

COURSE 1

ThinkBS

16 th November 2021

Lecturer PhD Math. Raluca Purnichescu Purtan

Department of Mathematical Methods and Models
Faculty of Applied Sciences, UPB

raluca.purnichescu@upb.ro

STATISTICAL METHODS WITH APPLICATIONS IN ENGINEERING AND MEDICINE

OUTLINE



- ✓ Statistical population, samples, sample size, parameters
- ✓ Descriptive statistics – central tendencies, variation, representations
- ✓ Data types and measurements
- ✓ Inferential statistics – estimations, confidence intervals, statistical tests (parametric and nonparametric)
- ✓ Applications in engineering and medicine



STATISTICAL POPULATION AND SAMPLES

- In statistics, a **population** is *the complete set of all units (elements, objects or people) of interest*. These units should have at least one *common characteristic*.

Examples:

People living in the same country;

The set of all possible hands in a poker game;

Students at our University;

Electrical cables with a specific diameter;

Patients suffering from a specific disease .

- A **sample** is a *subset of the population*, selected for statistical purposes (*with the proper technique*).

The statistical process of selecting a sample is known as *sampling*.

Number of elements in the sample is the **sample size (or volume)**.

● Basically, the **sampling techniques** are grouped in two major categories:

- Probability sampling

- Non-probability sampling

} difference: the sample selection is based on **randomization** or not;

every element gets equal chance to be picked up and to be part of sample for study.



PROBABILITY SAMPLING

Use randomization to make sure that every element of the population *gets an equal chance to be part of the selected sample*.

It's known as *random sampling or chance sampling*.

The advantages of using a probability sampling are:

A comparatively easier method of sampling;

High level of reliability of research findings;

High accuracy of sampling error estimation;

The absence of both systematic and sampling bias.

The main disadvantages are:

Higher complexity;

More expensive and time-consuming;

Chances of selecting specific class of samples only.

Types of probability sampling

Simple Random Sampling

The most well-known method to obtain an unbiased, representative sample;

All items in the population have an equal probability of being selected;

Minimize the bias and simplifies data analysis;

Creates samples that are very highly representative of the population;

Avoids the issue of consecutive data to occur simultaneously.

Other types of probability sampling

- *Cluster Random Sampling*
- *Systematic Sampling (Clustering)*
- *Multi-Stage Sampling*

Stratified Random Sampling

Is very appropriate when the population is heterogeneous;

Divides the elements of the population into small subgroups (strata) based on the similarity; the elements within the group are homogeneous and heterogeneous among the other subgroups formed;

The elements are randomly selected from each of the strata;

Leads to increased statistical efficiency.



NON-PROBABILITY SAMPLING

The samples are collected in a way that *does not give all the units in the population equal chances of being selected.*

Non-Probability sampling *does not involve random selection at all.*

The units in a non-probability sample are selected based on their accessibility.

The advantages of using a non-probability sampling are:

When a respondent refuses to participate, he may be replaced by another individual who wants to give information;

Very effective in terms of cost and time;

Easy to use.

The main disadvantages are:

*The possibility of gathering valuable data is reduced;
Impossible to estimate how well the researcher represent the population;*

Excessive dependence on judgment;

The researchers can't calculate margins of error;

Bias arises when selecting sample units;

The correctness of data is less certain.

Types of non-probability sampling

Purposive (Judgment, Authoritative) Sampling

*Where the researcher selects the units of the sample **based on their knowledge**;*

The units have the required characteristics to be representatives of the population;

Is used mainly when a restricted number of people possess the characteristics of interest;

It is a common method of gathering information from a very specific group of individuals.

Other types of non-probability sampling

- *Convenience Sampling*
- *Quota Sampling*
- *Referral / Snowball Sampling*

Example: *In a medical research article, there is written:*

“Using a retrospective institutional database, we identified 93 patients undergoing multimodal treatment for histologically diagnosed NELM at our center over a 15-year period, between 1st of January 2004 and 31st of December 2018”.



