Chapter 2

Point Estimation

We have seen that parametric estimation is a fundamental aspect of parametric statistics : it allows us to determine the parameters of the probability distribution governing the data of a random experiment and, consequently, to describe the characteristics of the underlying population.

There are three main forms of estimation :

1. point estimation,
2. interval estimation (confidence intervals),
3. hypothesis testing.

We begin with point estimation. Suppose the data

$$(x_1, x_2, \ldots, x_n)$$

are n realizations of independent and identically distributed random variables

$$(X_1, X_2, \ldots, X_n),$$

all following the same probability distribution.

From descriptive statistics, we may be able to identify a suitable family of probability distributions for the X_i, but the values of the parameters of that distribution are unknown. How can we estimate these parameters ?

In what follows, let θ denote an unknown parameter of interest.

Definition 2.0.1. *A* point estimator *of θ is a rule or function, based on the sample $(X_1, X_2, \ldots, X_n)$, that produces a single numerical value intended to approximate θ. The resulting value, obtained once the sample is observed, is called a* point estimate *of θ.*

Notation : Let X_i be a random variable with $1 \leq i \leq n$. Each X_i is characterized by its distribution function $F(x, \theta)$.

If X_i follows a continuous probability distribution, we denote its probability density function by $f(x, \theta) = \frac{d}{dx}F(x, \theta)$. If instead X_i follows a discrete distribution, then the elementary probabilities are given by $\mathbb{P}(X_i = x_i \mid \theta)$.

Definition 2.0.2. *An* estimator *of a parameter θ is a statistic*

$$S_n = S(X_1, X_2, \ldots, X_n),$$

whose values lie in the parameter space Θ. An estimate *of θ is the realization of this statistic when the sample is observed, that is*

$$S_n = S(x_1, x_2, \ldots, x_n).$$

Remark 2. *An estimator is a random variable, while an estimate is the numerical value it takes for a given sample.*

2.1 Properties of an estimator

We consider a sample $(X_1, X_2, \ldots, X_n)$ of n random variables following the same distribution with an unknown parameter θ.

1. An estimator T_n of the parameter θ is said to be ***unbiased*** if

$$\mathbb{E}(T_n) = \theta.$$

   More generally, an estimator T_n of a function $g(\theta)$ is unbiased if $\mathbb{E}(T_n) = g(\theta)$. If instead $\mathbb{E}(T_n) \neq g(\theta)$, the estimator is *biased*.

2. T_n is ***asymptotically unbiased*** for θ if

$$\lim_{n\to\infty} \mathbb{E}(T_n) = \theta.$$

3. T_n is ***consistent*** for $g(\theta)$ if it converges in probability to $g(\theta)$:

$$T_n \xrightarrow{\mathbb{P}} g(\theta).$$

   A sufficient condition for consistency is

$$\lim_{n\to\infty} \mathbb{E}(T_n) = g(\theta) \quad \text{and} \quad \lim_{n\to\infty} \mathrm{Var}(T_n) = 0.$$

Definition 2.1.1. *Let T_n be an estimator of θ :*

1. *The **bias** of T_n is*
$$\mathrm{Bias}(T_n) = \mathbb{E}(T_n) - \theta.$$
2. *The **mean squared error** (MSE) is*
$$\mathrm{MSE}(T_n) = \mathbb{E}\Big[(T_n - \theta)^2\Big] = \mathrm{Var}(T_n) + \Big(\mathrm{Bias}(T_n)\Big)^2.$$

Consequences :

— $T_n \xrightarrow{\mathbb{L}^2} \theta \;\Leftrightarrow\; \mathrm{MSE}(T_n) \to 0.$

— If T_n is unbiased, then
$$\mathrm{MSE}(T_n) = \mathrm{Var}(T_n),$$
and therefore $T_n \xrightarrow{\mathbb{L}^2} \theta \;\Leftrightarrow\; \mathrm{Var}(T_n) \to 0.$

— Among unbiased estimators, the **best** estimator is the one with the smallest variance.

— An unbiased estimator with high variance may give estimates very far from the true parameter value.

2.2 Point Estimation by the Method of Moments (MoM)

The basic idea of the *method of moments* is to estimate the parameters of the distribution of a random variable X by equating the theoretical moments (centered or non-centered) with their empirical counterparts computed from the data.

The theoretical p-th moment of X under the law $\mathbb{P}_\theta$ is defined by
$$m_p = \mathbb{E}_\theta(X^p), \qquad p \geq 1,$$
provided it exists.

