

Table of contents

- 1 **Statistics**
 - Introduction
 - Vocabulary
- 2 **Descriptive Statistics**
 - Variables
 - Graphical representations
 - Measures of the Central Tendency
 - The Mean
 - The Median
 - The Mode
 - Quartiles
 - Measures of variability
 - The range
 - Sample Variance and Sample Standard Deviation
 - The median, the quartiles and the interquartile range
 - Outliers
- 3 **Bibliography**

Some History

- The roots of word Statistics are Latin: Status (old Latin) which means (political) state, Statista (Italian) which means politician.
- In mid eighteenth century at a German University the word statistik was first used with the mean of political science of states: analysis of data concerning states.
- In Great Britain at the end of the eighteenth century the term statistics was introduced with a similar meaning: the science of states (or political arithmetic).
- The use of statistics without name it in this way dates back to the beginning of civilisations: early forms of population censuses, data concerning geographical and economical features etc.

Some History

- The first statistical study (which is widely considered to be the foundation of demography): in 1662 two Englishmen introduced some statistical methods as life tables and probabilities of survival to each age.
- Only in the nineteenth century results from probability theory started to be used in statistical reasoning.
- The mathematical foundations for statistics advanced due to the profound results in probability theory in the last century.
- Starting from the twentieth century new methods and theories were developed and the breakthrough of the computers had a major influence of advances in statistics.

Introduction

"Statistics has become the universal language of the sciences."

Elementary Statistics, R. Johnson, P. Kuby

A *typical statistical problem* must involve

- one or more random experiences which result in a series of data.
- the process of extracting information from the data and interpret the results.

The way in which the information is processed and interpreted gives the two main branches of *Statistics* as a science:

- *Descriptive Statistics* - collects, presents and describes the data (sometimes in a graphical form).
- *Inferential Statistics* - based on the collected data it makes decisions about the populations involved.

Introduction

Definiția 1

Statistics is the science of collecting, describing, interpreting data and making decisions based on these data.

The two areas already presented are the two steps of a statistical study:

- descriptive statistics has the role of
 - synthesize, summarize and display the data;
 - arrange the information and prepare it for making decisions.
- inferential statistics aims to
 - make decision based on gathered information;
 - estimate the parameters (such as expectation, variance etc);
 - verify statistical hypothesis.

Statistical Terminology

- Statistics has its own language and terminology.
- The most important concept in statistics is that of *population*: the complete collection of objects that are of interest for the collector.
- Example of populations: the set of all students in Iași, the set of all illiterate people in Romania, the set of all cans of cola produced in a month by a manufacture, the set of all tropical storms in 2021.

Definiția 2

A *population* is a set of objects (or individuals) whose properties are to be analyzed.

- A population can be finite (if it could be listed) or infinite (the population of all earthquakes in Vrancea area).

Statistical Terminology

- Because of the large size of the populations the statistical study focuses only on a small portions of a population. This is the *sample* which consists of individuals collected from the population.

Definiția 3

*A **sample** is a subset of a population. From a theoretical point of view we will require that each individual has the same chances to belong to the sample and any particular individual is independently chosen to belong to that sample. If these properties hold we have a **simple random sample**.*

- When we choose to study a population or a sample of it, a certain characteristic of the individuals is of interest.
- Such attributes could be: height, volume, Richter magnitude (for earthquakes), age, blood pressure, eyes color, area etc.

Statistical Terminology

Definiția 4

A **variable** or an **attribute** is a characteristic of interest of individuals from a population or a sample.

- When we choose a sample we have to measure the values of one or more associated variables. These are **data values**, or simply **data**, and can be real or integer numbers, words, letters etc.

Definiția 5

Data is the set of values for the variable(s) collected from each individual belonging to the sample.

- A population is numerically described by its **parameters** (such as expectation, variance, standard deviation). Parameters are the goals of a statistical study.

Statistical Terminology

Definiția 6

A **parameter** is a numerical value concerning the entire population.

- When the population is very large (which is very often the case) a parameter cannot be computed.
- A solution is to compute the value of the parameter only for a sample of the population. This is a **sample statistic**.
- For every parameter and every sample there exists a corresponding sample statistic.

