INTRODUCTION TO STATISTICS

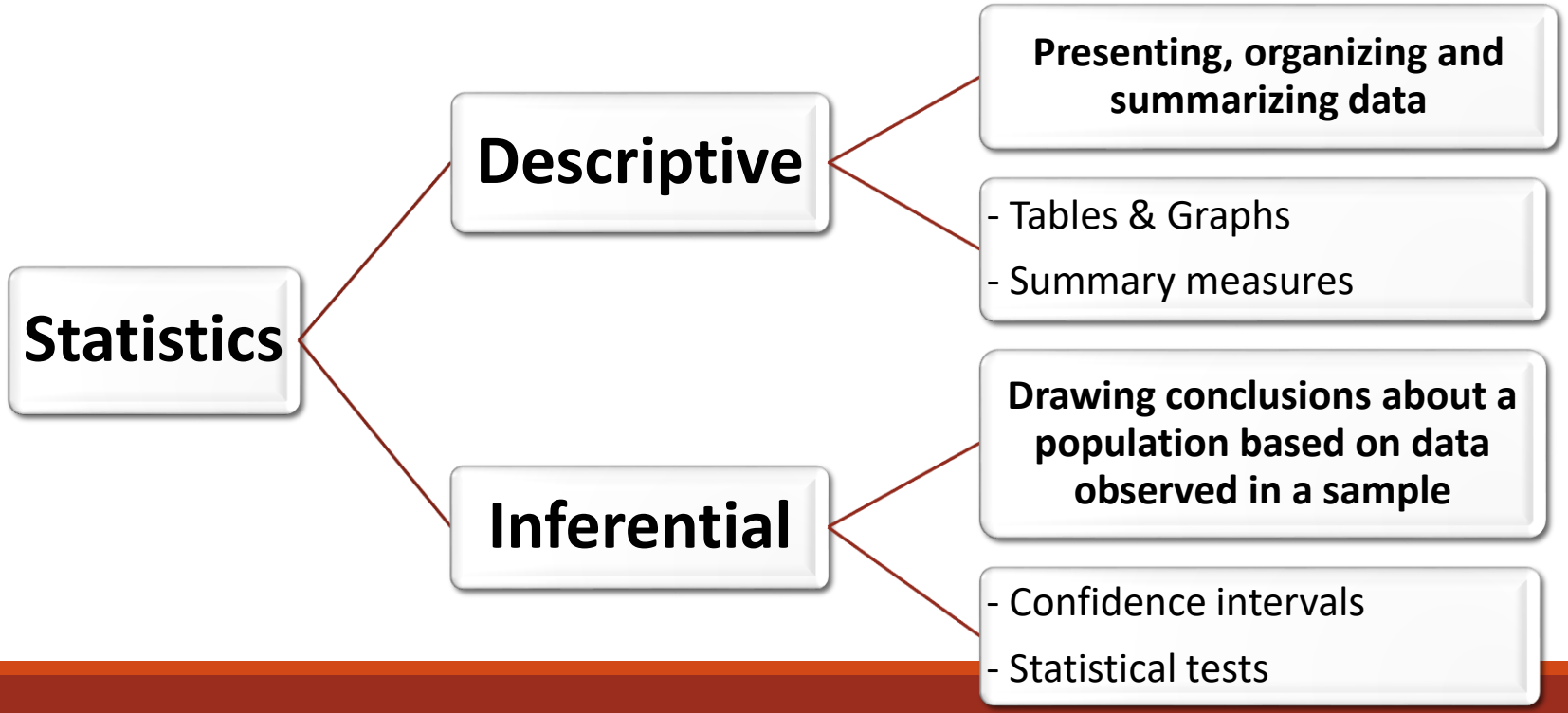
OBJECTIVES

- Types of medical data
- Quantification and accuracy
- Population and sample
- Sampling methods

DEFINITIONS

Statistics:

- Medical statistics deals with applications of biostatistics to medicine and the health sciences, including epidemiology, public health, forensic



Glosary

Variable is a characteristic of interest, a measurable characteristic to which is attributable many different values

- ***Dependent variable*** (also known as response variable) is the variable that is considered to vary depending on the other variable(s) called independent variables (term used in regression analysis).
- ***Independent variables*** are variables that are considered to influence the dependent variable or to explain the variations of the dependent variable in a regression model

Value is a quantitative measure or a characteristic associated to a variable.

Glosary

Variable.