Definition 2.2.1. *The **empirical p-th moment** based on the sample $(X_1, \ldots, X_n)$ is*
$$\hat{m}_p = \frac{1}{n}\sum_{i=1}^{n} X_i^p.$$

In particular, for the variance we have
$$\sigma_\theta^2 = \mathbb{E}_\theta(X^2) - \Big(\mathbb{E}_\theta(X)\Big)^2 = m_2 - (m_1)^2.$$

Replacing theoretical moments with empirical moments leads to the *empirical variance* :

$$S_n^2 = \hat{m}_2 - (\hat{m}_1)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean.

Remark 3. *The variance estimator S_n^2 defined above is* biased *for σ^2, but it is consistent. An unbiased estimator is obtained by using*

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Example. The expectation $\mu = \mathbb{E}(X)$ can be estimated by the first empirical moment,

$$\hat{\mu} = \bar{X}_n,$$

and the variance $\sigma^2 = \mathrm{Var}(X)$ by the empirical variance (second centered moment),

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

2.2.1 General Framework of the Method of Moments

Let $X_1, X_2, \ldots, X_n$ be a sample of size n, where each X_i takes values in a set $\mathcal{X}$.

Let $f : \mathcal{X} \to \mathbb{R}^k$ be a measurable function such that the mapping

$$\varphi : \Theta \longrightarrow \mathbb{R}^k, \qquad \theta \longmapsto \varphi(\theta) = \mathbb{E}_\theta\Big[f(X)\Big]$$

is invertible. Then

$$\theta = \varphi^{-1}\Big(\mathbb{E}_\theta\Big[f(X)\Big]\Big).$$

Definition 2.2.2. *The **method of moments estimator** $\hat{\theta}$ of θ is defined as the solution (when it exists and is unique) of*

$$\varphi(\hat{\theta}) \;=\; \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

In particular, if the distribution has two parameters (θ_1, θ_2) such that

$$\varphi(\theta_1, \theta_2) = \Big(\mathbb{E}(X),\ \mathrm{Var}(X)\Big),$$

then

$$\varphi(\hat{\theta}_1, \hat{\theta}_2) = \Big(\overline{X}_n,\ S_n^2\Big),$$

so that

$$(\hat{\theta}_1, \hat{\theta}_2) = \varphi^{-1}\Big(\overline{X}_n, S_n^2\Big).$$

Examples

1. **Bernoulli**(p) **:** $\mathbb{E}(X) = p$.

$$\hat{p} = \overline{X}_n.$$

2. **Poisson**(λ) **:** $\mathbb{E}(X) = \lambda$.

$$\hat{\lambda} = \overline{X}_n.$$

3. **Exponential**(λ) **:** $\mathbb{E}(X) = 1/\lambda$.

$$\hat{\lambda} = \frac{1}{\overline{X}_n}.$$

4. **Normal**(m, σ^2) **:** $\mathbb{E}(X) = m$, $\mathrm{Var}(X) = \sigma^2$.

$$\hat{m} = \overline{X}_n, \qquad \hat{\sigma}^2 = S_n^2.$$

5. **Gamma**(α, β) **:** $\mathbb{E}(X) = \alpha/\beta$, $\mathrm{Var}(X) = \alpha/\beta^2$. Solving gives

$$\hat{\alpha} = \frac{\overline{X}_n^2}{S_n^2}, \qquad \hat{\beta} = \frac{\overline{X}_n}{S_n^2}.$$

2.3 Estimation by the Maximum Likelihood Method

The statistician R.A. Fisher introduced a general estimation method, known as the **maximum likelihood method**. It is one of the most important and widely used statistical techniques for estimating the parameters of a probability distribution, i.e., of the proposed **statistical model**.