What type of sampling was used by the authors?

Purposive (Judgment, Authoritative) Sampling

Example: *In an engineering research article, there is written:*

“(…) the team decided to concentrate on reducing variation in crossbar length. To quantify the problem, they planned and executed a baseline investigation in which six consecutive parts were systematically sampled from the process each hour for five days”.



What type of sampling was used by the authors?



SUBPOPULATIONS

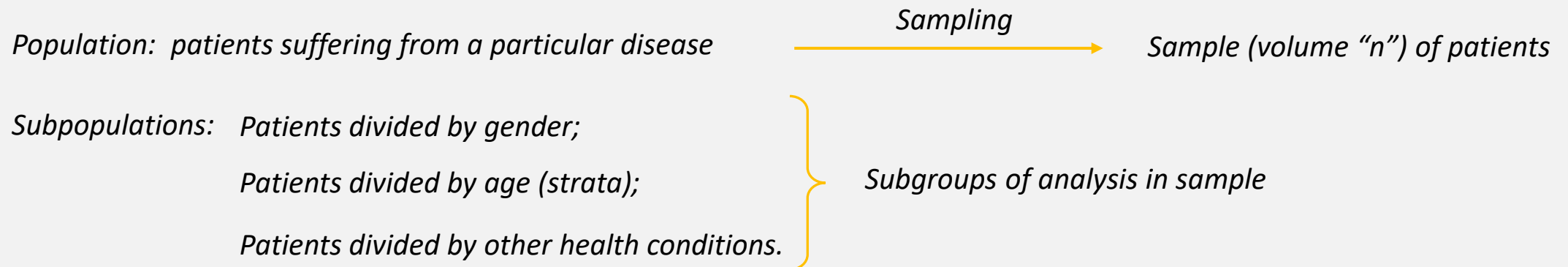
Subpopulations share additional attributes.

Subpopulations are particularly important when they have characteristics that are systematically different than the overall population.

When analyzing our data, we need to be aware of these deeper divisions. In fact, we can treat the relevant subpopulations as additional factors in later analyses.

Descriptive statistics may yield different results for different subpopulations.

Example: *A medical research is conducted on patients suffering from a particular disease, in order to study a new medicine drug and the benefits of it.*





POPULATION PARAMETERS AND SAMPLE STATISTICS

A parameter is a value that describes a characteristic of an entire population.



the value of a parameter is (usually) unknown – the entire population is almost never measured

*In a sample, the correspondent of the population's parameter is named "**statistic**" and it is an estimator of the population parameter's value.*

Example:

Population: *patients suffering from a particular disease* → *Sampling* → *Sample (volume "n") of patients*

Characteristics:

Parameter(s):

Statistic(s):

Age

mean, standard deviation

mean, standard deviation

Gender

proportion (quota)

proportion (quota)

BMI

mean, standard deviation, percentage

mean, standard deviation, percentage

Characteristics are represented by proper random variables

The parameters have a specific (theoretical) formulas and the correspondent statistics too.

In inferential statistics...

- we *use sample statistics* (mean, standard deviation or others) to estimate population parameters;
- we *perform hypothesis testing* on the sample estimate;
- we *create confidence intervals* to construct a range that actual population parameter value likely falls within.

Before any inferential statistic techniques are applied, the characteristics of the study population should be checked.

This step is vital for choosing the proper statistical method / model for the research study.



RANDOM VARIABLES

A *random variable* is a quantity which, upon the completion of a random experience, can take values from a well-defined set (the set of all possible outcomes of the experience).

Using random variables, a link can be established between the possible outcomes of a random phenomenon and the likelihood of these outcomes occurring.

A random variable is a function; In probability theory, there are two major types of random variables (depending on the nature of the described phenomenon): discrete or continuous variables.

In any statistical study, every characteristic of interest from a population must be described by a proper random variable.

The parameter(s) of every characteristic is the parameter of the correspondent random variable.

● **Discrete (simple) random variables**

The set of values of the discrete (simple) random variable is at most a countable set (in most real-life models, a finite set).