- ***qualitative (categorical) variable*** is a variable with modalities in the form of categories (such as *red*, *orange*, *green*, *blue*, *indigo*, *violet* for the variable ***color***) that can be represented on a **nominal** or an **ordinal** scale.
-
- ***quantitative variable*** is a variable with numerical modalities (such as weight, age, systolic blood pressure, etc.). Quantitative variables could be **discrete** variable (such as the number of medical interventions, number of white blood cells, pulse, etc.) or **continuous** variable (such as length, glycemia, etc.).

Glosary

Population refers to a collection of statistical units of the same nature whose quantifiable information are interested in. The population constitutes the reference universe during the study of a given statistical problem.

Target population = a population to which we would like to apply the results of an analysis.

Glosary

A **sample** is a subset of a population on which statistical studies are made to draw conclude relative to the population.

Sample size (n) is the number of individuals or elements belonging to a sample.

Sampling refers to the selection of part of a population (called sample), the study of certain characteristics *x* of this sample, and the fact of making inferences relative to the population. The sampling methods could be random methods (simple random sampling, stratifies sampling, systematic sampling, cluster sampling) or non-random methods (such as quota sampling).

Glosary

Estimation is the procedure used to determine the value of a particular parameter associated to a population. Two main types of estimators are used in medical statistics: *point estimation* and *interval estimation*.

- ***parameter (point estimation)***
- ***estimator (statistic)***

Variable and medical data

Variable = an entity that can take different values

Data = the value that is taken by a variable for a given patient

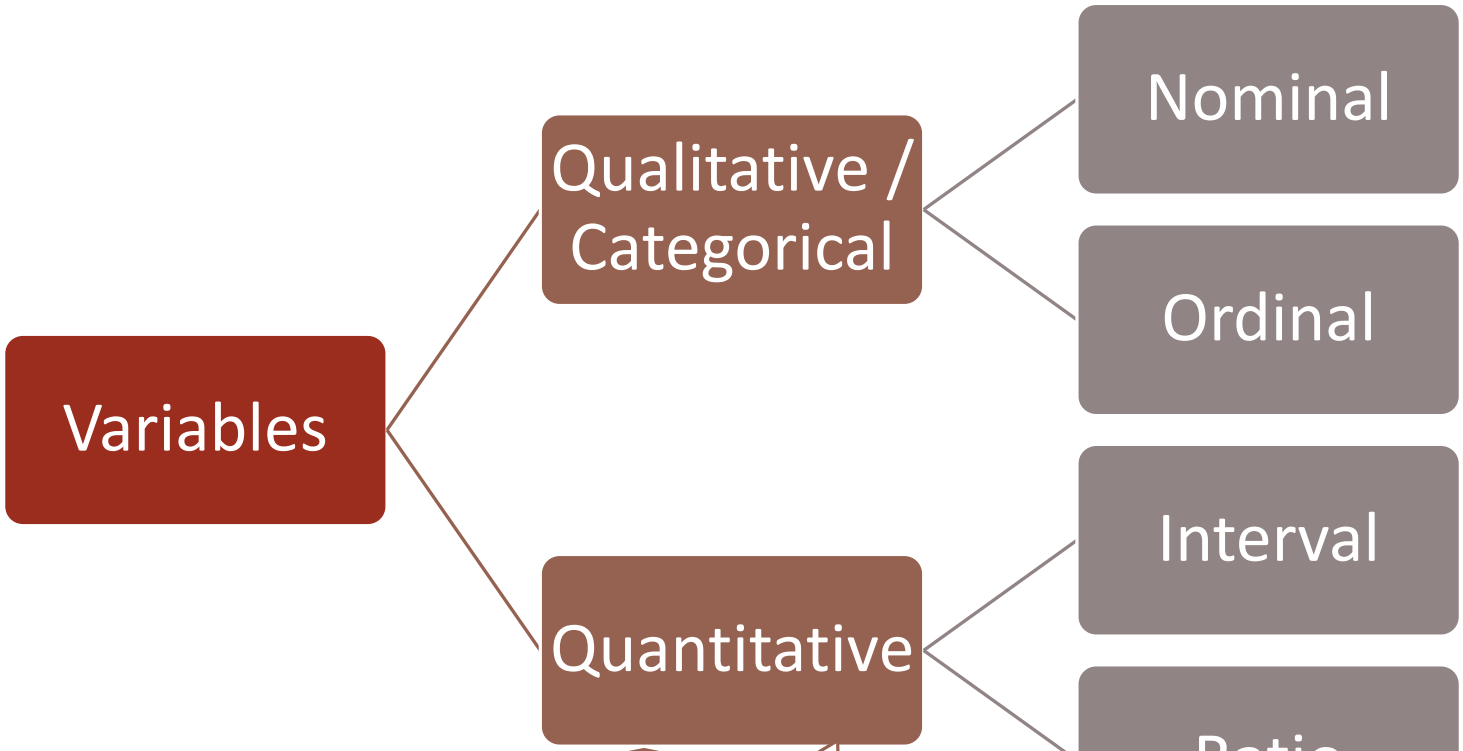
- It is also called statistical unit

Example:

What is the percentage of students smoking in the Faculty of Medicine?

