Inferential statistics

Sampling theory

Benchikh Tawfik

Faculty of Medicine
1st Year Medicine

February 01, 2026

PLAN DE COURS

- **Biostatistical assumption:** the population is much larger than the sample.
  - The goal is to choose a **representative sample**: it should resemble as closely as possible the population from which it is drawn. Individuals in the sample must share the same characteristics as those in the population.
  - Ideally, the sample is selected **at random**: every individual in the population has the same probability of being selected (sampling with replacement, also called non-exhaustive sampling).
- **Remark:** Sampling with replacement leads to **independent events** (a crucial point). If the population is very large or infinite, sampling can always be considered non-exhaustive: whether or not the selected element is replaced becomes negligible.

- **Statistical study:** analysis of characteristics (statistical variables) observed on individuals from a population or a sample.
- Individuals in a population vary $\Rightarrow$ samples also vary. How can this variability be handled?
  - The statistical variables describing the population can be viewed as **random variables**.
  - The distribution of a statistical variable X over the population is described by a **probability distribution**.

Definition of a Random Sample

- We study a characteristic X of a population, finite or infinite.
- The composition of the population with respect to X is completely determined by the probability distribution of X (density f_X or probability mass function p_X).
- Let E be the experiment consisting of randomly selecting an individual from the population. Before sampling, we aim to predict the value of the characteristic X that will be observed.
- Associated with experiment E is a random variable X with distribution f_X (or p_X).

Definition of a Random Sample (continued)

- We repeat the same experiment E independently n times, for instance by sampling with replacement.
- These n experiments define n **independent and identically distributed (i.i.d.) random variables** $X_1, X_2, \ldots, X_n$, all following the same distribution as X, called the **parent distribution**.
- By definition, the collection $(X_1, X_2, \ldots, X_n)$ of i.i.d. random variables is called a **random sample**, denoted briefly by $(X_i)_{i=1}^n$ or $\underline{X}$.
- A realization of the sample is denoted $(x_1, x_2, \ldots, x_n)$ and is called the **raw data set**; it is no longer random.

Purpose of a Random Sample

- **Objectives:**
  - Study of sampling fluctuations of an observed proportion,
  - Study of sampling fluctuations of an observed mean,
  - Estimation: prediction intervals and confidence intervals,
  - Hypothesis testing (goodness-of-fit, homogeneity, conformity to a distribution, etc.).

Example of a Random Sample

- The diameter X of a red blood cell is random, with standard deviation $\sigma = 0.617$, but its mean is unknown.
- The diameters of 200 red blood cells from a patient are measured, yielding a sample mean $\bar{x} = 7.07$.
- Measuring another 200 red blood cells would lead to a different mean.
- Hence, the mean of 200 observations is itself a **random variable**.
- This mean follows a probability distribution, called the **sampling distribution**.

Definition: Sampling Distribution

- **Definition:** The sampling distribution is the probability distribution of the sample means of all possible samples of size n drawn from the population.

DEFINITION OF A STATISTIC

- A **statistic** based on a random sample X of size n is any function of the random variables $X_1, \ldots, X_n$:

$$T(X_1, \ldots, X_n).$$

- Several statistics can be associated with the same sample.
  - The most common characteristics of a sample are the statistics $\overline{X}$ and S^2.
  - We assume that the first and second moments of the parent random variable X exist, and we denote $\mathbb{E}(X) = m = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Definition of a Statistic: Example

- The statistic $\overline{X}$ is defined as

$$\overline{X} = T(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

  which is the arithmetic mean of the n observations.
- The statistic

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

  is the sample variance.
- The unbiased sample variance is defined by

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

THE STATISTIC $\overline{X}$

- The statistic $\overline{X}$ is a function of the random variables X_i, defined by

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

  - A straightforward computation yields the following results:

$$\mathbb{E}(\overline{X}) = m, \qquad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

- **Asymptotic behavior:**
  - **Weak Law of Large Numbers:** The sample mean $\overline{X}$ converges in probability to m as $n \to \infty$.
  - Applying the **Central Limit Theorem** to the random variable Y_n,

$$Y_n = \frac{\sum_{i=1}^{n} X_i - nm}{\sigma\sqrt{n}} = \sqrt{n}\,\frac{\overline{X} - m}{\sigma},$$

we obtain that Y_n converges in distribution to a standard normal random variable $\mathcal{N}(0, 1)$ as $n \to \infty$.