In probability theory, these random variables are given in a table form:

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}, \text{ with } p_i \geq 0 \text{ and } \sum_{i=1}^n p_i = 1.$$


Here, the numbers x_i are the values of the random variable X (values associated with the outcomes of the experience) and p_i are the probabilities with which the random variable can take these values.

In statistics, the **theoretical probabilities** (sometimes unknown) are replaced by the **relative frequencies** – this is the “statistical definition” of the probability – based on the Law of Large Numbers !

Example:

We are interested in a study of family’s structures – and we need, for this study, to quantify the number of children in a family.

Population: Family (in a country, city or other strata)

Characteristic: number of children in the family  Discrete (simple) random variable $X = \begin{pmatrix} 0 & 1 & \dots & n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$

The number of children can be 0,1,2,3,... (finite!).

The probabilities cannot be theoretically assumed

In a large sample, these probabilities are the relative frequencies of the variable values

● **Continuous random variables**

The set of values of the random variable is an infinite set of real numbers.

A continuous random variable is given by a function which describes the distribution of its values. This function is called a **probability density function (PDF)** – and it has certain mathematical properties.



DATA TYPES AND MEASUREMENT LEVELS

Depending on the measured/observed/interrogated characteristic, a database can contain the following types of data and measurement levels of the variables:

➤ **Qualitative / categorical data– *non-numeric values***

These variables must be coded in a database, but the codes are just “labels” with a logical signification. They express a “quality”.

Measurement levels:

Ordinal - *variables that have two or more categories (levels) which can be ordered or ranked.*

Nominal - *variables that have two or more categories (levels), but do not have an intrinsic order.*

➤ **Quantitative or metric data (measurement levels: ratio, interval) – *numeric values***

These variables can be measured along a continuum (real interval) and they have numerical values.

Measurement levels – *very important when analyzing datasets (in order to apply the correct inference tool):*

Ratio level (or scale) *is a continuous scale where zero means “does not exist” (and the scale has no negative values).*

Interval level (or scale) *is a continuous scale with a meaning for “zero” and can include negative values.*

Dichotomous variables

Commonly referred as dummy variables - are variables with only two categories (levels). Usually, the values of a dichotomous variable are coded (assigned) as 0 and 1 and the variable is called **“binary”**.

categorical dichotomous variable
(nominal level of measurement)

“Survival status” is a “natural” discrete (nominal) dichotomous variable.

metric dichotomous variable
(ratio level of measurement)

“Exam pass” is a continuous (ratio level) dichotomous variable.

The line between discrete and continuous dichotomous variables is very thin.

Placing dichotomous variables into discrete or continuous categories becomes important when using some advanced statistical techniques. Usually, only nominal dichotomous variables are used in databases.

Care should be taken to place dichotomous variables into their “natural category” - see the examples above – trying to place the “survival status” into a continuous category can mislead the statistical results and interpretation.

Example:

In a database we have records of a population (students) with some characteristics of interest (listed below). Write the correspondent data types, measurement levels and coding.

<i>Characteristic (variable name in database)</i>	<i>Code</i>	<i>Data type</i>	<i>Measurement level</i>
<i>Age</i>	<i>Numbers (values)</i>	<i>metric</i>	<i>ratio</i>
<i>Gender (male, female)</i>	<i>0 1 (or other codes...)</i>	<i>qualitative</i>	<i>nominal (dichotomous)</i>
<i>Political orientation (left, center, right)</i>	<i>1 2 3 (or other codes...)</i>	<i>qualitative</i>	<i>nominal</i>
<i>The interaction with the staff members of the University is good (strongly disagree, disagree, agree, strongly agree)</i>	<i>1 2 3 4 (or other codes...) Likert scale</i>	<i>qualitative</i>	<i>ordinal</i>
<i>Height</i>	<i>Numbers (values)</i>	<i>metric</i>	<i>ratio</i>
<i>Staying with parents (yes, no)</i>	<i>0 1 (or other codes...)</i>	<i>qualitative</i>	<i>nominal (dichotomous)</i>
<i>Eye color (blue, green, brown, black)</i>	<i>1 2 3 4 (or other codes...)</i>	<i>qualitative</i>	<i>nominal</i>
<i>Exam pass (yes, no)</i>	<i>0 1 (or other codes...)</i>	<i>metric OR qualitative</i>	<i>dichotomous (ratio OR nominal)</i>



MEASURES OF CENTRAL TENDENCY

The *central tendency* is a representative value of a theoretical distribution or a data set.