- Variable: Smoking
- Data: the answer of Yes/No type (or number of cigarettes smoked per day) gave by each student

Types of variable



Continuous - take on any value in some range of values

Discrete - takes on a countable number of values

Measurement scale

Nominal	Ordinal
Variabile classified based on particular characteristics in discrete groups The groups could NOT be ordered	Ordered classification based on ranks (from smaller to larger) The distance between ranks is not specified
Interval	**Ratio**
The distance between 2 points on the scale has precise signification	The variable is quantitative continuous and has a true zero

Metric scale or continuous scale

Measurements scales: properties

NOMINAL

- Identity: expressed the membership of an element to a category
- Suppose a classification of the variable without indication an order or a quantity
- Could be noted with numbers (0-feminine; 1-masculine) BUT could NOT be processed regarding quantity or ordered values

ORDINAL

- Data are classified in conformity with an order or preferences
- Could be compared in term of “greater than”, “smaller than”, or “equal”

Measurements scales: properties

INTERVAL

- Quantitative data
- Identity and order
- The distance between two numbers has significance (allows comparison between numbers)
- 0 point is arbitrarily chosen

RATIO

- Quantitative data
- Has a 0 absolute that means the absence of the characteristic or the property

Types of variables- Qualitative / Categorical

- A variable is said to be **nominal** if is classified based on particular characteristics in discrete groups and the groups could NOT be ordered
  - **Dichotomial/Binary** is a qualitative variable which has **only two levels.**

- A variable is said to be **ordinal** if we have an ordered classification based on ranks

Measurements scales: examples

Nominal	Ordinal
Gender: 'Male' and 'Female' Hair color: 'Black hair', 'Brown hair ', 'Auburn hair', 'Red hair', 'Blond hair', 'Gray hair', 'White hair' Education: 'Primary', 'Secondary', 'Higher' Marital status: 'Married', 'Divorced', 'Widowed', 'Single' **Dichotomial / Binary …**	Pain: 'None', 'A bit', 'Quite a lot', 'More than I can bear'.

is a categorical variable which has only two levels.

Measurements scales: examples

Interval	Ratio
Number of lessions The temperature measured in degrees Celsius: ○ Measures of intelligence (IQ): it can not say that a person with an IQ of 150 is twice as intelligent as one with an IQ of 75. ○ Something that is at 20°C is NOT twice as hot as something at 10°C The two values cannot be properly expressed as a ratio as the zero-point on the Celsius scale is arbitrarily chosen (at the freezing point of water).	Weight in kilos (someone who weighs 80 kg is ***twice*** as heavy as someone who weighs 40 kg)

Types of variables - Quantitative

A variable is said to be ***discrete*** if it can assume only a finite or countable infinite number of values.

- Example: number of subject consulted in an Emergency Unit

In contrast, a variable is said to be ***continuous*** if it can assume any value within its range.

- Example: weight, height, etc.

Scale of measurements: transformation

It is possible to transform the interval and ratio scale into ordinal or nominal, BUT this transformation is performed with losing information

- Transformation of the scale associated with the age variable into ordinal scale “classes of age.”

It is NOT possible to transform the nominal or ordinal scale into an interval or ratio scale even if the numbers are attributed to different classes:

- Gender: M = 1, F = 0

POPULATION & SAMPLE

Population & sample

A population contains every member of a defined group of interest:

- All children aged between five and ten with caries living in Cluj-Napoca

A ***population*** is a collection of measurements made on items defined by some characteristic of the items.

Population & sample

A ***sample*** is a subset of the population containing one or more items.

A sample is the section of a population that we investigate it.

- Descriptive statistics are the techniques we use to *describe* the main features of a sample.
  - Example: we described the average number of times the children in the sample brushed their teeth.
- Statistical inference is the process of using the value of a sample statistic to make an informed guess about the value of a population parameter.

