Statistical inference

Prof. REMITA Mohamed Riad

National School of Artificial Intelligence.

2024-2025

Statistique inférentielle

1. Sampling
2. Estimation
    a. Point estimates
    b. By Confidence Interval
3. Hypothesis testing
    a. Parametric
    b. Non-parametric

Introduction

Statistical inference consists in inferring the unknown characteristics of a population from a sample drawn from that population. Once the characteristics of the sample are known, they reflect those of the population, within a certain margin of error.

1. Sampling

Introduction

The notion of sampling is associated with a subset (of size n) of individuals drawn from a population. A value is associated with each individual drawn, and the set of values obtained is denoted by $(x_1, \cdots, x_n)$.

- Knowing the value of a parameter (mean, variance, ...), we look for information on the value that can be taken by this parameter. This is the sampling problem.

- We know the value of a parameter in a sample and we're looking for information on this parameter in the population. This is an estimation problem.

So, taking a random sample of size n means considering n realizations of an r.v. X, or considering n independent random variables $X_1, \cdots, X_n$ with the same law as X.

Sampling methods

Sampling is used for several reasons;

- When the population is infinite, only a part of it can be observed.
- Sampling is less expensive than a census.
- There's no other way.

There are several methods for selecting a sample.

Random sampling

1. **Elementary method**
   In a population of size N, where each individual has a probability $\left(\frac{1}{N}\right)$ of being chosen, n individuals are drawn at random. The draws are made by generating n random numbers.
   Advantages: Simple and the sample represents the population well.
   Disadvantages: Requires a good sampling frame (a complete, up-to-date list of all individuals in the population, without repetition), can be time-consuming when generating a large sample.

Random sampling

2. **Elementary method**
   It consists in drawing an individual every $k = \frac{N}{n}$ individuals encountered. Only the first individual is selected by generating a random number between 1 and N.
   Advantages: Selection of a single random number, fast, good sample distribution in the sampling frame.
   Disadvantages: You need a good sampling base.

Random sampling

3. **Stratified sample**
   Individuals are drawn from homogeneous groups within the population, known as strata.
   Advantages: The sample accurately represents each of the population's characteristics.
   Disadvantages: You need to know every characteristic of the population, may be difficult to reach individuals belonging to a small stratum of the population, often costly.

Random sampling

4. **Cluster Sampling**
   The population is divided into heterogeneous clusters of similar sizes. The sample size n is decided, and then the number of clusters required is determined. Finally, the desired number of clusters is selected using simple random sampling.
   Advantages: Reduces travel and costs when the population is spread over a large area.
   Disadvantages: If the clusters are homogeneous, the sample produced will not accurately represent the population.

Empirical Sampling

This method is characterized by the closest possible resemblance to the population, achieved through prior knowledge of the population's composition. Examples include:
Blindly Sampling: Individuals are selected in a completely arbitrary (non-random) manner.
Voluntary Sampling: Individuals are selected by calling for volunteers.

Empirical Sampling

Quota Sampling: This method involves creating a sample of size n, where the proportions of individuals match those of the population. The selection of individuals for the sample is not random.

Snowball Sampling: This method begins by arbitrarily selecting a small group of individuals who possess the characteristics required for the study. These individuals are then asked to identify others in their network who share the same characteristics. These new participants, in turn, select others in the same manner, and the process continues until the sample reaches the desired size.

Determining sample size

For the sample to accurately reflect the characteristics of the population to be represented, it must include a certain number of individuals. To calculate the ideal sample size, we need to define a set of values.

- Specify the population size N;
- Define the margin of error e;
- Define the confidence level c;
- Determine the standard deviation σ;
- Set the Z-score z as a function of the confidence level ($c = 90\% \Longrightarrow z = 1,645; c = 95\% \Longrightarrow z = 1,96; c = 99\% \Longrightarrow z = 2,326$);

Determining sample size

If the population is small or medium-sized and we know all the important values, we use the standard formula

$$n = \frac{\frac{z^2 \cdot \sigma^2}{e^2}}{1 + \frac{z^2 \cdot \sigma^2}{N \cdot e^2}} \left(= \frac{\frac{z^2 \cdot p(1-p)}{e^2}}{1 + \frac{z^2 \cdot p(1-p)}{N \cdot e^2}} \text{ if } p \text{ is a proportion} \right).$$

If the population is very large or unknown, we use the formula

$$n = \frac{z^2 \times \sigma^2}{e^2}. \left(= \frac{z^2 \cdot p\,(1-p)}{e^2} \text{ if } p \text{ is a proportion} \right)$$

Slovin's formula is a very general equation used when estimating the size of a population,

$$n = \frac{N}{1 + N \cdot e^2}.$$

1. Estimation

Introduction

This time, the aim is to estimate certain statistical characteristics of the law through a series of observations $x_1, x_2 \cdots, x_n$.
From the characteristics of a sample, what can we deduce about the characteristics of the population from which it is drawn?
Estimation consists in giving approximate values to the parameters of a population, using a sample of n observations from that population. The exact value may be wrong, but the "best possible value" that can be assumed is given.
Estimation problems fall into two categories:
- Point estimation: based on the information provided by the sample, gives a single value for the parameter.
- Confidence interval estimation: involves constructing an interval within which the parameter lies with a given probability.