Definiția 7

A **sample statistics** is a parameter computed for a sample instead of the entire population.

Use of Terminology

- *Population*: the set of all freshmen (first year students) in Iași.
- *Sample*: the freshmen from CS Department. (Note that this sample is not a simple random sample.)
- *Variable/attribute*: their vocabulary size.
- *Data*: 4200, 3520, 1800, ... - measure the size for each freshmen.
- *Parameter*: the expectation (average) of the vocabulary size of a student in Iași.
- *Statistic*: the average size of vocabulary for freshmen from CS Department.

Types of Variables

- The first classification views a variable as a quantitative/qualitative attribute. Therefore we have
- Variables that result in *qualitative information*, like the eye color of students, the genre of books in a library (fiction, science, motivational etc), the personality types of people from a community (sanguine, choleric, melancholic or phlegmatic), the level of satisfaction of costumers in a mall shop etc.
- Variables that result in *quantitative information*, like the height of students, their weight, the money each student spends for science books in a year etc.

Types of Variables

Definiția 8

A **qualitative** (or **categorical**) **variable** is a variable that describes an individual from a population (according to a categorization).

A **quantitative** (or **numerical**) **variable** is a variable that measures or counts something about an individual from a population.

- Qualitative variables can be **nominal** or **ordinal**.
- Nominal variables are: the eye color of students, the personality type of people, the names of people from a community etc.
- Examples of ordinal variables are: the level of satisfaction of customers, the level of educational experience (high-school graduate, college graduate, Ph. D.) etc.

Types of Variables

Definiția 9

A **nominal variable** is a variable that names or describes an individual from a population and a natural order cannot be assigned to the values of this variables.

An **ordinal variable** is a qualitative variable whose values can be naturally ordered.

- Quantitative variables can be **discrete** or **continuous**. The two types of variables can be distinguished by deciding if they are related to counting or to measuring.

Types of Variables

- A discrete variable usually counts: the number of credits of a student, the number of pages of a book etc; sometimes such a variable gather a score or a grade which cannot be continuous.
- A continuous variable measures: the volume, the height, the speed, the pressure etc.

Definiția 10

A **discrete variable** is a variable that have a countable number of values; such a variable can assume any value from a set of isolated points along a line interval.

A **continuous variable** is a variable that have a uncountable number of values; such a variable can assume any values along a line interval (including every possible value between any two values).

Graphical representations

- A first exploration of data is the use of graphical representations which can reveal patterns of behavior of the variable being studied.
- The type of graphical display depends basically on the type of the variable.
- For qualitative data the graphical representations used are pie charts and bar graphs.
- For quantitative data the purpose of the graphical representation is to find the shape of its distribution.

Graphical representations - qualitative data

- The qualitative data is first transformed in frequencies.
- The *frequency* of an observation (a value of the variable) is the number of times that observation occurs in the sample.
- The *relative frequency* of an observation is the ratio between the the number of times that observation occurs in the sample and the size of the entire sample.
- The *frequency distribution* of a qualitative variable is the set of all distinct observations and the set of their relative frequencies.

Graphical representations - quantitative data

- For quantitative data we can use the frequencies and the relative frequencies or the *grouping* to form a frequency distribution:
 - we *group* data in classes (or *bins*) which usually are interval with the same width; the classes must not overlap each other.
 - a rule for the common width of the bins is $1 + \log n / \log 2$ where n is the size of the sample.
 - then we *sort* the data in the classes: we find the number of observations belonging to each class - these are the frequencies.
 - the sum of all frequencies must equal the size of the sample (n).
 - we can find the relative frequencies by dividing each frequency to n .

Graphical representations - quantitative data

- The most use graphical representation for quantitative data is the *histogram*.
- Another convenient way, for relative small samples is the *stem-and-leaf* plot.

Looking at the data

- When we look at the graphical representation or simply at the data of a sample, we may ask the following questions.
- What are the central/average values?
- How much are the data spread (around its average values)?
- What is the shape of the distribution?
- Are there awkward values, that is data points which doesn't make sense in the general picture?