This tendency can be assessed through different measures, the most common being the **mean** (or *average*, or *expected value*).

In probability theory, there are specific formulas for calculating the mean, variance, standard deviation, median of a discrete or continuous variable.

● **The mean**

For a classical theoretical distribution, the mean has a known form;

The mean, as a measure of central tendency, is very sensitive to outliers (extreme values of the distribution)

For a data set with “n” records, the mean is calculated as the arithmetic mean of the variable’s values:

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

In practice, when analyzing data from a sample and making statistical inference based on the available dataset, the mean value should be presented with the correspondent standard deviation (“mean ± sd”)



For qualitative / categorical data in a dataset, can we calculate the mean?

● **The median**

For a theoretical distribution, the median represents the 50th percentile (the quantile for $\alpha = 1/2$)

For a data set, it is the “middle value” and can be calculated by first sorting the values and then either picking the middle value (in the case of an odd number of values), or computing the arithmetic mean of the two middle values (in the case of an even number of values)

The median, as a measure of central tendency, is very robust to outliers (extreme values of the distribution) and can be reported instead of the mean value for statistical inference.

In practice, when analyzing data from a sample and making statistical inference based on the available dataset, the median value should be presented with the correspondent range of values (min-max).

● **The variance (dispersion) and standard deviation**

For a classical theoretical distribution, the variance has a known form;

For a data set with “n” records, the variance should be calculated using the formula:

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

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the sample mean

instead of the (very popular) formula: $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

unbiased estimator

*The **standard deviation (SD)** is defined as the square root of the variance.*

In practice, we do not use the variance when interpreting statistical results because the variance uses the squared values of the variable.

We use the standard deviation which is expressed in the same unit of measurement as the variable and the mean or median.

● The coefficient of variation (CV)

The coefficient of variation is a statistical measure of the dispersion of data points around the mean and is a **unit-free** and **dimension-free** value, regardless of the sample size.

This coefficient is widely used in statistics:

- to describe the dispersion of data (in population or samples)
- to compare the degree of variation of two or more data sets, for a specific characteristic of interest (even if the mean values are quite different)

If the standard deviations are equal, then the data set with a smaller mean will have a greater variation of values.

The **coefficient of variation** is the real number given by:

$$CV = \frac{\sigma}{m}$$

(population)

$$CV = \frac{s}{x}$$

(sample)

another term for this coefficient is
“**relative standard deviation**”

CV is also calculated and used as percentage: $CV = \frac{s}{x} \cdot 100\%$

The coefficient of variation **should only be used to compare positive data on a ratio scale**. It has little or no meaning for measurements on an interval scale.

Also, if the mean value is near zero, the coefficient of variation is sensitive to small changes in the mean. If the mean value is zero, the coefficient cannot be calculated (it has no meaning).



OUTLIERS

Outliers are ***unusual values of a variable***, in a dataset, and they can distort statistical analyses and violate their assumptions.

Outliers *increase the variability* in our data, which *decreases statistical power*.

