# INTRODUCTION TO STATISTICS

## Sorana D. Bolboacă

# OBJECTIVES

- Definitions

- Stages of Scientific Knowledge

- Quantification and Accuracy

- More on Types of Medical Data

- Population and sample

- Sampling

19-Oct-15

# DEFINITIONS

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

- Why?:
  - To provide students with an introduction to statistical <u>techniques</u> and <u>concepts</u> used in <u>research</u>.

19-Oct-15

# STAGES OF SCIENTIFIC KNOWLEDGE

- We gather data because we want to know something
  - The data are useful if they provide information about what we want to know.
- Stages of knowledge development:
  1. Description: describe the medical event
  2. Explanation: explain these events
  3. Prediction: predict the occurrence of these events

19-Oct-15

# STAGES OF SCIENTIFIC KNOWLEDGE

**Description**:

- We seek to describe the data-generating process in cases for which we have data from the process

- Questions by Examples:

  - What is the range of tumor volume for a sample of patients with head and neck tumors?

  - What is the difference in average tumor volume between patients with negative biopsy results and those with positive results?

19-Oct-15

# STAGES OF SCIENTIFIC KNOWLEDGE

**Explanation**: explain these events

- We seek to infer characteristics of the data-generating process when we have only part (usually a small part) of the possible data.

- Question by example:
  - For a sample of patients with head and neck tumors, we can expect the average of tumor volumes of patients with positive biopsy results to be less than those of patients with negative biopsy results for all Romanian patients with head and neck tumors?

- Usually take the form of tests of hypotheses.

# STAGES OF SCIENTIFIC KNOWLEDGE

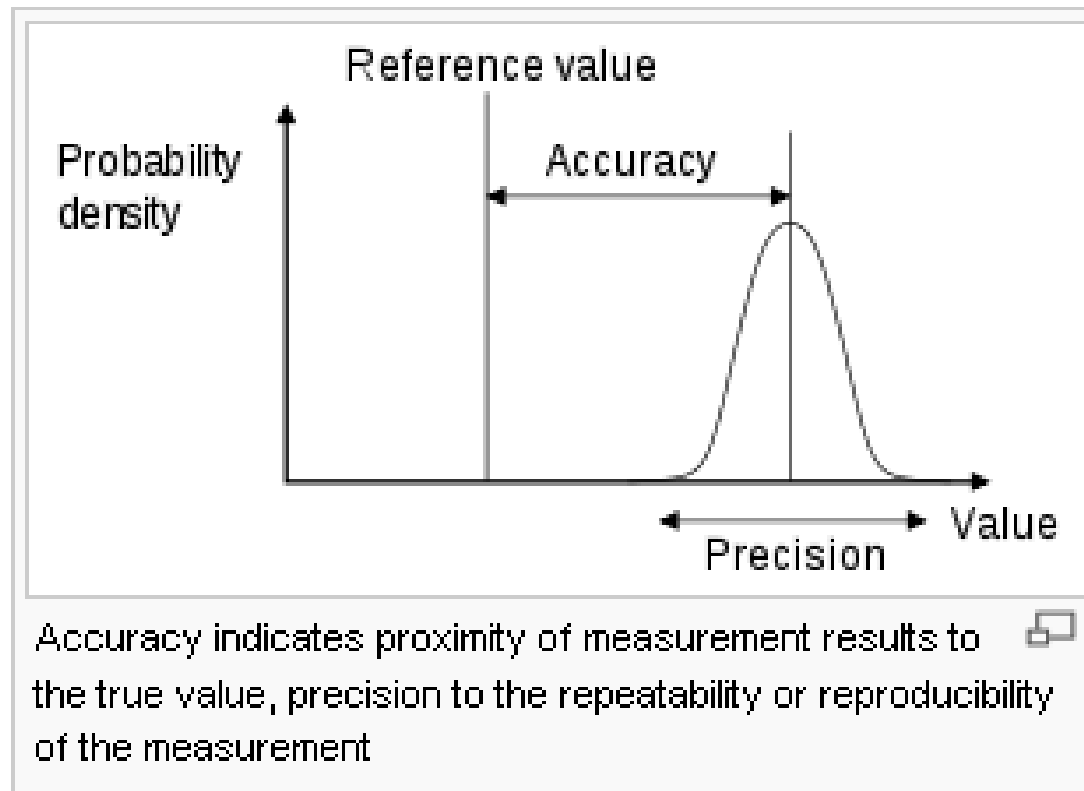**Prediction**: predict the occurrence of these events

- We seek to make predictions about a characteristic of the data-generating process on the basis of newly taken related observations.