Central Tendency

- The *central tendency* or the *center of the distribution* is the (abstract) center of the data. All the measures of the central tendency are related in a way or another to the term average.
- Different ways to define the central tendency:
 - The point at which the distribution is in balance.
 - The number that minimizes the sum of the absolute deviations of all the values.
 - The number that minimizes the sum of squared deviations of all the values.
 - The most frequently occurring value.

The (Arithmetic) Mean

- Suppose that the values from the sample are x_1, x_2, \dots, x_n .

left

Definiția 11

The sample mean is the arithmetic average of all sample values:

$$\bar{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

- The formula for the mean of the entire population is essentially identical.
- The *population mean* is denoted by μ .

The Mean, the Sample and the Population

- In terms of probabilistic language the population mean is the expectation of the random variable (X) whose values are the attributes of the individuals. Hence $\mathbb{E}[X] = \mu$.
- The sample mean is a statistic that estimates the population mean.
- Suppose that X_1, X_2, \dots, X_n are the variables behind each individual from our sample and x_i is just a value of X_i .
- Then X_i is a random variable having the same distribution as X . Moreover, the variables $(X_i)_{1 \leq i \leq n}$ are jointly independent.
- These observations lead to the fact that the sample mean can be viewed as a random variable, the already computed arithmetic mean is just one of its possible values (each sample gives a value for the sample mean).

The Mean, the Sample and the Population

- If the sample mean is a random variable we can compute its expectation:

$$\begin{aligned}\mathbb{E}[\bar{x}_n] &= \mathbb{E}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \\ &= \frac{1}{n}\mathbb{E}[X_1 + X_2 + \cdots + X_n] = \\ &= \frac{\mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]}{n} = \mu\end{aligned}$$

- The expectation of the sample mean is the mean of the population.
- We call such a statistic an *unbiased estimate* of the corresponding parameter.
- The sample mean is an unbiased estimate of the population mean.

The Mean for Grouped Data

- The formula from the above definition holds for ungrouped data. For such data all the observations contribute to the value of the mean.
- For the mean of grouped data we can use an weighted formula:

$$M = \frac{m_i * f_i}{\sum_i f_i}$$

where m_i is the middle of the interval of the i th class, and f_i is the number of observations that belong to this class.

- In the above formula the observations do not directly contribute to the mean; however this formula is very useful: for very large samples it is easier to use this formula than the arithmetic mean.

Properties of the Mean

- We return now to the main definition of the mean.
- Small variations in the sum on the numerator doesn't change much in the final mean. We say the the mean is stable to small data variations.
- *Outliers* can have a big influence on the value of the mean; that is introducing very large, or very small values in the sample can dramatically change the mean.
- The mean is a *linear* function as it is the expectation of a random variable.
- The *deviations* from the mean are $(x_i - \bar{x}_n)$; their sum is zero:

$$\sum_i (x_i - \bar{x}_n) = 0.$$

Properties of the Mean

- (*Variational definition*) It can be proved that the mean is the number M that minimizes the sum of the squared deviations:

$$\sum_i (x_i - M)^2.$$

- There are other types of mean besides the arithmetic mean (A): the geometric mean (G), the harmonic mean (H).

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}, H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}.$$

- Suppose that a car goes four times between two cities with the following speeds 80km/h, 90km/h, 60km/h, and 120km/h, respectively. What was its average speed?
- Using the arithmetic mean we get 87.5 km/h; but the appropriate mean here is the harmonic mean which gives 82.3km/h.

The Median

- The median is an *order statistic*; the computation of such a statistic requires to arrange the data in increasing order.

Definiția 12

The median (Me) is the value which holds the middle position when the data are ranked in increasing order.

- The median splits the data in two halves: one half contains data greater or equal with the median, and the second half consists of the values less than or equal with the median.
- The value of the median is an observation or the average of two observations (for an even sized sample).
- As a statistic the median is much less influenced by the existence of outliers.

The Mode

Definiția 13

The **mode** is the value that occurs most frequently.

- For grouped data we first choose the class having the largest frequency, this will be the **modal class**. Let i be the index of that class, and a_i the left bound of its interval, L be the length of the interval(s).
- Then the mode can be computed using the following formula

$$\text{mode} = a_i + \frac{L * (f_i - f_{i-1})}{(f_i - f_{i-1}) + (f_i - f_{i+1})}$$

- The **anti-mode** is the less frequent value.

