General Introduction to Biostatistics

Benchikh Tawfik

Faculty of Medicine

1^{st} Year Medicine

September 28, 2025

Plan du cours

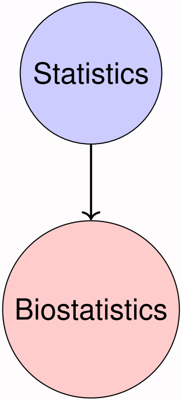
1. General Introduction
2. Basic Statistical Terminology and General Approach

What is Statistics?

- Science of collecting, analyzing, interpreting, and presenting data.

What are Biostatistics?

- Application of statistical methods to biology and health sciences.
- Helps design medical studies, analyze patient data, support decision-making.
- Example: Comparing effectiveness of two treatments.



Objectives of a Statistical Study in Medicine

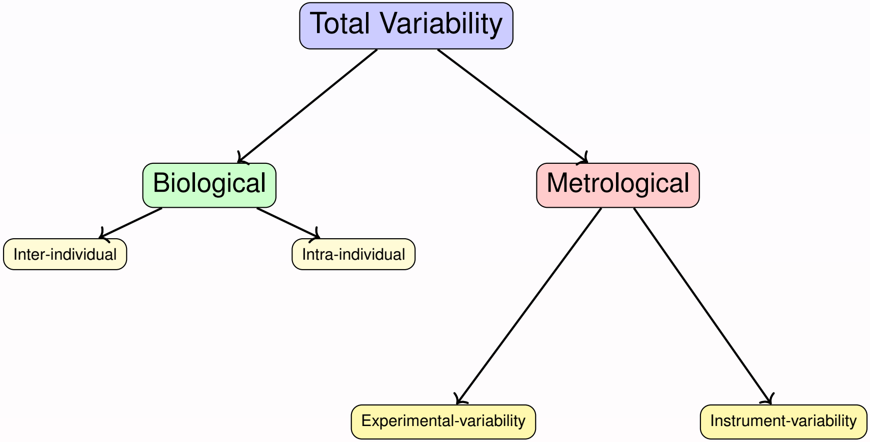
- Identify health problems (e.g., prevalence of diabetes).
- Evaluate treatments (drug A vs. drug B).
- Support public health decisions (vaccination programs).
- Predict outcomes (survival, recovery rates).

Medical Question → Data Collection → Analysis → Conclusion

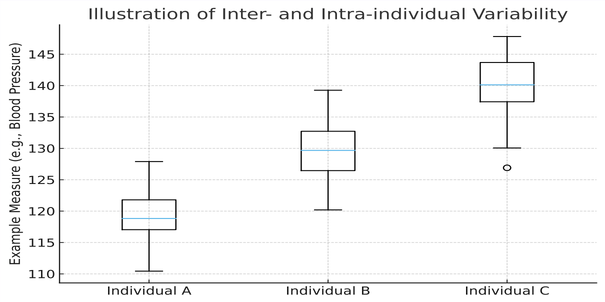
Variability and Uncertainty

- Data are influenced by uncertainty and show variability.
- Causes include: unpredictable outcomes, measurement errors, use of samples instead of full populations, and other random factors.
  - Example: we cannot always know whether a patient has a disease just by observing symptoms.

Sources of Variability



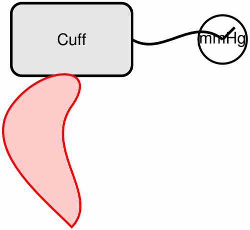
Biological Variability: Examples



- The differences between the three boxplots represent inter-individual variability (each individual has a different typical level).
- The spread within each boxplot shows intra-individual variability (measurements fluctuate for the same person).

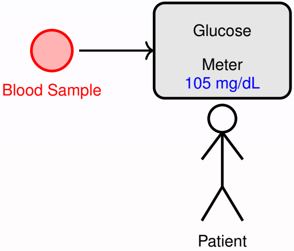
Metrological Variability: Example

- **Example 1:** Blood pressure readings in the same patient may differ depending on:
  - **Experimental conditions:** rest time, body position, stress.
  - **Instrument precision:** accuracy of the blood pressure cuff/gauge (tensiomÃ¨tre/manomÃ¨tre).
- Recommendation: measure after $\geq$15 minutes of rest in a calm setting.



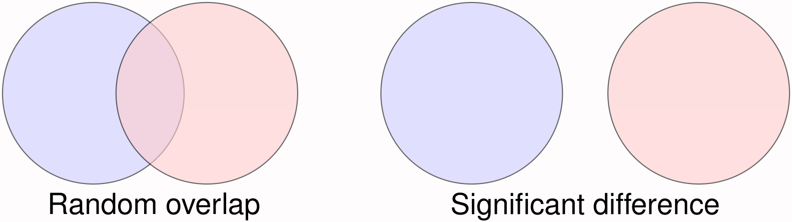
Metrological Variability: Another Example

- **Example 2:** Blood glucose levels in the same patient can differ due to:
    - **Experimental conditions:** fasting vs. after meals, stress, exercise.
    - **Instrument precision:** accuracy of the glucometer (device sensitivity, calibration).
- Recommendation: measure under standardized conditions (e.g., morning fasting).



