STAT0008 Lecture 3
Properties of Estimators

Dr. Aidan O'Keeffe

Department of Statistical Science
University College London

15th October 2018

Outline

- Point estimation of a parameter
  - Unbiasedness
  - Precision
  - Mean squared error
- Consistency
- Relative efficiency of two estimators
- Score function
- Fisher information
- Cramér-Rao lower bound

Criteria for Point Estimation

We assume the following

- There is a single, unknown, parameter θ that we wish to estimate
- We shall estimate θ using some statistic $T(\mathbf{X})$, with $\mathbf{X} = (X_1, \ldots, X_n)$. This is known as **point estimation**.
- Here, T is a function of the random sample $\mathbf{X}$.

Note that $T(\mathbf{X})$ is a random variable that has a **sampling distribution**.

Properties of the sampling distribution of T may help us to assess the quality of T as an estimator for θ (i.e. how 'good' is T as an estimator of θ – this depends on what we mean by 'good'!)

Unbiasedness

Definition: Unbiased Estimator

Suppose that $T(\mathbf{X}) = T(X_1, \ldots, X_n)$ is an estimator for θ. We say that $T(\mathbf{X})$ is an unbiased estimator for θ if

$$\mathbb{E}(T(\mathbf{X})) = \theta$$

The **bias** of an estimator of θ is defined as the difference

$$\text{Bias}(T(\mathbf{X})) = \mathbb{E}(T(\mathbf{X})) - \theta.$$

Note here that expectations are taken with respect to the sampling distribution of $T(\mathbf{X})$.

Unbiased Estimators

An unbiased estimator is *fair* in the sense that such an estimator does not consistently over- or under- estimate the unknown parameter of interest.

Unbiasedness is **not invariant**. In other words, if T is an unbiased estimator for θ then $g(T)$ is not necessarily an unbiased estimator for $g(\theta)$ for some function $g(.)$.

Often, unbiased estimators can be determined by adjusting obvious estimators.

Example: Unbiasedness

Suppose that $X_1, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2. Find a unbiased estimator for μ^2.

Example: Unbiasedness

Example: Unbiasedness

Comments

We see that a 'method' for finding an unbiased estimator might be to

1. 'Guess' at a sensible estimator.
2. Derive its expectation.
3. Adjust the estimator accordingly to remove the bias and recover an unbiased estimator.

Precision and Mean Squared Error

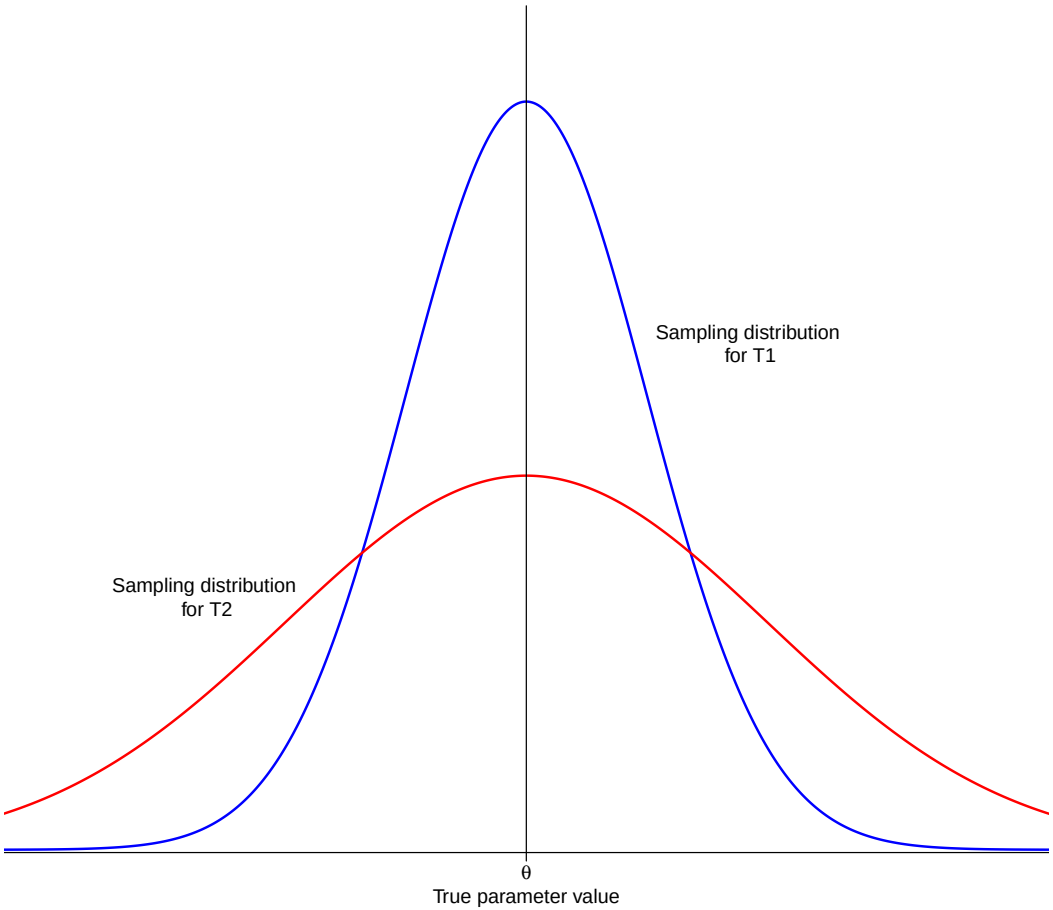
Unbiasedness is often considered to be desirable. However, unbiasedness alone does not guarantee that an estimator is 'good'. We need to consider the variability of an estimator.