Let X be a random variable with a parametric distribution (discrete or continuous), whose parameter θ we wish to estimate. Define

$$f(x;\theta) = \begin{cases} f_\theta(x), & \text{if } X \text{ is continuous with density } f_\theta, \\ \mathbb{P}_\theta(X = x), & \text{if } X \text{ is discrete with probability mass function } \mathbb{P}_\theta. \end{cases}$$

Definition 2.3.1. *Given a sample of i.i.d. random variables* $(X_1, \ldots, X_n)$*, the* ***likelihood function*** *of* θ *for the realization* $\mathbf{x} = (x_1, \ldots, x_n)$ *is*

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

Definition 2.3.2. *The* ***maximum likelihood estimator (MLE)*** *of* θ *is defined by*

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\mathbf{x};\theta).$$

Since the likelihood is always positive and the logarithm is strictly increasing, maximizing $L(\mathbf{x};\theta)$ is equivalent to maximizing the **log-likelihood function** :

$$\ell(\mathbf{x};\theta) = \ln L(\mathbf{x};\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta).$$

Optimization conditions

In practice, finding $\hat{\theta}$ is an optimization problem. If $\ell(\theta)$ is differentiable, then :

1. **Necessary condition (first-order) :**

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \quad \text{or equivalently} \quad \frac{\partial \ell(\theta)}{\partial \theta} = 0.$$

2. **Second-order condition :** At $\theta = \hat{\theta}$, if

$$\frac{\partial^2 L(\theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} < 0 \quad \text{or equivalently} \quad \frac{\partial^2 \ell(\theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} < 0,$$

then $\hat{\theta}$ is a local maximum of $\ell(\theta)$ (and hence of $L(\theta)$).

One must still check that the obtained solution corresponds to a *global maximum* over the parameter space Θ.

Example 2.3.3. *We wish to estimate the parameter λ of a Poisson distribution from an n-sample. We have*

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \mathbb{P}_\lambda(X = x_i) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}.$$

The likelihood function is written as

$$L(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-\lambda n} \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!}.$$

Since the likelihood is positive, it is simpler to work with its logarithm :

$$\begin{aligned} \ln L(x_1, \ldots, x_n; \lambda) &= \ln e^{-\lambda n} + \ln \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} \\ &= -\lambda n + \ln \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} \\ &= -\lambda n + \sum_{i=1}^{n} x_i \ln \lambda - \sum_{i=1}^{n} \ln(x_i!). \end{aligned}$$

The first derivative is

$$\frac{\partial \ln L(x_1, \ldots, x_n; \lambda)}{\partial \lambda} = -n + \sum_{i=1}^{n} \frac{x_i}{\lambda},$$

which vanishes at

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The second derivative is

$$\frac{\partial^2 \ln L(x_1, \ldots, x_n; \lambda)}{\partial \lambda^2} = -\sum_{i=1}^{n} \frac{x_i}{\lambda^2},$$

which is always non-positive.

Thus, the maximum likelihood estimator *is*

$$\hat{\Lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

and for a given realization $(x_1, \dots, x_n)$*, the* estimate *is*

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}.$$

Since the second derivative is negative, this indeed corresponds to a maximum.

Example 2.3.4 (Continuous Distribution)**.** *We wish to estimate the parameters* μ *and* σ *of a normal distribution from an n-sample. The normal law* $\mathcal{N}(\mu, \sigma)$ *has density function*

$$f(x; \mu, \sigma) = f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Let us write the likelihood function for a realization of an n-sample of independent variables :

$$f(x_1, \dots, x_n; \mu, \sigma) = \prod_{i=1}^{n} f(x_i; \mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right).$$

We can decompose the sum of squares as

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2,$$

where $\bar{x}$ *denotes the sample mean. Thus, the likelihood function can be written as*

$$f(x_1, \dots, x_n; \mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2}\right).$$

For μ*, we compute*

$$\frac{\partial}{\partial \mu} \ln L = \frac{\partial}{\partial \mu} \left(\ln\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} - \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2} \right) \right) = 0 - \frac{-2n(\bar{x} - \mu)}{2\sigma^2}),$$

$$\frac{\partial}{\partial \mu} \ln L = 0 \Rightarrow \frac{2n(\bar{x} - \mu)}{2\sigma^2} = 0 \quad \Longrightarrow \quad \hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

For the second parameter, we compute

$$\frac{\partial}{\partial \sigma} \ln L = \frac{\partial}{\partial \sigma} \left(\frac{n}{2} \ln \frac{1}{2\pi\sigma^2} \right) - \frac{\partial}{\partial \sigma} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^3},$$

hence

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2,$$

which can also be written as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

We now check that these correspond to local maxima :

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2} \leq 0,$$

$$\frac{\partial^2 \ln L}{\partial \sigma^2} = \frac{n}{\sigma^4} - \frac{3}{\sigma^4}\Big(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\Big).$$

