2. 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 $4, \mu_4$. We have

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

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

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

Proposition

If the draw is made without replacement, we have

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

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 $(x_1, x_2 \cdots, x_n)$.

Definition

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

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:

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

Definition

An unbiased estimator T verifying the equality

$$Var(T) = \frac{1}{nI(T)} \text{ where } I(T) = \mathbb{E}\left[\left(\frac{\partial \ln L(X_1, \cdots, X_n, T)}{\partial T}\right)^2\right]$$

is said to be **efficient**. The function $I(T)$ is called Fisher information of the estimator T and L is the likelihood.

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.

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.

- $\overline{X}$ is an unbiased estimator of the mean μ. Its estimation $\overline{x}$ is the observed mean in a realisation of the sample.
- $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is a biased estimator of σ^2.
- $\widetilde{S}^2 = \frac{n}{n-1}S^2$ is an unbiased estimator of σ^2. Its estimation is $s^2 = \frac{n}{n-1}s_e^2$ where s_e^2 is the observed variance in a realisation of the sample.
  If the mean μ of X is unknown, $T = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$ is better estimator of σ^2 than S^2.
- If p is the frequency of a character, F is an unbiased estimator of p. Its estimation is noted f.

Maximum likelihood method

The distribution of the random vector $(X_1, \cdots, X_n)$ is called the sample likelihood, denoted $L(x_1, \cdots, x_n)$. The purpose of the maximum likelihood method is to choose the most likely value for estimating θ la valeur le plus vraisemblable. The likelihood function is denoted by $L(x_1, \cdots, x_n; \theta)$.
The maximum likelihood estimator is given by the maximum of the likelihood function

$$L(x_1, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

where $f(x, \theta)$ is the distribution of the population.

Maximum likelihood method

The maximum is obtained by cancelling the derivative of this function

$$\frac{dL\left(x_1, \cdots, x_n; \theta\right)}{d\theta} = 0$$

or by canceling the derivative of its logarithm

$$\frac{d\left[\ln L\left(x_1, \cdots, x_n; \theta\right)\right]}{d\theta} = 0.$$

Maximum likelihood method

Example

In a population, consider an r.v. $X \rightsquigarrow \mathcal{P}(\lambda)$. We want to estimate λ.
To do this, a sample of size n is drawn. Assuming $n = 6$ and the realization is $(0, 2, 2, 3, 1, 2)$, find the estimate of λ by this method.

Example

On souhaite estimer les paramètres et d'une loi normale à partir d'un n-échantillon.

Interval estimation

Definition

Point estimation gives a parameter θ to be estimated a unique value which gives a slightly different estimate of the parameter to be estimated, even if it is unbiased. It would be interesting to construct an interval $[a, b]$ in which the parameter θ lies with a given probability.

To determine this interval, we give ourselves a confidence level denoted $1-\alpha$. The value α measures the probability that the value of θ does not lie within the interval $[a, b]$. We will calculate the bounds of the interval, called confidence limits, in such a way that $\mathbb{P}\left(a \leq \theta \leq b\right) = 1-\alpha$.
The interval $[a, b]$ is called the confidence interval.

Confidence interval of a proportion

It is assumed that the draw is random and that the sample size n is large $(n \geq 30)$. In the population, a proportion p of individuals possess a certain characteristic. A confidence interval for p is sought from the value f_n : frequency of individuals possessing the characteristic in the sample. We know that the variable $X = nF_n$ follows a binomial distribution $\mathcal{B}(n, p)$ and as n is large we have $\frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow \mathcal{N}(0,1)$. We have

$$\mathbb{P}\left(-u_{1-\frac{\alpha}{2}} \leq \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \leq u_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

Confidence interval of a proportion

hence

$$f_n - u_{1-\frac{\alpha}{2}} \sqrt{\frac{p\,(1-p)}{n}} \leq p \leq f_n + u_{1-\frac{\alpha}{2}} \sqrt{\frac{p\,(1-p)}{n}}$$

we note that the bounds contain p, which is to be estimated. To do this, we simply replace p by f_n and the confidence interval is then written as follows

$$f_n - u_{1-\frac{\alpha}{2}} \sqrt{\frac{f_n\,(1-f_n)}{n}} \leq p \leq f_n + u_{1-\frac{\alpha}{2}} \sqrt{\frac{f_n\,(1-f_n)}{n}}.$$

Confidence interval of a mean

Known σ

If the distribution of the a.v. X is normal, or if X follows any distribution with large n $(n \geq 30)$, we can say that $\overline{X}$ follows $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. The confidence interval is given by

$$\mathbb{P}\left(-u_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}\right) = 2\Phi\left(u_{1-\frac{\alpha}{2}}\right) - 1,$$

this means that $\Phi\left(u_{1-\frac{\alpha}{2}}\right) = \frac{1+(1-\alpha)}{2}$, where Φ is the cumulative function of the distribution $\mathcal{N}(0,1)$.

