen as a type of compression information.

Correlation matrix

If we have p lists (or variables, representing p features) $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p \in \mathbb{R}^N$, we may construct the $p \times p$ correlation matrix whose entries are the correlation coefficients associated with all possible pairs of variables $\mathbf{x}_i$, *i.e.*,

$$\operatorname{corr}(\mathbf{x}_i, \mathbf{x}_j) = \rho_{ij} = \frac{\operatorname{cov}(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{\sigma^2(\mathbf{x}_i)\sigma^2(\mathbf{x}_j)}}, \qquad i, j = 1, \dots, p.$$

Correlation matrix is a symmetric and positive semi-definite matrix: on the main diagonal the entries are equal to 1 (each variable is trivially correlated to itself).

Covariance matrix

$$\begin{bmatrix} \sigma^2(\mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_2) & \dots & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_p) \\ \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_1) & \sigma^2(\mathbf{x}_2) & \dots & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\mathbf{x}_p, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_p, \mathbf{x}_2) & \dots & \sigma^2(\mathbf{x}_p) \end{bmatrix}$$

Correlation matrix

$$\begin{bmatrix} 1 & \mathrm{corr}(\mathbf{x}_1, \mathbf{x}_2) & \dots & \mathrm{corr}(\mathbf{x}_1, \mathbf{x}_p) \\ \mathrm{corr}(\mathbf{x}_2, \mathbf{x}_1) & 1 & \dots & \mathrm{corr}(\mathbf{x}_2, \mathbf{x}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{corr}(\mathbf{x}_p, \mathbf{x}_1) & \mathrm{corr}(\mathbf{x}_p, \mathbf{x}_2) & \dots & 1 \end{bmatrix}$$

Covariance matrix vs. Correlation matrix

- The covariance matrix measures how changes in one variable are associated with changes in another, but without standardizing the values.
- The correlation matrix represents the strength and direction (positive/negative) of variables relationships on a standardized scale from -1 to $+1$.

Example

The manager of a clothing store is analyzing the sales of classic blue sweaters in 10 different periods of the year. The manager has the number $\mathbf{x}_1$ of sweaters sold, the price (in euros) variations $\mathbf{x}_2$, the cost (in euros) $\mathbf{x}_3$ of the advertising that appeared in newspapers, and the number of hours $\mathbf{x}_4$ the sales assistant was present in the store. The data are:

$$\mathbf{x}_1 = \{230, 181, 165, 150, 97, 192, 181, 189, 172, 170\}, \quad \mathbf{x}_2 = \{125, 99, 97, 115, 120, 100, 80, 90, 95, 125\},$$
$$\mathbf{x}_3 = \{200, 55, 105, 85, 0, 150, 85, 120, 110, 130\}, \quad \mathbf{x}_4 = \{109, 107, 98, 71, 82, 103, 111, 93, 86, 78\}.$$

The manager believes that the price variations influences the number of sweaters sold. Is it correct?

Covariance matrix

$$\begin{bmatrix} 1152.46 & -88.91 & 1589.67 & 301.6 \\ -88.91 & 244.27 & 102.33 & -101.76 \\ 1589.67 & 102.33 & 2915.56 & 233.67 \\ 301.6 & -101.76 & 233.67 & 197.07 \end{bmatrix}$$

Correlation matrix

$$\begin{bmatrix} 1 & -0.17 & 0.87 & 0.63 \\ -0.17 & 1 & 0.12 & -0.46 \\ 0.87 & 0.12 & 1 & 0.31 \\ 0.63 & -0.46 & 0.31 & 1 \end{bmatrix}$$