4 COURSE

Estimation

1 Estimation

4.1.1 Estimators

Definition 2.1

To **estimate a parameter** means to find an approximate value based on the results obtained from a sample.
An **estimator** $\hat{\theta}$ of the unknown parameter θ is a function that assigns to each set of observations $(x_1, ..., x_n)$ an estimated value $\hat{\theta}$:

$$\hat{\theta} : (x_1, ..., x_n) \longmapsto \hat{\theta} = f(x_1, ..., x_n)$$

Hence, $\hat{\theta}$ is a random variable. We can compute its expectation $E(\hat{\theta})$ and variance $Var(\hat{\theta})$. These quantities measure the quality of the estimator for the parameter θ.

Example 2.1

Estimating the average height of a population from the empirical mean of a sample taken from that population.

Definition 2.2

An estimator $\hat{\theta}$ is said to be **unbiased** if the mean of its sampling distribution equals the true value of the parameter θ:

$$E(\hat{\theta}) = \theta.$$

Otherwise, it is said to be **biased**.
The **bias** of an estimator is defined as:

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

Remark

The absence of bias does not necessarily imply that an estimator is efficient. A parameter can have multiple unbiased estimators. In such cases, efficiency is compared using their variances: an estimator with smaller variance provides estimates closer to the true value of θ.

Definition 2.3

An unbiased estimator $\hat{\theta}_1$ is said to be **efficient** if, for any other unbiased estimator $\hat{\theta}_2$:

$$E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta \quad \text{and} \quad Var(\hat{\theta}_1) < Var(\hat{\theta}_2).$$

Definition 2.4

An estimator $\hat{\theta}$ is said to be **consistent** (or convergent) if its distribution tends to concentrate around the true value θ as the sample size increases, i.e.:

$$\lim_{n \to +\infty} Var(\hat{\theta}) = 0.$$

Common Estimators

(A) Quantitative Characteristic Let X be a random variable with mean μ and standard deviation σ defined on a parent population Ω, and let $(X_1, ..., X_n)$ be a random sample.

Properties

1. $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an unbiased and consistent estimator of μ $(E(\overline{X}) = \mu)$.
2. $V = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$ is a biased estimator of σ^2.
3. $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n}{n-1} V$ is an unbiased and consistent estimator of σ^2.

(B) Qualitative Characteristic

Property

For a qualitative characteristic with population proportion p, the sample proportion F is an unbiased and consistent estimator of p.

4.1.2 Confidence Intervals

Definition 2.5

Rather than determining a single approximate value of a parameter θ, we may seek an **interval** that contains the true value of θ with a specified probability.
Let X be a random variable whose distribution depends on the parameter θ. A **confidence interval of risk** α for θ is defined by random variables A_n and B_n such that:

$$P(A_n \leq \theta \leq B_n) = 1 - \alpha.$$

The realized interval $[a, b]$ is obtained from a sample $(x_1, ..., x_n)$ as:

$$a = A_n(x_1, ..., x_n), \quad b = B_n(x_1, ..., x_n).$$

Remarks

1. The quantity $1 - \alpha$ is called the **confidence level** of the interval $[a, b]$, i.e. $P(a \leq \theta \leq b) = 1 - \alpha$.

2. In practice, we often have only one sample that provides a single confidence interval $[a, b]$.
3. The parameter to be estimated may be a mean, a variance (for quantitative variables), or a proportion (for qualitative ones).

Confidence Interval for a Mean

We consider the case where X follows a normal distribution $N(\mu, \sigma)$, or when the sample size is large ($n > 30$) so that $\overline{X}$ approximately follows the same law.

Given a sample $(x_1, ..., x_n)$, we define:

$$m = \frac{x_1 + ... + x_n}{n}, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2.$$

(A) Case σ known

$$IC = \left[m - t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\,;\, m + t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

(B) Case σ unknown

$$IC = \left[m - t_{1-\alpha/2,\,n-1}\frac{s}{\sqrt{n}}\,;\, m + t_{1-\alpha/2,\,n-1}\frac{s}{\sqrt{n}}\right]$$

where $t_{1-\alpha/2,\,n-1}$ is the quantile of order $1 - \dfrac{\alpha}{2}$ of Student's t distribution with $n - 1$ degrees of freedom.

Remark

If $n > 30$, then $t_{1-\alpha/2,\,n-1} \approx t_{1-\alpha/2}$.

Confidence Interval for a Variance

(A) Case μ known

$$IC = \left[\frac{nv}{\chi^2_{1-\alpha/2}(n)}\,;\, \frac{nv}{\chi^2_{\alpha/2}(n)}\right]$$

where $v = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \mu)^2$, and $\chi^2_{1-\alpha/2}(n)$ and $\chi^2_{\alpha/2}(n)$ are the chi-squared quantiles of orders $1 - \dfrac{\alpha}{2}$ and $\dfrac{\alpha}{2}$ respectively.

(B) Case μ unknown

$$IC = \left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\,;\, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right]$$

Remark

If $n > 30$, we can approximate:

$$\chi^2_{\alpha}(n-1) \approx \frac{1}{2}\left(t_\alpha + \sqrt{2n-3}\right)^2.$$

Hence:

$$IC = \left[\frac{2(n-1)s^2}{(t_{1-\alpha/2} + \sqrt{2n-3})^2} \, ; \, \frac{2(n-1)s^2}{(t_{\alpha/2} + \sqrt{2n-3})^2} \right]$$

and the symmetry of the standard normal law ensures that $t_{\alpha/2} = -t_{1-\alpha/2}$.

Confidence Interval for a Proportion

From the approximation $F \sim N(p, \sqrt{\frac{pq}{n}})$ with $q = 1 - p$, we deduce:

$$IC = \left[f - t_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \, ; \, f + t_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \right]$$

where f is the sample proportion.