Application: Distribution of a Percentage

- Let K be the random variable representing the number of successes in a sequence of n independent trials, each with probability of success p.
- The random variable K follows a binomial distribution $\mathcal{B}(n, p)$.
- Define

$$F = \frac{K}{n},$$

the empirical frequency (percentage) of successes. Then

$$\mathbb{E}(F) = p, \qquad \mathrm{Var}(F) = \frac{pq}{n}, \quad q = 1 - p.$$

- By applying the Central Limit Theorem for large n, we show that the random variable F converges in distribution to a normal distribution:

$$\mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right).$$

The Statistic S^2

- The sample variance S^2 of a random sample of size n is defined by

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

MAIN PROPERTIES OF THE STATISTIC S^2

- **Expectation of S^2:**

$$\mathbb{E}(S^2) = \frac{n-1}{n}\,\sigma^2 = \left(1 - \frac{1}{n}\right)\sigma^2.$$

  - When the sample size n is large, the expectation of S^2 converges to σ^2.

- **Variance of S^2:**

$$\mathrm{Var}(S^2) = \frac{n-1}{n^3}\left[(n-1)\mu_4 - (n-3)\sigma^4\right],$$

  where μ_4 denotes the fourth central moment of the random variable X.

  - For large n, the variance of S^2 converges to

$$\frac{\mu_4 - \sigma^4}{n}.$$

GAUSSIAN SAMPLES

- If the random variables X_i follow a normal distribution $\mathcal{N}(m, \sigma^2)$, then:
  - The sample mean $\overline{X}$, being a linear combination of n independent Gaussian random variables, is itself Gaussian. Therefore, for any n,

$$\overline{X} \sim \mathcal{N}\left(m, \frac{\sigma^2}{n}\right).$$

  - From the decomposition of the sample variance S^2, we obtain

$$\frac{nS^2}{\sigma^2} \sim \chi^2(n-1).$$

  - If the mean m is known, we use the statistic

$$T = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2,$$

and we have

$$\frac{nT}{\sigma^2} \sim \chi^2(n).$$

REMINDERS

- If $X \sim \mathcal{N}(0,1)$, then $X^2 \sim \chi^2_1$.
- If $X_1, \ldots, X_n$ are independent and identically distributed as $\mathcal{N}(m, \sigma^2)$, then:
  - $\displaystyle\sum_{i=1}^{n} X_i \sim \mathcal{N}(nm, n\sigma^2)$,
  - $\displaystyle\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$,
  - $\displaystyle\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - m)^2 \sim \chi^2_n$,
  - $\displaystyle\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2_{n-1}$.
- The Student and Fisher-Snedecor distributions are typically used through statistical tables or software; explicit density formulas are therefore not required.

Summary

- Why is sampling important in statistical studies?
  - Sampling is crucial because it reduces the cost and time of data collection.
  - Moreover, a well-designed sample can provide reliable results that reflect the population, enabling informed, data-driven decisions.

STATISTICAL APPROACH: A REMINDER

- The statistical approach consists of:
  - drawing a **representative sample** from the population (a random sample) using appropriate techniques; the methods used belong to **sampling theory**,
  - studying the main characteristics of a sample with respect to a given statistical variable X, assumed to arise from a population with a known probability distribution,
  - generating **samples of random variables** in order to validate conclusions, which constitutes the core of **inferential statistics**.

- Probability distributions are generally:
  - **Completely unknown**: nothing is known about the underlying distribution, this leads to **non-parametric inferential statistics** (outside the scope of this course),
  - **Partially known**: the family of distributions is known, but one or more parameters are unknown (e.g., X follows a normal distribution, but neither its mean nor its variance is known), this leads to **parametric inferential statistics**.

Inferential Statistics: Objectives

- **Deduction:** predicting, from a known or assumed population, the characteristics of samples drawn from it.
- **Induction (inference):** inferring the characteristics of an unknown population from statistics computed on a representative sample, that is, extrapolating sample observations to the entire population.

The Three Main Pillars of Parametric Inferential Statistics

- Statistics allow us to:
  - estimate an unknown parameter: **point estimation** (e.g., estimating $\mathbb{E}(X)$ from a sample, such as the mean blood glucose level),
  - provide a range where a parameter is likely to lie: **confidence interval estimation**,
  - make decisions: **hypothesis testing** (e.g., deciding whether the claim $\mathbb{E}(X) = 0.9$ is plausible based on the observed sample).
- The validity of these methods relies primarily on the representativeness of the sample.