Comparing the different measures

- The mode and the mean are more stable in the presence of outliers.
- The mean incorporates all the values, and cannot be computed for open distributions (first, or the last interval being open).
- The median and the mode are not linear functions of the sample.
- The mode is computed mainly for grouped data.
- For asymmetric distributions the mode gives the most real image of the central tendency.

Comparing the different measures

- When the sample contains unusual large (or small) values the median is the preferred measure because its stability makes it more representative.
- For symmetric distributions the three measures are almost equal.
- The shape of the distribution can be related to the mean and median. The shape is
 - Left skewed if $\bar{x}_n < Me$;
 - Symmetric if $\bar{x}_n = Me$;
 - Right skewed if $\bar{x}_n > Me$;

Quartiles

- Related to the measures of central tendency are the *measures of position* which are order statistics (like the median).

Definiția 14

The quartiles are the values that divide the ranked data in four equal parts.

- The first quartile, Q_1 , is a value such that at most 25% of the data are smaller in value than Q_1 and at most 75% are larger.
- The third quartile, Q_3 , is a number such that at most 75% of the data are smaller in value than Q_3 and at most 25% are larger.

Quartiles

- The second quartile, Q_2 , is a value such that at most 50% of the data are smaller in value than Q_2 and at most 50% are larger. For this reason the second quartile is in fact the median: $Me = Q_2$.
- The quartiles have similar properties to those the median. The most important being that they are stable in the presence of outliers.
- Similar order statistics are: the *deciles*, the *percentiles* etc. All these statistics split the ranked data in equal subsamples.
- For example there are nine deciles that divide the sorted data into ten equal parts, each part representing 10% of the sample.

Measures of variability

- After locating the "center" of the data the search for more information turns to the variability or the spreading of the values.
- Within the data sample the values usually differ one from one another and from the "center" value.
- The extent to which the "center" value is a good representative of the values in the sample depends upon the variability or dispersion in the data.
- Samples are said to have high dispersion when they contain values considerably higher and/or lower than the average value.
- Because basically we have two measures of the central tendency (mean and median), we will have two main types of measuring the variability.

The range

Definiția 15

The range is the difference between highest and the lowest values.

$$\text{range} = \max - \min.$$

- Since the range is based only on the two most extreme values, if one of these is either very high or low it will result in a range that is not typical of the variability within the sample.
- We say that the outliers have a very deep (and direct) influence on the range.

Sample Variance

We present next the measures of variability about the mean.

- We already noted that the deviations from the mean are $(x_i - \bar{x}_n)$.
- The deviation $(x_i - \bar{x}_n)$ is positive (negative) when x_i is larger (smaller) than the sample mean.
- In order to describe an "average" value of all deviations we might use the arithmetic mean deviation. However the sum of all deviations is zero, therefore their mean will be zero.
- We can remove this effect by squaring the deviations, and then computing a certain kind of quadratic mean.

Sample Variance

Definiția 16

The **sample variance**, s^2 , is the mean of the squared deviations using $(n - 1)$ as the denominator (instead of n), n being the size of the sample:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{n - 1},$$

- The sample variance is non-negative, and it will be zero if and only if the sample contains the same constant value.
- The sample variance is the statistic corresponding to the **variance of the population**, usually denoted by σ^2 .

Sample Variance

- The reason we use $(n - 1)$ as a denominator in the definition of the sample variance is that we want to get an unbiased estimate.
- The expectation of the sample variance (viewed as a random variable) is

$$\begin{aligned} \mathbb{E}[s^2] &= \mathbb{E} \left[\frac{\sum_{i=1}^n (X_i - \bar{x}_n)^2}{n - 1} \right] = \mathbb{E} \left[\frac{\sum_{i=1}^n \left(nX_i - \sum_{j=1}^n X_j \right)^2}{n^2(n - 1)} \right] = \\ &= \frac{\sum_{i=1}^n \mathbb{E} \left[n^2 X_i^2 - 2nX_i \left(\sum_{j=1}^n X_j \right) + \left(\sum_{j=1}^n X_j \right)^2 \right]}{n^2(n - 1)} \end{aligned}$$