At the point $\hat{\sigma}$*,*

$$\frac{\partial^2 \ln L}{\partial \sigma^2}(\hat{\sigma}) = \frac{n}{\hat{\sigma}^4} - \frac{3}{\hat{\sigma}^4}\Big(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\Big) \leq 0.$$

The method provides an unbiased estimator of the mean, $\mathbb{E}(\hat{\mu}) = \mu$*. However, the variance estimator is biased, i.e.* $\mathbb{E}(\hat{\sigma}^2) \neq \sigma^2$*. Nevertheless, the estimator is asymptotically unbiased.*

2.4 Fisher Information

To simplify the notation, we assume in this section that the parameter θ belongs to $\mathbb{R}$. The results remain valid—with the necessary adaptations—when θ is multidimensional (these extensions will be mentioned at the end of the section).

We denote by $L'(\theta; x)$ (respectively $L''(\theta; x)$) the first (respectively second) derivative with respect to θ of the likelihood function $L(x; \theta)$, for the observed value x in the considered parametric model.

In what follows, we assume that the parametric model $(E, \mathcal{E}, \{\mathbb{P}_\theta : \theta \in \Theta\})$, with generic random variable X, satisfies the following assumptions :

H1. The parameter space Θ is an open set.

H2. The probability laws $\mathbb{P}_\theta$ all have the same support, which therefore does not depend on θ.

H3. The first and second derivatives $L'(x;\theta)$ and $L''(x;\theta)$ of the likelihood function exist for each $x \in E$.

H4. The functions $L'(x;\theta)$ and $L''(x;\theta)$, viewed as functions of the variable x (i.e., as densities), are integrable for all $\theta \in \Theta$, and integration and differentiation can be interchanged :

$$\frac{\partial}{\partial\theta}\int_A L(x;\theta)\,dx = \int_A L'(x;\theta)\,dx, \qquad \frac{\partial^2}{\partial\theta^2}\int_A L(x;\theta)\,dx = \int_A L''(x;\theta)\,dx,$$

for all $A \in \mathcal{E}$.

Consider the random variable

$$S(X,\theta) = \frac{\partial}{\partial\theta}\ln L(X;\theta) = \frac{L'(X;\theta)}{L(X;\theta)},$$

which, as a function of θ, is sometimes called the *score function*. Under the above assumptions, this random variable satisfies :

$$E_\theta\Big(S(X,\theta)\Big) = \int_E \frac{\partial}{\partial\theta}\ln L(x;\theta)L(x;\theta)\,dx = \int_E L'(x;\theta)\,dx = \frac{\partial}{\partial\theta}\int_E L(x;\theta)\,dx = 0,$$

since $\frac{\partial}{\partial\theta}\ln L(x;\theta) = \frac{L'(x;\theta)}{L(x;\theta)}$ and $\int_E L(x;\theta)\,dx = 1$.

We now make the additional assumption :

H5. The score function is square-integrable.

Definition 2.4.1. ***(Fisher Information).*** *The Fisher information is defined as the variance of the score function, that is,*

$$I(\theta) = \mathrm{Var}_\theta(S(X,\theta)) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln L(X;\theta)\right)^2\right].$$

We can express the Fisher information in an equivalent form.

Proposition 2.4.2. *The Fisher information can also be written as*

$$I(\theta) = -\,\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}S(X,\theta)\right) = -\,\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\ln L(X;\theta)\right).$$

Proof. We note that

$$\frac{\partial}{\partial\theta}S(x,\theta) = \frac{\partial}{\partial\theta}\frac{L'(x,\theta)}{L(x,\theta)} = \frac{L''(x,\theta)}{L(x,\theta)} - \frac{(L'(x,\theta))^2}{(L(x,\theta))^2} = \frac{L''(x,\theta)}{L(x,\theta)} - S(x,\theta)^2.$$

Therefore,

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}S(X,\theta)\right] = \mathbb{E}_\theta\left[\frac{L''(X,\theta)}{L(X,\theta)}\right] - I(\theta).$$

Now, observe that

$$\mathbb{E}_\theta\left[\frac{L''(X,\theta)}{L(X,\theta)}\right] = \int_E L''(x,\theta)\,dx = \frac{\partial^2}{\partial\theta^2}\int_E L(x,\theta)\,dx = 0,$$

which follows from assumption **H4**. Hence, the stated result holds. □

Example 2.4.3. ***Fisher Information in the Case of a Real Gaussian Model with Known Variance*** σ^2**.**