- Question by example:
  - On a basis of a patient's negative clinical examination, tumor specific antigen at high level and head or neck tumor volume, what is the probability that he has malign tumor?

- Take the form of a mathematical model of the relationship between the predicted (dependent) variable and the predictor (independent) variable.

# QUANTIFICATION AND ACCURACY

- Knowledge gained from data are more informative and accurate if the data are quantitative.

- Statistics: development of probabilistic knowledge using observed quantities:
  - Counting the number of patients with heart disease
  - Is interested in the process that generate the data

- Quantification:
  - Quantitative form naturally.
  - Measurement base

# MEASUREMENTS

- Accuracy: refers to how well the data-gathering intent is satisfied

- Precision: refers to consistency of measurement



Reference value

Probability density

Accuracy

Precision

Value

Accuracy indicates proximity of measurement results to the true value, precision to the repeatability or reproducibility of the measurement

19-Oct-15

# 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 statistics unit

**Example**:

■ What is the percentage of students smoking in the Faculty of Dentistry?

❑ Variable: Smoking

❑ Data: the answer of Yes/No type (or number of cigarettes smoked per day) obtained from each student
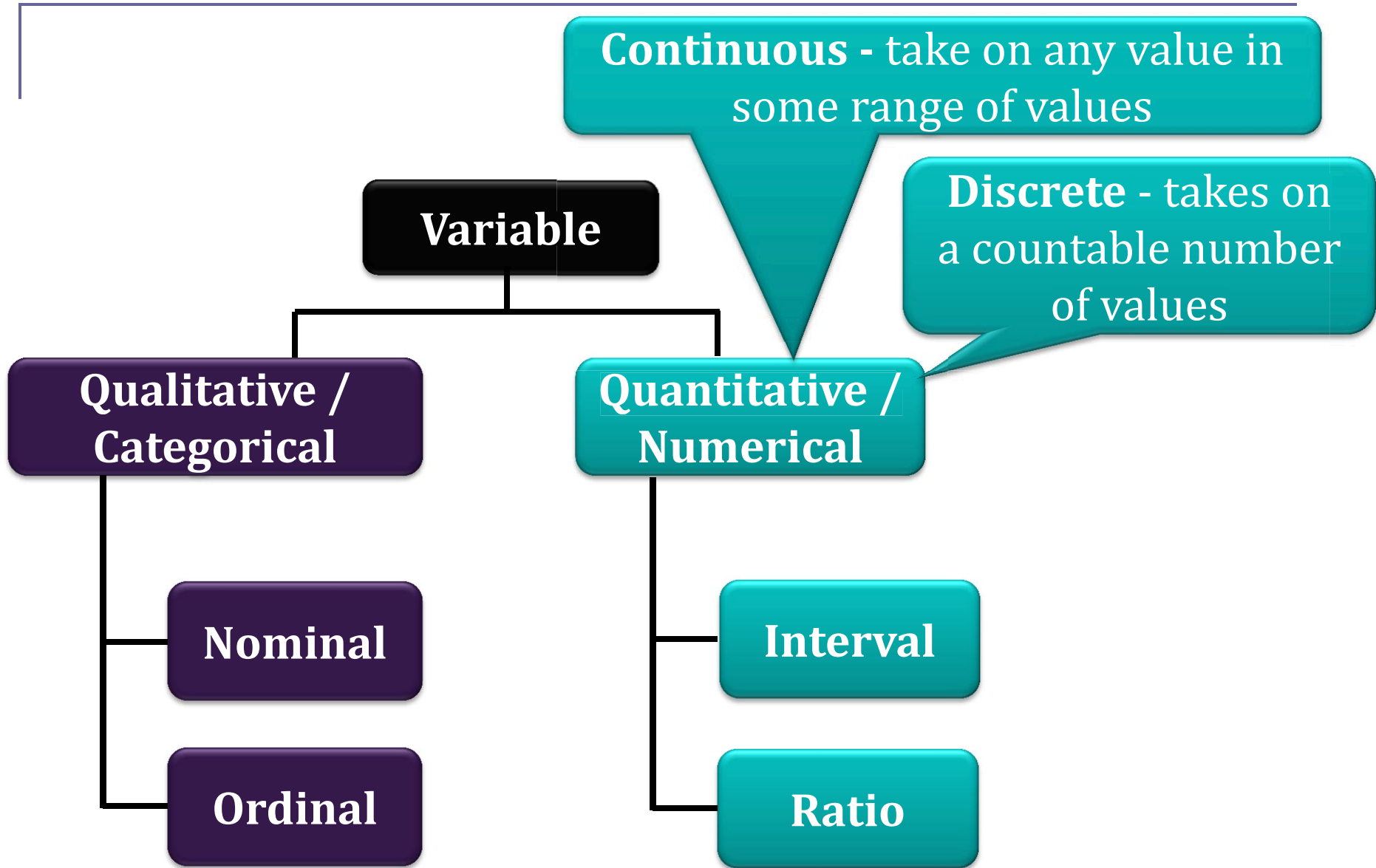
# TYPES OF VARIABLES