Other two basic concepts in statistics

A ***parameter*** is a numerical characteristic of the population; its value is a function of the values of the variables of the items in the population.

- Average of weight in the whole population

A ***statistic*** (singular form) is a numerical measure of a sample; its value is a function of the values of the variables of the items in the sample.

- Average of weight in the sample

Population

Population = a (large) set of entities (items, persons, objects, things, etc.) that have at least a common attribute – form the purpose of a statistical analysis

- Population size = number of the elements of the population
- Statistical unit = an element of the population
- Inclusion criteria
- Exclusion criteria

Population: example

All children aged between five and ten with caries living in Cluj-Napoca

- A particular characteristic (or variable) of the population that we wish to know about is called a population parameter.
- If we want to know how often they brush their teeth, we could ask every child with caries in this age group how often they brush their teeth and calculate the average.
- The average number of times a day that teeth are brushed is thus the population parameter.
- This is impractical, so we study a sample of them.

Sample: why?

Sample = a finite subset of a population.

- A sample usually contains a smaller number of items or subjects as compared to the population.
- We might decide to select 50 children aged between five and ten in Cluj-Napoca with caries and ask them how often they brush their teeth.
- The value of a particular characteristic of a sample is called the sample statistic.

Sample: why?

1. Samples can be studied more quickly than populations.
2. A study of a sample is less expensive than the study of the entire population
3. Sometimes, the all items of population are destroyed in the research process
4. Sample results are often more accurate than results obtained by examination the whole population
5. The correct extraction of the participants in the study of a specific population, the researcher can analyze the sample and make inferences about the feature of the population from which the sample was extracted.

Disadvantages of sampling

The sample could not be representative of the population

Change for over- / under-estimation

…

Sample: characteristics

A sample must be representative for the population regarding:

- Size
- Characteristics

Sample size calculation:

- The risk of rejecting the null hypothesis (H_0) when H_0 is correct = the significance level; alfa, $\alpha = 5\% = 0.05$
- The power of the study (probability of rejecting the null hypothesis when it is true)

Steps in choosing the sample

Define the target population:

- All children aged between five and ten with caries.

Define the accessible population:

- All children aged between five and ten with caries living in Cluj-Napoca

Find the sample size needed to be studied.

Factors in choosing the sample

Accuracy: real value + error

- As the sample volume is higher, the probability of error is smaller.

Costs:

- As the sample volume is higher, the costs will be higher.

Population homogeneity:

- The members of the population are similar from the point of view of studied characteristics.
- The volume of sample size increased as the variability in population increases.

Factors in choosing the sample

Other factors:

- Uncontrollable variables exist.
- We desire to study the sample by groups (we will need a higher sample size)
- We expect a high number of patients lost from observation (we will need a higher sample size)
- We desire a higher statistical power of the results (we will need a higher sample size)

Sample size: empirical rules

Size of population	Size of sample (% from size of population)
0 – 100	100
101 – 1000	10
1001 – 5000	5
5001 – 10000	3
> 10000	1

Sampling methods

Some sampling methods [White C. Sampling in Medicinal Research. Br. Med. J. 1953; (4849):1284–1288] & [Suresh K, Thomas SV, Suresh G. Design, data analysis and sampling techniques for clinical research. Ann Indian Acad Neurol. 2011;14(4): 287–290.]: probability sampling

- Simple random sample
- Systemic random sample
- Stratified random sample
- Cluster sample

Simple random sample

Subjects are randomly extracted from statistics population

Each subject has the same chance to be included in the sample

How:

- Generating random integers:
  - http://www.graphpad.com/quickcalcs/randomN1.cfm
  - http://stattrek.com/Tables/Random.aspx
- Using random tables:
  - http://www.morris.umn.edu/~sungurea/introstat/public/instruction/ranbox/randomnumbersII.html
- Using Excel function: RANDBETWEEN

Systemic random sample

It is selected to be included in the sample each of k^{th} element from the population

The value of k is obtained by dividing the size of the population to the desired sample size

$$\text{sample size} = (\text{population size})/k$$

$$n=N/k$$

- Example:
  - Population size N = 10000
  - Desire sample size n = 1000
  - k = 10000/ 1000 = 10

It is not indicated to be used when a periodicity appear into the population

Stratified random sample

The population is splitting before sampling in relatively homogeneous subgroups called **strata**

The strata are extracted randomly to be included into the sample

Each stratum must be representative in the sample as it is in the population

- Proportionate allocation uses a sampling fraction in each of the strata that are proportional to that of the total population. If the population consists of 60% in the male stratum and 40% in the female stratum, then the relative size of the two samples (three males, two females) should reflect this proportion.
- Optimum allocation: Each stratum is proportionate to the standard deviation of the distribution of the variable.
  - Larger samples are taken in the strata with the greatest variability to generate the least possible sampling variance.