Consider the model :

$$\mathcal{P} = \left\{\mathcal{N}(\mu,\sigma^2) : \mu \in \mathbb{R}\right\},$$

where σ^2 *is assumed to be known.*

The log-likelihood for an observation x *is given by :*

$$\ln L(x;\mu) = -\ln(\sigma\sqrt{2\pi}) - \frac{(x-\mu)^2}{2\sigma^2}.$$

Hence,

$$S(x,\mu) = \frac{x-\mu}{\sigma^2} \quad \textit{and} \quad \frac{\partial}{\partial\mu}S(x,\mu) = -\frac{1}{\sigma^2}.$$

Since the latter expression is constant in x*, the Fisher information in this model is therefore*

$$I(\mu) = \frac{1}{\sigma^2}.$$

Of course, we obtain the same result by computing

$$\mathbb{E}_\mu\left[S^2(X,\mu)\right] = \mathbb{E}_\mu\left[\left(\frac{X-\mu}{\sigma^2}\right)^2\right] = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

We note that the Fisher information increases as the variance σ^2 *decreases.*

Fisher Information in a Sampling Model.

Now consider the Fisher information in a sampling model, where we observe an independent and identically distributed (i.i.d.) sample $X_1, \ldots, X_n$ with the same distribution as X. Then,

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} L(x_i; \theta), \quad \text{and} \quad \ln L(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \ln L(x_i; \theta).$$

Differentiating twice with respect to θ, we obtain :

$$\frac{\partial}{\partial \theta} S(X_1, \ldots, X_n; \theta) = \frac{\partial^2}{\partial \theta^2} \ln L(X_1, \ldots, X_n; \theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln L(X_i; \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} S(X_i; \theta),$$

which proves the following proposition.

Proposition 1. *The Fisher information for a sampling model—that is, for the sample $X_1, \ldots, X_n$—is n times that of the generic random variable X from which the sample is drawn. In other words,*

$$I_n(\theta) = n\, I(\theta),$$

where $I_n(\theta)$ denotes the Fisher information of the sample $X_1, \ldots, X_n$, and $I(\theta)$ denotes the Fisher information of a single observation X.

Example 2.4.4. ***Fisher Information for a Sample in a Real Gaussian Model with Known Variance σ^2 (Continuation of Example 2.4.3).***

From the result of Example 2.4.3 and the previous proposition, the Fisher information for a sample in this model is

$$I_n(\theta) = n\, I(\theta) = \frac{n}{\sigma^2}.$$

Remark 4. ***Case Where the Parameter θ is Multidimensional.***

- *The functions $L'(x; \theta)$ and $L''(x; \theta)$ correspond respectively to the gradient and the Hessian matrix of the likelihood function $L(x; \theta)$.*
- *The score function is then a random vector, given by the gradient of the log-likelihood evaluated at the random variable X :*

$$S(X, \theta) = \nabla_\theta \ln L(X; \theta).$$

- *The Fisher information becomes a matrix corresponding to the covariance matrix of the score $S(X, \theta)$. It is also equal to the negative expected value of the Hessian matrix of the log-likelihood :*

$$I(\theta) = \Sigma_{S(X,\theta)} = -\,\mathbb{E}_\theta\Big[\nabla^2_\theta \ln L(X; \theta)\Big].$$

2.5 Cramér–Rao Bound (or Fréchet–Darmois–Cramér–Rao Inequality)

Theorem 2. *Let $(E, \mathcal{E}, \{\mathbb{P}_\theta : \theta \in \Theta\})$ be a parametric model, where $\Theta \subset \mathbb{R}$, with generic random variable X, satisfying assumptions **H1–H4** from the previous section (and possibly **H5** if we wish to use the alternative expression of the Fisher information). Let $T(X)$ be an unbiased and square-integrable estimator of $g(\theta) \in \mathbb{R}$.*

Assume that the function $x \mapsto T(x)L'(x;\theta)$ is integrable over E, and that differentiation and integration can be interchanged, i.e.,

$$\frac{\partial}{\partial\theta}\int_E T(x)L(x;\theta)\,dx = \int_E T(x)L'(x;\theta)\,dx.$$

Finally, suppose that the Fisher information $I(\theta)$ is strictly positive for all $\theta \in \Theta$.

Then the function g is differentiable, and for all $\theta \in \Theta$, we have

$$\mathrm{Var}(T(X)) \geq \frac{[g'(\theta)]^2}{I(\theta)}.$$

The bound $\frac{[g'(\theta)]^2}{I(\theta)}$ is called the Cramér–Rao bound (or Fréchet–Darmois–Cramér–Rao bound).