If an unbiased estimator has a large variance then we say that the estimator is not very precise.

On average, the estimator would result in the true parameter value, θ, but because the estimator is not very precise, individual estimates of θ calculated using the estimator are likely to differ substantially from each other.

Precision and Mean Squared Error

Consider density plots of the sampling distributions of two estimators $T1$ and $T2$, below.



Precision and Mean Squared Error

Both $T1$ and $T2$ are unbiased, but we can see that $T1$ is more precise than $T2$ because the density of $T1$ is more concentrated around the true parameter value, θ.

Typically, we should seek an estimator that is unbiased and has a relatively small variance, when compared to other competing estimators.

One measure that assesses both the bias and variance of an estimator is the **mean squared error**.

Mean Squared Error

Often, we might seek an estimator with a small **mean squared error**.

Definition: Mean Squared Error

Suppose that $T = T(X_1, \ldots, X_n)$ is an estimator for θ. The mean squared error of T is defined:

$$\text{MSE}(T; \theta) = \mathbb{E}\left[(T - \theta)^2\right].$$

We can show that the mean squared error is related directly to the bias and variance of T.

Bias–Variance Decomposition

$$
\begin{aligned}
\mathsf{MSE}(T;\theta) &= \mathbb{E}\left[(T-\theta)^2\right] \\
&= \mathbb{E}\left[T^2 - 2\theta T + \theta^2\right] \\
&= \mathbb{E}(T^2) - 2\theta\mathbb{E}(T) + \theta^2 \\
&= \mathbb{E}(T^2) - \left[\mathbb{E}(T)\right]^2 + \left[\mathbb{E}(T)\right]^2 - 2\theta\mathbb{E}(T) + \theta^2 \\
&= \left\{\mathbb{E}(T^2) - \left[\mathbb{E}(T)\right]^2\right\} + \left\{\left[\mathbb{E}(T)\right]^2 - 2\theta\mathbb{E}(T) + \theta^2\right\} \\
&= \mathsf{Var}(T) + \left\{\left(\mathbb{E}(T) - \theta\right)^2\right\} \\
&= \mathsf{Var}(T) + \left[\mathsf{Bias}(T)\right]^2
\end{aligned}
$$

We see that the mean squared error of an estimator is the sum of its variance and squared bias. Hence, the mean squared error will be large if *either* the bias *or* variance is large (or both!).

As a result, we should like an unbiased estimator for θ with small variance (i.e. a precise, unbiased estimator).

Dominant and Admissible Estimators

Suppose that T_1 and T_2 are two estimators for θ. If

$$\text{mse}(T_1;\theta) \geq \text{mse}(T_2;\theta) \quad \text{for all } \theta \in \Theta$$

$$\text{There exists some } \theta \in \Theta \text{ such that } \text{mse}(T_1;\theta) > \text{mse}(T_2;\theta)$$

then we say that T_2 dominates T_1 (with respect to the mean squared error).