Then the confidence interval is $\left[\overline{X} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \overline{X} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$.

Confidence interval of a mean

If we take $\alpha = 0,05$ we get $\Phi\left(u_{1-\frac{\alpha}{2}}\right) = \frac{1+(1-0,005)}{2} = 0,975$. The table gives $u_{1-\frac{\alpha}{2}} = 1,96$. We obtain then

$$\mathbb{P}\left(\overline{X} - 1,96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95$$

Confidence interval of a mean

If we take $\alpha = 0,05$ we get $\Phi\left(u_{1-\frac{\alpha}{2}}\right) = \frac{1+(1-0,005)}{2} = 0,975$. The table gives $u_{1-\frac{\alpha}{2}} = 1,96$. We obtain then

$$\mathbb{P}\left(\overline{X} - 1,96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1,96\frac{\sigma}{\sqrt{n}}\right) = 0,95$$

hence the confidence interval

$$\overline{x} - 1,96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + 1,96\frac{\sigma}{\sqrt{n}}.$$

Confidence interval of a mean

Unknown σ (any population with large n or normal population) In most cases, when μ is unknown in a population, σ is also unknown. To estimate the parameter $\theta = \mu$, the previous relationship is no longer valid. We use the r.v. $T = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n-1}}} \rightsquigarrow \mathcal{T}_{n-1}$ (Student with $n-1$ degrees of freedom). We obtain then

$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n-1}}} \leq t_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

where $t_{1-\frac{\alpha}{2}}$ is read from the Student table with $n-1$ degrees of freedom.

This gives us the confidence interval

$$\overline{x} - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}} \leq \mu \leq \overline{x} + t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n-1}}.$$

If n is large $(n \geq 30)$ we can replace $t_{1-\frac{\alpha}{2}}$ by $u_{1-\frac{\alpha}{2}}$.
In the case of a random draw, the standard deviation of $\overline{X}$ is $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ and we replace $\frac{s}{\sqrt{n-1}}$ by $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ in the confidence interval.

Confidence interval of a mean

Example

The average height of a random sample of 40 people taken from a population of 780 is $1.70m$. The standard deviation for the whole population is $24cm$. Find the 95% confidence interval for the mean height of the population.

Example

500 students sit an exam. A random sample of 38 marks gives a mean equal to 8.65 and a standard deviation equal to 2.82. Find the confidence interval for the population mean scores at $90\%, 95\%$ and 99%.

Confidence interval of a variance

The population distribution is assumed to be normal. the r.v. $\frac{nS^2}{\sigma^2}$ follows the χ^2_{n-1} distribution. Let's determine the confidence interval from $\mathbb{P}\left(s_1^2 \leq \sigma^2 \leq s_2^2\right) = 1 - \alpha$.
Let's consider a and b as the limits of the interval such that $\mathbb{P}\left(a \leq \frac{nS^2}{\sigma^2} \leq b\right) = 1 - \alpha$, we deduce that
$s_1^2 = \frac{nS^2}{b} \leq \sigma^2 \leq \frac{nS^2}{a} = s_2^2$.
We then look for s_1^2 and s_2^2 such that

$$\mathbb{P}\left(\sigma^2 \leq s_1^2\right) = \mathbb{P}\left(\sigma^2 \leq \frac{nS^2}{b}\right) = \mathbb{P}\left(b \leq \frac{nS^2}{\sigma^2}\right) = \frac{\alpha}{2}$$

and

$$\mathbb{P}\left(\sigma^2 \geq s_1^2\right) = \mathbb{P}\left(\sigma^2 \geq \frac{nS^2}{a}\right) = \mathbb{P}\left(a \geq \frac{nS^2}{\sigma^2}\right) = \frac{\alpha}{2},$$

the a and b values are determined by reading the χ^2 table.