## Independent versus Dependent Variables

- Independent (predictors):
  - The variable imposed and controlled by the researcher

- *Example*:
  - The study of three teaching techniques

- Dependente (outcome variable):
  - The variable that is measured in order to demonstrate the effects of independent variables
- *Example*:
  - Accumulated knowledge
  - Attitude in regards of teaching techniques

19-Oct-15

# TYPES OF VARIABLES

## Measurement scale

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

Metric scale or continuous scale

19-Oct-15

# MEASUREMENTS SCALES: PROPERTIES

## Nominal

- Identity: expressed the membership of an element to a category

- Suppose a classification of variable without indication an order or an quantity

- Could be noted with numbers (0-feminine; 1-masculine) BUT could NOT be processed in terms of quantity or ordered values. ordine.

## Ordinal

- Data are classified in conformity with an order or preferences

- Could be compared in term of "greater than", "smaller than", or "equal"

19-Oct-15

# MEASUREMENTS SCALES: PROPERTIES

**Interval**

- Quantitative data
- Identity and order
- Distance between two numbers has significance (allows comparison between numbers)
- 0 point is arbitrary chosen

**Ratio**

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

# MEASUREMENTS SCALES: EXAMPLES

| Nominal | Ordinal |
|---|---|
| Sex: 'Male' and 'Female' | Pain: 'None', 'A bit', 'Quite a lot', 'More than I can bear'. |
| Hair colour: 'Black hair', 'Brown hair ', 'Auburn hair', 'Red hair', 'Blond hair', 'Gray hair', 'White hair' | |
| Education: 'Primary', 'Secondary', 'Higher' | |
| Marital status: 'Married', 'Divorced', 'Widowed', 'Single' | is a categorical variable which has only two levels. |
| Dichotomial / Binary ... | |

19-Oct-15

# MEASUREMENTS SCALES: EXAMPLES

| Interval | Ratio |
|---|---|
| Number of teeth a patient has<br><br>Temperature measured in degrees Celsius:<br><br>■ 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.<br><br>■ Something that is at 20°C is NOT twice as hot as something at 10°C<br><br>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

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

  - Number of subject consulted in a Emergency Unit

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

  - 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 loosing information
  - Transformation of the scale associated to age variable into ordinal scale "classes of age"
- It is NOT possible to transform the nominal or ordinal scale into interval or ratio scale even if the numbers are attributed to different classes:
  - Gender: M = 1, F = 0

19-Oct-15

# MEASUREMENTS

- Measurement = expressing observations in numbers

  - are always an imperfect reflection of the underlying phenomenon:
    - Incompleteness
    - Inaccurate

- How do we assess and improve quality of measurement?

  Measure (x)= true score (X) + error (e)

19-Oct-15

# MEASUREMENT: ERROR

- Error terms may be:
  - Random: Average to 0 & Uncorrelated with X.
  - Systematic: Do NOT average to 0 (bias) & Are correlated with X & Are correlated with other *e* or *X* variables in the system
- Random errors:
  - Lead to ***unreliability*** (the observations are unstable in an unpredictable way)
  - ***Unreliability***:
    - Will decrease with repeated measurement.
    - Can be assessed with repeated measurement.
    - Can be repaired.

19-Oct-15

# MEASUREMENT: ERROR

- Systematic:
  - Invalidity (bias) = lack of systematic error (the size of error can be predicted).

- Systematic (correlated) bias can be estimated and repaired, but we need to measure the source of the bias.

- Multiple measurement can be a big help to fight systematic bias.

19-Oct-15

# RELIABILITY VERSUS VALIDITY

- Reliability = consistency of a measurement procedure

  - If a measurement device or procedure consistently assigns the same score to individuals or objects with equal values, the device is considered reliable

- Validity = a measurement device is valid if it really (and cleanly) measures what it is supposed to measure

  - **Internal validity** of a study refers to the integrity of the experimental design.