Sample Variance

$$\begin{aligned}
 &= \frac{\sum_{i=1}^n \mathbb{E} [n^2 X_i^2] - 2n \sum_{i=1}^n \mathbb{E} \left[\sum_{j=1}^n X_i X_j \right] + n \mathbb{E} \left[\sum_{j=1}^n X_j^2 + 2 \sum_{i < j} X_i X_j \right]}{n^2(n-1)} \\
 &= \frac{n^2 \sum_{i=1}^n \mathbb{E} [X_i^2] - 2n \sum_{i=1}^n \mathbb{E} [X_i^2] - 2n \sum_{i=1}^n \sum_{j \neq i} \mathbb{E} [X_i X_j]}{n^2(n-1)} + \\
 &\quad + \frac{n \sum_{j=1}^n \mathbb{E} [X_j^2] + 2n \sum_{i < j} \mathbb{E} [X_i X_j]}{n^2(n-1)} =
 \end{aligned}$$

Sample Variance

$$\begin{aligned}
 & n(n-1) \sum_{i=1}^n \mathbb{E}[X_i^2] - 2n \sum_{i < j} \mathbb{E}[X_i X_j] \\
 &= \frac{n(n-1) \sum_{i=1}^n \mathbb{E}[X_i^2] - 2n \sum_{i < j} \mathbb{E}[X_i X_j]}{n^2(n-1)} \\
 &= \frac{n^2(n-1) \sum_{i=1}^n \mathbb{E}[X_i^2] - 2n \sum_{i < j} \mathbb{E}[X_i] \mathbb{E}[X_j]}{n^2(n-1)} \\
 &= \frac{\sum_{i=1}^n \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2}{n} = \sigma^2.
 \end{aligned}$$

Sample Standard Deviation

- A more simple formula (left as an exercise) for the sample variance is

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

Definiția 17

The sample standard deviation, s , is the square root of the sample variance.

- The sample standard deviation is a biased estimate of the standard deviation of the population, σ .
- It can be proved that the sample standard deviation is an underestimate of the standard deviation of the population, that is $\mathbb{E}[s] < \sigma$.

Five number summary

- We look now to the measures of variability about the median. First, a very effective way of describing data is the 5-number summary.

Definiția 18

The **five number summary** is composed of the following

- 1 *min*, the smallest value in the data set;
- 2 Q_1 , the first quartile;
- 3 Me , the median;
- 4 Q_3 , the third quartile;
- 5 *max*, the largest values in the data set.

Interquartile range

- A way of displaying the 5-number summary is the *box-and-whiskers* display.

Definiția 19

The **midquartile** is the middle value between the first and the third quartiles:

$$\text{midq} = \frac{Q_1 + Q_3}{2}.$$

The **interquartile range** is the difference between the third and the first quartiles:

$$\text{IQR} = Q_3 - Q_1.$$





Outliers

- *Outliers* are the values in the data set that can be considered too small or too large for the general "picture" of the sample.
- Obviously the outliers are related to the notion of variability of the data. Usually these values come from errors of measurement, but, however, they can have natural causes.
- Sometimes, when appropriate, we first delete the outliers of a sample before performing any other statistical analysis.
- We will give two rules of detecting the outliers using the two types of measuring the variability.

Outliers

- The first rule is related to the mean. We can consider as outliers the values in the data set which don't belong to the interval $(\bar{x}_n - 2s, \bar{x}_n + 2s)$.
- The second rule called $1.5 * IQR$ rule says that any value outside the interval $(Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR)$ is an outlier.

Bibliography

-  Freedman, D., R. Pisani, R. Purves, *Statistics*, W. W. Norton & Company, 4th edition, 2007.
-  Johnson, R., P. Kuby, *Elementary Statistics*, Brooks/Cole, Cengage Learning, 11th edition, 2012.
-  Shao, J., *Mathematical Statistics*, Springer Verlag, 1998.
-  Spiegel, M. R., L. J. Stephens, *Theory and Problems of Statistics*, Schaum's Outline Series, McGraw Hill, 3rd edition, 1999.