Random vs. Significant Variability

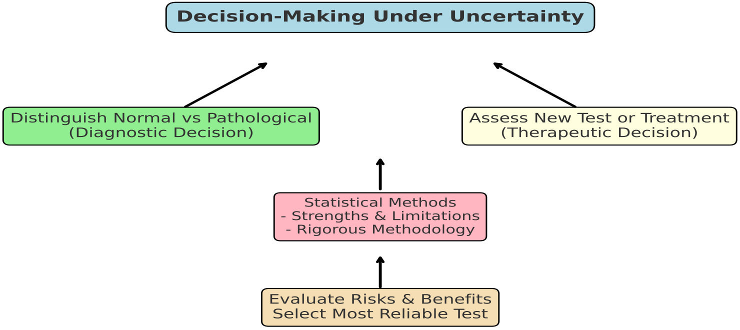
- **Random variability**: natural differences, chance, measurement error.
- **Significant variability**: systematic differences, evidence of a true effect.
- Example:
  - Random: weight differences among healthy adults.
  - Significant: lower blood pressure with Drug A vs placebo.



Decision-Making Under Uncertainty

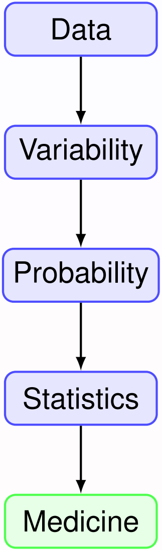
- **Medical decisions** (diagnostic or therapeutic) must account for natural variability:
  - Distinguishing what is **normal** from what is **pathological**.
  - Evaluating the reliability of a test or the effectiveness of a treatment.
- **Statistics in practice:** Understanding both the strengths and limitations of statistical methods is essential for modern medicine.
- **Research foundation:** Every medical research finding is based on clinical or biological experimentation analyzed through rigorous statistical methods.
- **Statistical reasoning:** Helps doctors weigh risks and benefits and choose the most reliable test for the situation

Decision-Making Under Uncertainty



ROLE OF STATISTICS:

- **Role of statistics:** Provides tools to account for uncertainty arising from biological and metrological variability.
- **Implication:** Describing or comparing data requires methods that explicitly handle uncertainty.
- **Foundation:** These methods rely on probability theory.



Conclusion

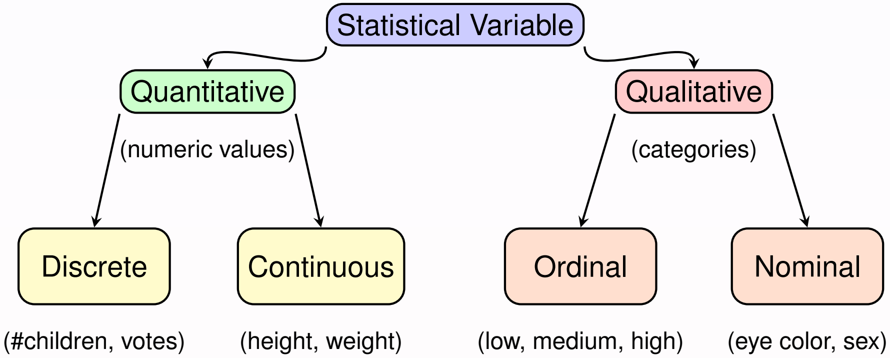
- **Conclusion:** Statistics and probability are indispensable in health and medical research.

Fundamental Definitions

1. **Statistical population** Ω: the complete set of individuals or objects under study, i.e., the collection of homogeneous objects or individuals under investigation.
2. **Individual** or **statistical unit**: a single element of the population ($\omega \in \Omega$).
3. **Sample**: a restricted group or subset drawn from the population, on which the observations are actually carried out.

1. **Sample size** n: the cardinality (number of elements) of the sample.
2. **Statistical survey:** an operation consisting of observing, measuring, or questioning the individuals of a sample.
3. **Census:** survey covering the whole population (an exhaustive survey).
4. **Sample survey:** survey on part of the population (a non-exhaustive survey).

Classification of Variables



Statistical Variable: Definition

- A statistical variable is a characteristic defined on the population (and related to the statistical survey), observed on the sample. Examples: age, income, height, gender, etc.
- Mathematically, it is a mapping defined on the population:

$$X : \Omega \to \begin{cases} \mathbb{R} & \text{if quantitative,} \\ \mathcal{E} & \text{if qualitative.} \end{cases}$$

Notations

- A statistical (or random) variable is denoted by a capital letter X, Y, while its observed values are written with lowercase letters $x_1, x_2, \ldots, y_1, y_2, \ldots$.
- One variable studied $\Rightarrow$ univariate data series.
- Two variables studied $\Rightarrow$ bivariate data series.
- More than two variables $\Rightarrow$ multivariate analysis techniques are required.

