Introduction to Statistics

Sorana D. BOLBOACĂ, Ph.D., M.Sc., M.D., Lecturer

OUTLINE

- Definitions
- Stages of Scientific Knowledge
- Quantification and Accuracy
- Types of Data
- Population, sample and randomization

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 techniques and concepts used in dental research.

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

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?

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.

Phases of Studies

- **Phase I**: to discover if a treatment is safe and to understand it in order to design formal studies
  - A new drug is assessed to learn the level of dosage and if this level is safe in its main and side effects.
- **Phase II**: preliminary investigation of effectiveness of treatment.
  - Effectiveness = a measure of the benefit resulting from an intervention for a given health problem under usual conditions of clinical care for a particular group.

Phases of Studies

- **Phase III**: large-scale verification of the early findings
  - Is the step from “some evidence” to “proof”
- **Phase IV**: an established treatment is monitored to detect any changes in the treatment
  - Long-term toxicity
  - Evolution of a microorganism to partially immune to a drug that killed the microorganism

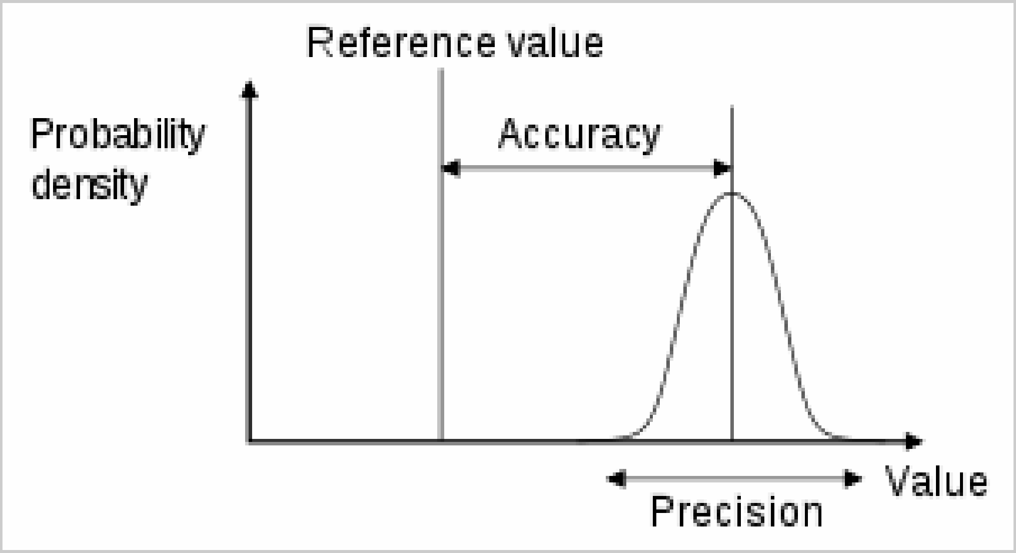
Quantification and Accuracy

- 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



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

Types of 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

Types of Variables

Measurement scale

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

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"

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' 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’ Dichotomial …	Pain: 'None', 'A bit', 'Quite a lot', 'More than I can bear‘.

Measurements Scales: Examples

Interval	Ratio
Number of teeth a patient has 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 Data

Quantitative (metric data)

- Continuous
  - An infinite number of possible values are in the interval of two observed values
- Discrete
  - Integers

Atribute (qualitative / categorical data)

- NOT metric data
- Can take a finite number of values

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 attibuted to different classes:
  - Sex: M = 1, F = 0

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?

$$\text{Measure } (x) = \text{true score } (X) + \text{error } (e)$$

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.

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.

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.

Population, sample and randomization

- 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 sample is the section of a population that we actually study
  - 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.

Elementary Concepts in Statistics

- Statistics population
- Sample
- Unit Statistics
- Variable

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

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

Sample: Why?

- 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.

Sample: Characteristics

- Representative for the population:
  - Size
  - Characteristics
- Sample size calculation:
  - The risk of rejecting the null hypothesis if it is corrected (significance level; alfa, α = 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 population are similar from the point of view of studied characteristics
  - The volume of sample size increased as the variability in population increases
- **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)

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

- Sampling:
  - Random
  - Systemic
  - Stratified
  - Cluster

Random Sampling

- 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

Systemic Sampling

- It is selected to be included into the sample each of kth element from the population
- The value of k is obtained by dividing the size of the population to the desire sample size

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

  - 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 Sampling

- 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 Sampling

- **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 Sampling

- 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“): number of subjects in the cluster study and the number of subjects in the population or other cluster

Tasks

- Using the PubMed database identify
  - An article for each sampling method
  - An article which presents the formula for sample size calculation