Causes:

- *data entry or measurement errors*
- *sampling problems and unusual conditions*
- *natural variation (unusual values, but a normal part of the data distribution).*

Identification:

- *in-depth knowledge about all the variables when analyzing data (knowing what values are typical, unusual, and impossible)*
- *visual methods (boxplots, histograms, ascending data-sorting)*
- *coefficient of variation (the percentage formula): if the value exceeds 60-70%*

Decision: *well documented and explained (due to the influence on the study results and inference)*

- *remove the corresponding data points from the dataset*
- *perform the analysis keeping the outliers*
- *perform the analysis replacing the outlier(s) value with the median value*



INFERENCE STATISTICS

Central problem: *estimating population's parameters using a (representative) sample*

*mean, SD, median ...
(quantitative data)*

*proportions ...
(qualitative data)*

There are two types of estimators:

- **Point estimators** – *we determine an estimated value for the studied parameter, based on the parameter value calculated from the selection:*

unbiased estimators for the mean and variance:

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

There are several mathematical methods developed to find point estimators for parameters, including, but not limited to:

*{ the maximum likelihood method
the least square method
the moment method*



implemented in specialized software (for example, MATLAB, SPSS) or have a robust algorithm which can be coded using any programming language

- **Interval estimators** – we determine a set in which the parameter value can be found. These types of estimators are known as **confidence intervals**.

“the interval $[\theta_1, \theta_2]$ contains the value of the parameter θ with a probability of $1 - \alpha$ ”

“the set $[\theta_1, \theta_2] \subset \mathbb{R}$ is the $1 - \alpha$ confidence interval for the parameter θ ”

The value α is called the **significance level or significance threshold**.

The process of obtaining a confidence interval for a parameter is complicated and requires sturdy mathematical knowledge. Confidence intervals for some of the most common distributions have already been determined and implemented in many specialized software (SPSS, STATA, R).

Remarks:

- *The confidence interval differs from one selection to another (because the selection values, on which the statistics depend, are different);*
- *The confidence interval depends on the **selection (sample) size** and **standard error**:*
 - *as the sample size decreases, the confidence interval has a larger range;*
 - *as the standard error increases, the confidence interval has a larger range.*

The *standard error* for a selection with size n and standard error s is defined by:
$$Err = \frac{s}{\sqrt{n}}$$

How do we interpret a confidence interval?

“The confidence interval is the interval in which the true value of the parameter θ can be found with probability $1-\alpha$ ”.

Wrong !

Let's assume we have one selection and we find the confidence interval of level $1-\alpha$ for the parameter θ based on it. Now, we make more selections (from the same population) and compute the empirical value of parameter θ for each selection (keep in mind that the parameter θ is usually the mean or variance, so it can be calculated with ease). In $100 \cdot (1-\alpha)\%$ of the cases, the value we calculate can be found in the interval we determined based on the initial selection.

“Considering multiple selections, in $100 \cdot (1-\alpha)\%$ of the cases, the value of the parameter θ will be included in the confidence interval found based on one selection”.

Correct !

STATISTICAL ANALYSIS

1) PRELIMINARY ANALYSIS

Descriptive statistics

For metric data, check for normality (or other known distributions)

2) COMPARATIVE ANALYSIS

Statistical Tests (parametric, nonparametric)

Other methods (like OR,RR) – nonparametric tests



PRELIMINARY ANALYSIS

Descriptive statistics

For quantitative (metric) data:

Mean \pm SD, calculate CV

Median and range

Outliers (identification, decision)

Check for normality (or other distribution) for metric data

For qualitative data:

Absolute frequency (for samples smaller than 100)

Absolute and / or relative frequency, percentage (for samples larger than 100)

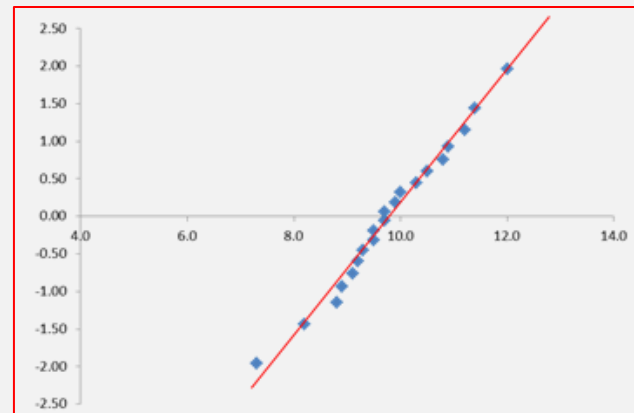
PRELIMINARY ANALYSIS - check for normality (or other distribution) for metric data