  - **External validity** of a study refers to the appropriateness by which its results can be applied to non-study patients or populations.

19-Oct-15

# POPULATION & SAMPLE

# POPULATION & SAMPLE

- A <u>population</u> 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.

19-Oct-15

# POPULATION & SAMPLE

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

- A <u>sample</u> is the section of a population that we actually 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.

19-Oct-15

# OTHER TWO BASIC CONCEPTS IN STATISTICS

- A ***parameter*** is a numerical characteristic of the <u>population</u>; its value is a function of the values of the variables of the items in the population.

  - ❑ Average of weight in whole population

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

  - ❑ Average of weight in the sample

19-Oct-15

# POPULATION

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

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

19-Oct-15

# 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 <u>population parameter</u>.

  - 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 clearly impractical so we study a sample of them.

19-Oct-15

# SAMPLE: WHY?

- Sample = a finite subset of a population.

  - A sample usually contains a smaller number of items or subjects like 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 <u>sample statistic</u>.

19-Oct-15

# 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 process of study destroys the items of population
4. Sample results are often more accurate than results obtained by study the entire population

19-Oct-15

# SAMPLE: WHY?

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.

19-Oct-15

# DISADVANTAGES OF SAMPLING

- The sample could not be representative for the population

- Change for over / under estimation

- ...

19-Oct-15

# SAMPLE: CHARACTERISTICS

- Representative for the population:
  - 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

19-Oct-15

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

19-Oct-15

# FACTORS IN CHOOSING THE SAMPLE

- **Other factors**:
  - ❑ Uncontrollable variables exists
  - ❑ We desire to study the sample by groups (we will need a higher sample size)
  - ❑ We expect a great 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)

19-Oct-15

# 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 |

19-Oct-15

# 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

19-Oct-15

# SIMPLE RANDOM SAMPE

- Subjects are randomly extracted from statistics population

- Each subject has the same chance to be included into 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

19-Oct-15

# SYSTEMIC RANDOM SAMPLE

- It is selected to be included into 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 desire 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

19-Oct-15

# Stratified Random Sample

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

- The strata are extracted randomly to be included into the sample

- Each strata must to be representative in the sample as it is in the population

  - Proportionate allocation uses a sampling fraction in each of the strata that is 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 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
- <u>Advantages</u>: Can be cheaper than other methods - e.g. fewer travel expenses, administration costs
- <u>Disadvantages</u>:
    - Higher sampling error ("design effect"): number of subjects in the cluster study and the number of subjects in the population or other cluster

19-Oct-15

# 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
    - ...

19-Oct-15

# 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.

19-Oct-15

# WHY DOES THE TYPE OF VARIABLE MATTER?

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

**Type of variable**

**Descriptive method**

**Inferential method**

19-Oct-15
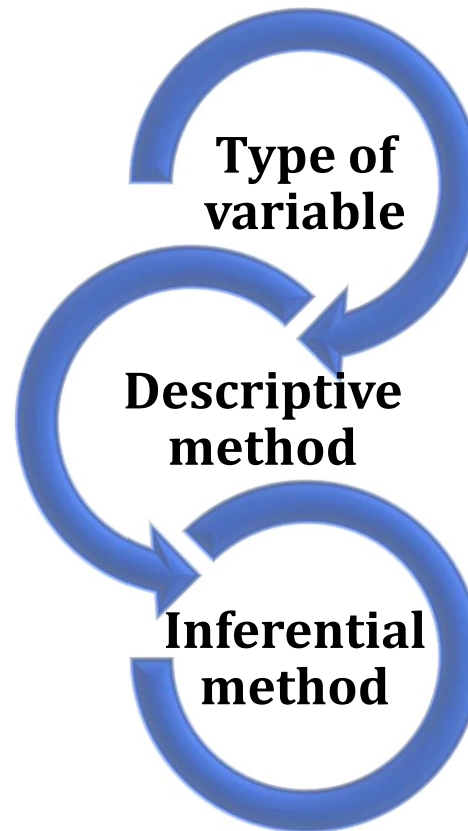
# SUMMARY

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

- Research is almost never conducted in populations

- Medical research is conducted on **samples**

- Sample is a subset of population

- Sampling allow generalizability of results

19-Oct-15

# TASK

- Using the <u>PubMed</u> database identify
  - An article (you will need a full text) in which to fully present the applied randomization procedure.

19-Oct-15

# Thank you for your attention!

19-Oct-15