An estimator T' is said to be **admissible** if there is no other estimator that dominates T'.

In words we'd say that there is no uniformly better estimator than T' (with respect to the mean squared error).

Consistency

We have seen that, in most cases, an estimator of some parameter of interest, θ, is a function of a random sample $\{X_1, \ldots, X_n\}$.

We may be interested in knowing about the behaviour of an estimator for θ as the size of the sample, n, changes. We define

$$T_n = T(X_1, \ldots, X_n)$$

to be an estimator for θ, constructed using a sample of size n: $X_1, \ldots, X_n$.

In words (mathematical definition to follow!), we say that T_n is **consistent** for θ if T_n approaches θ as n tends to infinity.

Consistency

A more formal (mathematical) definition:

Definition: Consistency

Let T_n be an estimator for some parameter θ based on a sample of size n. Then T_n is said to be consistent for θ if T_n converges in probability to θ as n tends to infinity. That is, for all $\epsilon > 0$

$$\mathbb{P}\left(|T_n - \theta| > \epsilon\right) \to 0 \quad \text{as } n \to \infty.$$

Strong Consistency

Furthermore, we say that T_n is **strongly consistent** for θ if

$$\mathbb{P}\left(\lim_{n\to\infty} T_n = \theta\right) = 1.$$

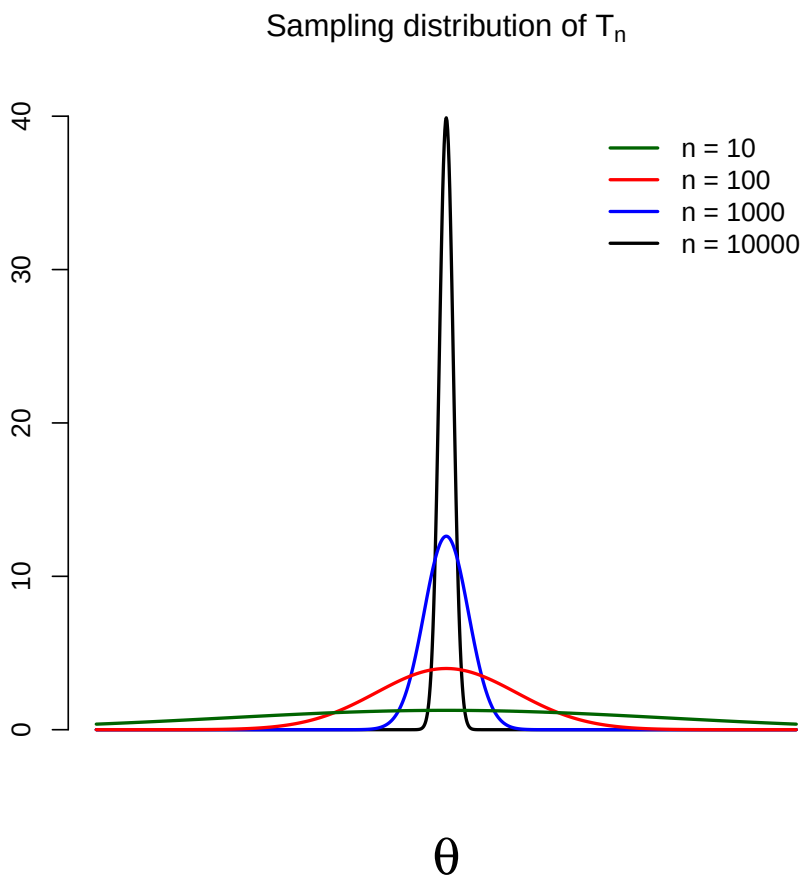
In other words, T_n converges *almost surely* to θ as $n \to \infty$.

Typically, we shall be concerned with consistency rather than strong consistency. We note here that, in some textbooks, 'consistency' may sometimes be referred to as '**weak** consistency'.

Strong consistency implies consistency but (weak) consistency does not necessarily imply that an estimator is also strongly consistent.

Consistency

Consistency conveys the idea that the **location** of the sampling distribution of the estimator $T_n(\mathbf{X})$ is concentrated around the true parameter value, θ, for large n.