Stratified random sample

Advantages:

- Focuses on important subpopulations and ignores irrelevant ones.
- Allows the use of different sampling techniques for different subpopulations.
- Improves the accuracy/efficiency of estimation.
- Allows greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size.

Cluster sample

A sampling technique used when "natural" groupings are evident in a population.

- The total population is divided into these groups (or clusters)
- A sample of the groups is selected

One version of cluster sampling is area sampling or geographical cluster sampling: Epidemiological Studies

Advantages**:** Can be cheaper than other methods - e.g. fewer travel expenses, administration costs

Disadvantages:

- Higher sampling error ("design effect“): the number of subjects in the cluster study and the number of subjects in the population or another cluster

Sampling methods

Nonprobability sampling [Saumure K, Given LM. Nonprobability Sampling. In: LM Given (Ed). The SAGE Encyclopedia of Qualitative Research Methods. 2008 . Doi:10.4135/9781412963909]:

- Qualitative research
- Participants are chosen because they meet pre-established criteria
- Types:
  - Convenience sampling
  - Snowball sampling
  - Purposive sampling
  - …

Sampling methods

Convenience sampling:

- participants are selected because they are accessible
  - relatively easy for the researcher to recruit.

Snowball sampling:

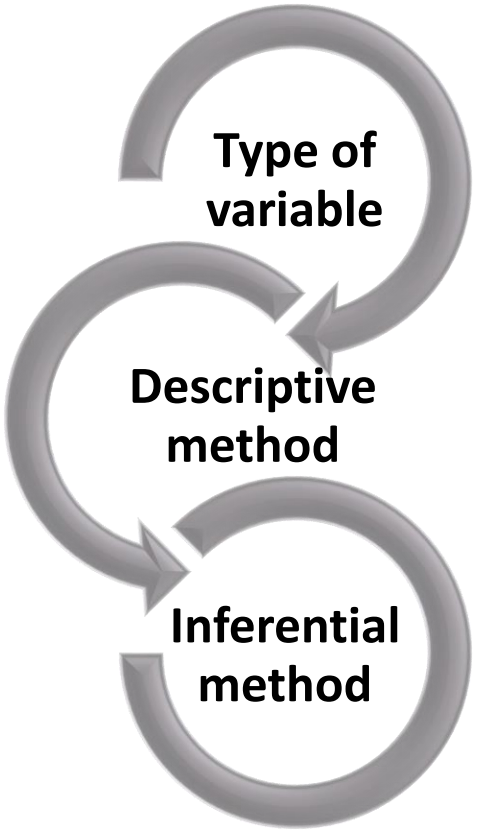
- current participants refer other potential participants to the researcher (e.g., as they are members of the same group or share similar interests, etc.)

Purposive sampling (judgmental, selective or subjective sampling)

- a group of sampling techniques that rely on the judgment of the researcher
- Examples: maximum variation sampling, homogeneous sampling, typical case sampling, extreme case sampling, total population sampling, and expert sampling.

Why does the type of variable matter?

It is essential to assess how you will measure the effect of interest and know how this determines the statistical methods you can use.



Summary

For correct classification of the variable, we need to know the units or measurements or the values that variable can take!

Research is almost never conducted in populations.

Medical research is conducted on samples.

The sample is a subset of the population.

Sampling allows generalizability of results.

QUESTIONS

The manager of a hospital wishes to monitor the satisfaction of patients treated in the hospital. A survey was conducted on 100 randomly chooses patients treated in the hospital during 6 months. What type of variable is “satisfaction scale” (1 = not satisfied at all; 5 = very satisfied)?

1. quantitative
2. qualitative
3. continuous
4. ordinal
5. discrete

QUESTIONS

Consider the next variableS describing a data set of individuals who want to donate blood. Which of the following are dichotomial variables?

Gender (M=1, F=2)	Age (years)	No. of children	Smoker (yes/no)	Higher education (yes/no)
2	25	1	Yes	No
1	35	3	No	Y

1. Gender
2. The number of children
3. Higher education
4. Smoker
5. Age


THANK YOU
FOR YOUR
ATTENTION