Définitions

Definition

Let X be a r.v. on a space $(\Omega, \mathcal{F}, \mathbb{P})$. A sample of X of size n is a $n-$tuplet $(X_1, \cdots, X_n)$ of independent r.v.with the same distribution as X which will be referred to as the mother distribution. A realization of this sample is a $n-$tuplet $(x_1, \cdots, x_n)$ where $X_i(\omega) = x_i$.

Definition

We call statitic on a $n-$sample a function of $(X_1, \cdots, X_n)$.

Moyenne empirique

Definition

The sample's mean or empirical mean is the statistic noted $\overline{X}$ and defined by

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

Remarque

For a realization $(X_1, \cdots, X_n)$, *the statistic* $\overline{X}$ *will take the value* $\overline{x} = \frac{1}{n} \sum_{i-1}^{n} x_i$ *(which is the arithmetic mean as we know it). For another realization, under the same conditions, a second sample will yield the realization* $(x'_1, \cdots, x'_n)$, *and* $\overline{X}$ *will take the value* $\overline{x'} = \frac{1}{n} \sum_{i-1}^{n} x'_i$.

Empirical mean

Proposition

Let X be a r.v. with mean μ and standard deviation σ. We have

$$\mathbb{E}\left[\overline{X}\right] = \mu, Var\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

Furthermore, by the Central Limit Theorem, $\overline{X}$ converges in distribution to $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ as n goes to infinity.

Empirical mean

Remarque

The variance of $\overline{X}$ is calculated for the case of a sample of i.i.d. random variables (a sample drawn with replacement from a finite population or a sample drawn with or without replacement from an infinite population).
If the sample is drawn without replacement from a finite population (exhaustive sampling), the random variables are no longer independent. In this case, we have: $Var\left(\overline{X}\right) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$, where $\frac{N-n}{N-1}$ is called the finite population correction factor (or exhaustivity factor).

Empirical variance

Definition

We call empirical variance, the statistic noted $\widetilde{S}^2$ defined by

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

Proposition

Let X be a r.v. with standard deviation σ and a centred moment of order 4, μ_4. We have

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

Distribution of frequencies

Let $(X_1, \cdots, X_n)$ a random sample of size n and following a Bernoulli distribution with parameter p as mother distribution.
Then,

$$F = \frac{X_1 + \cdots + X_n}{n}$$

is the frequency of the value 1 in the sample and nF follow the binomiale with parameters n and p.
Thus

$$\mathbb{E}[F] = p, Var(F) = \frac{pq}{n}.$$

Proposition

If the draw is made without replacement, we have

$$Var(F) = \frac{pq}{n}\frac{N-n}{N-1}.$$

Point estimation

We aim to estimate a parameter θ of a population (which could be its mean, standard deviation, or a proportion p). An estimator of θ is a statistic T, whose realization is considered as a possible value of the parameter θ. The estimation of θ associated with this estimator refers to the observed value during the experiment, i.e., the value taken by the function at the observed point $(x_1, x_2 \cdots, x_n)$.

Point estimation

Quality of an estimator

Definition

An estimator T is said **convergent** if T converges in probability to θ when n goes to infinity.

Theorem

If T is convergent and with variance going to 0 when n goes to infinity then T is said **efficient**.

Point estimation

Quality of an estimator

Definition

We define the **bias** of T for θ as the value $b_\theta(T) = \mathbb{E}[T] - \theta$.
An estimator T is said to be **unbiased** if $\mathbb{E}[T] = \theta$.
We say that the T is an asymptotically unbiased estimator if:

$$\lim_{n \longrightarrow \infty} b_\theta(T) = 0.$$

Point estimation

Quality of an estimator

Definition

Let T be an estimator of a parameter θ with distribution $\mathbb{P}_\theta$ of an observed random variable X. We suppose that there exist two functions $a = a(\theta, n)$ and $b = b(\theta, n)$ such that:

$$\lim_{n \longrightarrow \infty} \frac{T - a}{b} \sim N(0, 1).$$

We then say that T is an asymptotically normal estimator.

Point estimation

Quality of an estimator

The quality of an estimator is also measured by the **mean squared error** (or quadratic risk), defined as $\mathbb{E}\left[(T-\theta)^2\right]$.

Theorem

Let T an estimator of the studied parameter θ. We have

$$\mathbb{E}\left[(T-\theta)^2\right] = Var\,(T) + (\mathbb{E}\,[T]-\theta)^2\,.$$

Remarque

Between two unbiased estimators, the better one is the one with the minimal variance. The estimator with the smallest variance is said to be more efficient.