Remark 5. *The integrability assumption on the function $T(\cdot)L'(\cdot;\theta)$ and the possibility of interchanging differentiation and integration are ensured whenever there exists an integrable function h that dominates $T(\cdot)L'(\cdot;\theta)$, i.e.*

$$|T(x)L'(x;\theta)| \leq h(x), \quad \forall x \in E, \quad \text{and} \quad \int_E h(x)\,dx < +\infty.$$

Proof. (of Theorem 2) Since the estimator $T(X)$ is integrable and unbiased, we have :

$$\mathbb{E}_\theta[T(X)] = \int_E T(x)L(x;\theta)\,dx = g(\theta).$$

By the theorem's assumption, we can differentiate under the integral sign, which gives :

$$g'(\theta) = \int_E T(x)L'(x;\theta)\,dx,$$

and the integral on the right-hand side is well defined.

Now, recalling that $S(x,\theta) = \frac{L'(x;\theta)}{L(x;\theta)}$, we can write :

$$g'(\theta) = \int_E T(x)S(x,\theta)L(x;\theta)\,dx = \mathbb{E}_\theta[T(X)S(X,\theta)].$$

Since we have shown that $S(X,\theta)$ is a centered random variable, it follows that :

$$g'(\theta) = \mathbb{E}_\theta[T(X)S(X,\theta)] - \mathbb{E}_\theta[T(X)]\,\mathbb{E}_\theta[S(X,\theta)] = \mathrm{Cov}_\theta(T(X), S(X,\theta)).$$

By the Cauchy–Schwarz (or Hölder) inequality, we have :

$$\Big(g'(\theta)\Big)^2 = \Big(\mathrm{Cov}_\theta(T(X), S(X,\theta))\Big)^2 \leq \mathrm{Var}_\theta(T(X))\ \mathrm{Var}_\theta(S(X,\theta)).$$

Rearranging terms, we obtain :

$$\mathrm{Var}_\theta(T(X)) \geq \frac{[g'(\theta)]^2}{I(\theta)},$$

since by definition the variance of $S(X,\theta)$ is the Fisher information $I(\theta)$.

□

Remark 6. — *In the case of a sample model of size n, the Cramér–Rao bound naturally becomes*

$$\frac{[g'(\theta)]^2}{I_n(\theta)}.$$

— *The Cramér–Rao bound depends only on the parametric model, on the function $g(\theta)$ being estimated, and on the sample size. The larger the sample size, the smaller the bound becomes (which is intuitive and desirable !).*

— *There is no guarantee that the minimum variance is actually attained.*

— *In the case where the parameter θ is multidimensional, taking values in $\Theta \subset \mathbb{R}^p$, and the function g takes values in $\mathbb{R}^k$, with the Fisher information matrix $I(\theta)$ invertible, the Cramér–Rao lower bound for the covariance matrix $\Sigma_{T(X)}$ of an unbiased estimator $T(X)$ is given by :*

$$\Sigma_{T(X)} \geq \nabla'_\theta g(\theta)\, I^{-1}(\theta)\, \nabla_\theta g(\theta),$$

where $\nabla_\theta g(\theta)$ denotes the gradient of g with respect to θ.

Definition 2.5.1 (Efficient Estimator)**.** *An unbiased estimator that attains the Cramér–Rao lower bound is said to be efficient.*

It is said to be asymptotically efficient if

$$\lim_{n \to +\infty} \frac{[g'(\theta)]^2}{I_n(\theta)\ \mathrm{Var}_\theta(T_n(X))} = 1,$$

where $T_n(X)$ denotes the estimator based on a sample of size n.

Note that an unbiased efficient estimator is necessarily optimal within the class of unbiased estimators. The converse, however, is false, since the Cramér–Rao bound is not necessarily attained.

Example 2.5.2. *Fisher Information for a sample in the case of a real Gaussian model with variance σ^2 (continuation of Examples 2.4.3 and 2.4.4).*

We have seen that the variance of the empirical mean estimator $\bar{X}_n$ is σ^2/n. This quantity is equal to the Cramér–Rao lower bound in the real Gaussian model with known variance (where only the expectation μ is unknown). Since the estimator $\bar{X}_n$ is unbiased, it is therefore efficient in this model for estimating the parameter μ. $\diamond$