● Graphical methods

QQ-plot

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- Q-Q plots are commonly used to compare a data set to a theoretical distribution.
- The points plotted in a Q-Q plot are always non-decreasing when viewed from left to right. If the two distributions being compared are identical, the Q-Q plot follows the line $y = x$. If the two distributions agree after linearly transforming the values in one of the distributions, then the Q-Q plot follows some line, but not necessarily the line $y = x$.

Example: We have registered the values of hemoglobin for 20 patients. We want to check the data for normality (normal or gaussian distribution). In our database are listed the values: **8.8, 9.2, 7.3, 8.2, 9.1, 8.9, 9.3, 9.5, 9.6, 9.7, 9.8, 9.9, 10.0, 10.3, 11.2, 10.8, 10.9, 11.4, 10.5, 12.0.**



Checking the Q-Q plot, we can assume that the data comes from a normal distribution.

PRELIMINARY ANALYSIS - check for normality (or other distribution) for metric data

● Statistical tests

Smirnov-Kolmogorov test

➤ If our data comes from a preset, known distribution, the p -value of the test is greater than 0.05.

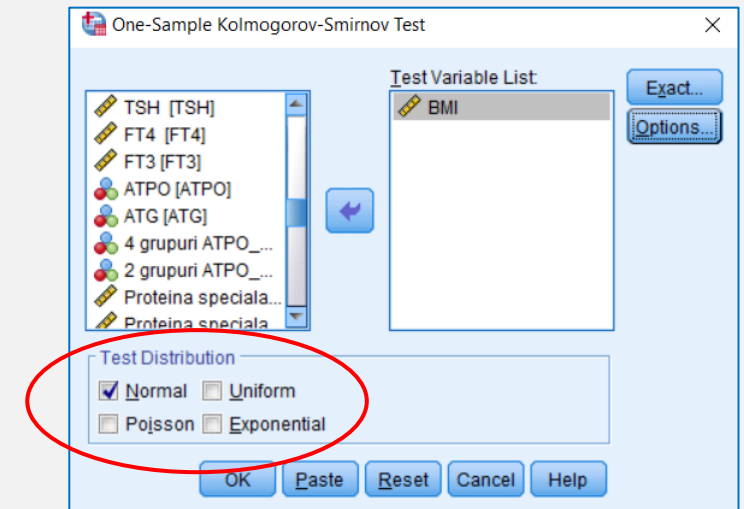
Example: We have registered the values of BMI for 35 patients. We want to check the data for normality (normal or gaussian distribution).

Descriptive statistics for BMI variable (metric, ratio level)

Variable	N	Mean ± SD	Median	Range	p-value
BMI (kg/m ²)	35	22,24 ± 10,02	18,41	12,10 – 46,47	0,001

$$CV = \frac{s}{x} \cdot 100\% = \frac{10.22}{22.24} \cdot 100\% = 45.9\%$$

p -value < 0.05, the BMI data is not normally distributed.



Database in created in SPSS.

For the variable BMI, median and range must be reported, NOT mean ± SD (the mean is not a good measure for the central tendency of this variable). Non-parametric tests should be used in comparative analysis for this variable.



COMPARATIVE ANALYSIS – STATISTICAL TESTS

● **Statistical hypothesis**

- Is a supposition related to *the distribution of one or more variables* from the studied statistical population or to *the values of some parameters* of a known distribution;
- It is studied using a sample (selection) of volume “n” and *a certain test*. Based on these, we accept or reject, with a certain probability, the considered hypothesis.
- In order to test our theoretical hypothesis (supposition), we must formulate two statistical hypothesis:
 - The **null hypothesis** (denoted by H_0) – it states that there are no differences between the studied parameters or that there are no differences between the tested distributions (*it is formulated with “=”*);
 - The **alternative hypothesis** (denoted by H_A) – it states that there is a difference or a relation (*it is formulated with “≠”, “>” or “<”*);

● **Significance level**

The maximum probability we reject the null hypothesis with is called the **significance level** and is denoted by α .