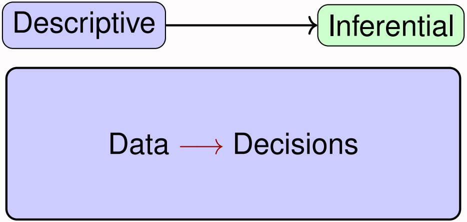
Problem Identification

- The preliminary stage of any statistical study aims to define and identify the problem through a set of guiding questions that help delimit the scope of the investigation and the possible approaches:
  - What are the objectives:
    - What is the population or sample to be studied?
    - What are the characteristics and variables of interest?
    - What insights can a statistical study provide?

General Approach to Descriptive Statistics

Any statistical study can be broken down into at least two main phases:

- **Data collection**: gathering or recording the relevant information.
- **Data analysis and interpretation**, which includes:
  - Descriptive statistics: summarizing and presenting data using tables, graphs, and summary measures (e.g., mean, median, variance).
  - Inferential statistics: draw conclusions about a population from a sample.



Statistical Inference and Statistical Risk

- The core idea of **statistical inference** is to draw **reliable conclusions** about a **population** that cannot be fully observed (often because it is too large).
- These conclusions are based on information from a **random sample**, considered representative of the population.
- Such generalizations always involve **risk of error**, which is quantified and controlled using probability theory.
- Thus, statistical inference is essentially a form of **probabilistic reasoning**.

Inference vs. Deduction

- **Deduction:** From the general to the particular. Predicting the characteristics of samples when the population is already known.
- **Induction (Inference):** From the particular to the general. Predicting the characteristics of an unknown population based on a representative sample.

Modalities: Definition

- The possible values of a qualitative variable are called **categories** (or **modalities**).
- The set of modalities is denoted by

$$\mathcal{E} = \{e_1, \ldots, e_m\}.$$

- Example 1: number of children per family $\mathcal{E} = \{0, 1, 2, \ldots, 9\}$.
- Example 2: eye color $\mathcal{E} = \{\text{green}, \text{blue}, \text{brown}, \text{gray}, \text{black}, \ldots\}$.

Statistical Series: Definition

- A **statistical series** is the list of values taken by a variable X across the observation units.
- The number of observation units is denoted n (or N): the sample size.
- The observed values of X are written

$$x_1, x_2, \ldots, x_i, \ldots, x_n.$$

Statistical Series: Example

- Consider the variable *marital status*, denoted by X, and the statistical series of its values for a sample of 20 individuals. The coding is as follows:

S:	single
M:	married
W:	widowed
D:	divorced

Statistical Series: Example

- The population consists of "individuals";
- The sample contains 20 individuals;
- The statistical variable is "marital status";
- We obtain the following raw statistical series:

M	M	D	S	S	M	S	S	S	M
S	M	W	M	W	D	S	S	S	M

Here, $n = 20$, $x_1 = M$, $x_2 = M$, $x_3 = D$, $x_4 = S$, $x_5 = S$, $\ldots$, $x_{20} = M$.

- The set of possible categories (modalities) of X is $\{S, M, W, D\}$.

Statistical Data

- **Statistical data** = observations collected on a set of individuals (sample) for one or more variables.
- Organized in a data table: rows = individuals, columns = variables.
- Example:

	Sex	Age	Income
1	1	55	20680
2	1	41	46870
3	1	28	12350
4	2	64	19410
5	2	32	24560

Statistical Data: 2

	X_1	X_2	$\dots$	X_j	$\dots$	X_p
individual 1 (ω_1)	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1p}
individual 2 (ω_2)	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
individual i (ω_i)	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
individual n (ω_n)	x_{n1}	x_{n2}	$\dots$	x_{nj}	$\dots$	x_{np}

$$x_{ij} = X_j(\omega_i)\,.$$

Data Collection

- Define the variables to study and the target population.
- Select a **sample**: size, method, and representativeness.
- Establish a **protocol**: who collects the data, how, and when.
- Data can be collected by:
  - **Observation**: recording natural phenomena.
  - **Experimentation**: controlling and provoking conditions.
- Key requirement: the sample must be **representative**, otherwise results are biased.
- To obtain a representative sample, it should be constructed in a random manner-for instance, by performing a true random draw (with replacement).

Summary

- Statistics: general science of data.
- Biostatistics: applied to biology and medicine.
- Steps in medical studies:
  - Population $\longrightarrow$ Characteristics $\longrightarrow$ Sample $\longrightarrow$ Analysis.
- Key: distinguish between random and significant variation.