In practice, the most usual values we encounter are $\alpha = 0.05$ and $\alpha = 0.001$, but 0.1, 0.025 or 0.005 are also used.



COMPARATIVE ANALYSIS – STATISTICAL TESTS

● **Testing errors and the power of a test**

There are two possible testing errors:

- **Type I error**, also known as *false positive* – *we reject the null hypothesis when it is true*. In other words, we decide that there is a difference, when, in fact, there are no differences;
- **Type II error**, also known as *false negative* – *we do not reject the null hypothesis when it is not true*. In other words, we decide that there is no difference, when, in fact, there is a difference.

The power of a test - *is the probability of correctly rejecting the null hypothesis when it is false*; it is inversely proportional to the probability of a type II error.

● **Statistical significance and p-value**

The **statistical significance** of a result is the probability that the relation or difference found in the parameters or distribution of the sample also exists in the population from which the selection was made.

The statistical significance is expressed through the **p-value**, which represents *the probability of making a type I error*.

A test result is considered statistically significant if $p \text{ value} < \alpha$, where α is the significance level of the test.



COMPARATIVE ANALYSIS – STATISTICAL TESTS

● **Types of statistical tests – choosing the correct test**

When deciding what statistical test is suitable for our analysis, we shall consider several aspects:

- **Data type and the theoretical distribution of data:**

Parametric tests – for *metric data with normal distribution*;

Nonparametric tests – for *metric data with non-normal distribution* (or unknown distribution) and for *qualitative data*.

- **Number of groups for simultaneous comparison** (2 or more)

- **Groups types:**

Independent groups: each statistical unit (person, object) is only in one condition of the test variable (is only in one group);

Dependent groups (matched, pairs): each statistical unit (person, object) in one group can be paired with an observation in the other group, or the groups are made on the same statistical unit, evaluated on the test variable at different time moments.

- **Number of statistical units in each comparison group**

For small groups (less than 30 units), nonparametric tests should be considered, regardless of the data type and distribution.



T-TEST

● ***It is suitable for :***

- *metric data, with normal distribution;*
- ***2 independent comparison groups;***
- *each group has more than 30 statistical units (cases), relative equal number of cases in each group.*

● ***What is it testing?***

The difference of the means, for the interest variable, between the independent groups.

● ***What is it called?***

Two sample t-test, Student's t-test, independent sample t-test;

● ***What do we read from descriptive analysis, what do we report?***

- *Mean \pm SD for the interest variable in each group;*
- *CI (confidence interval);*
- *p-value of the test.*

● ***What do we interpret?***

Based on the p-value, if there are (or not) significant differences between the means of the interest variable, in our comparison groups.

T-TEST - example

We have a database containing data for 89 patients with peritoneal dialysis. We need to perform a comparison of hemoglobin values of the patients with infectious complications and those with non-infectious complications for a confidence level of 95%.

We have 2 variables involved in this comparison:

Hemoglobin – interest variable

Type of complications – split variable (the criterion for making two groups for analysis)

In the database, “hemoglobin” is recorded as a metric variable (ratio level) and “type of complications” is recorded as nominal dummy variable (with two label codes)

In the preliminary analysis we check for normality the metric variable (Smirnov-Kolmogorov test) and we conclude that we have a normal distribution for this variable (p -value > 0.05).

The conditions for T-test are satisfied and we can report the mean \pm SD for the interest variable, in both groups:

Variable of interest	Infectious complications (n=49)	Non-infectious complications (n=40)	p-value (T-test)
Hemoglobin g/dl	10,91 \pm 2	10,69 \pm 1,4	0,561

Interpretation (statistical and ... clinical)

There is no statistically significant difference of the hemoglobin value ($p > 0.05$, T-test) in the analyzed groups.

Clinical...