Consistency and Mean Squared Error

It can be shown that if

$$\text{mse}(T_n; \theta) \to 0 \quad \text{as} \quad n \to \infty$$

then both Bias(T_n) and Var(T_n) must tend to zero and T_n is consistent for θ.

We note that this property is sufficient but not necessary for consistency.

Hence, a straightforward way to establish consistency for an **unbiased** estimator is to assess whether the variance of the unbiased estimator tends to zero as $n \to \infty$.

Consistency: Example 1

Suppose that $X_1, \ldots, X_n$ are iid Exp$(\frac{1}{\mu})$ random variables. Find the maximum likelihood estimator for μ and show that this estimator is consistent.

Consistency: Example 1

Consistency: Example 1

Consistency: Example 2

Now suppose that $X_{(1)} = \min(X_1, \ldots, X_n)$. Then

$$\begin{aligned}\mathbb{P}(X_{(1)} > x) &= \mathbb{P}(X_1 > x \cap X_2 > x \cap \ldots \cap X_n > x) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i > x) \\ &= \exp\left(-\frac{nx}{\mu}\right)\end{aligned}$$

So $X_{(1)} \sim \text{Exp}(\frac{n}{\mu})$ which implies that $\mathbb{E}(X_{(1)}) = \frac{\mu}{n}$. Hence

$$nX_{(1)} \text{ is unbiased for } \mu.$$

Then

$$\text{Var}(nX_{(1)}) = n^2\text{Var}(X_{(1)}) = n^2\frac{\mu^2}{n^2} = \mu^2.$$

So $\text{Var}(nX_{(1)})$ does not tend to zero as $n \to \infty$ and $nX_{(1)}$ and we cannot apply our previous result.

Relative Efficiency

Suppose that T_1 and T_2 are unbiased estimators of some parameter, θ.

We say that T_1 is **more efficient** than T_2 if $\text{Var}(T_2) > \text{Var}(T_1)$.

The efficiency of T_2 relative T_1 is defined as

$$\text{Eff}(T_2, T_1) = 100\frac{\text{Var}(T_1)}{\text{Var}(T_2)}$$

and this quantity is typically quoted as a percentage.

Relative Efficiency: Example

Suppose that $X \sim \mathcal{N}(\theta, 4)$ and $Y \sim \mathcal{N}(\theta, 9)$ with X and Y independent. Let $U = \frac{1}{2}(X + Y)$ and $V = \frac{1}{3}(X + 2Y)$. Determine the variances of U and V and compute the relative efficiency of these estimators.

Relative Efficiency: Example

Example: Linear Unbiased Estimator

Let us now consider the class of estimators that is all linear combinations of X and Y that are unbiased estimators of θ. This class is written

$$W = aX + (1-a)Y \quad \text{for } 0 \leq a \leq 1.$$
$$\text{Var}(W) = 4a^2 + 9(1-a)^2$$

We consider the value of a that gives the smallest variance of W.

$$\frac{\partial}{\partial a}\text{Var}(W) = 8a - 18(1-a).$$

Hence $\text{Var}(W)$ is minimised where $a = \frac{18}{26} = \frac{9}{13}$. The **best linear unbiased** estimator of θ is

$$W = \frac{9}{13}X + \frac{4}{13}Y$$

which has a variance of 2.77.

Note

We note that minimising the variance of an estimator alone is not necessarily a useful criterion for choosing an appropriate estimator.

For example, if $T(\mathbf{X}) = c$ for some constant c then, regardless of the data $\mathbf{X}$, $\text{Var}(T) = 0$ but, typically, T will not be a very good estimator for θ.

Now we shall consider properties of the variance of an estimator and derive a key result known as the **Cramér-Rao lower bound**.

Score Function

Recall that a log-likelihood function is defined

$$\ell(\theta; \mathbf{X}) = \log \mathcal{L}(\theta; \mathbf{X}).$$

The **score function** is the gradient of the log-likelihood function (i.e. here, the first derivative with respect to θ). We write the score function as

$$\begin{aligned} U(\theta; \mathbf{X}) &= \frac{\partial}{\partial \theta} \ell(\theta; \mathbf{X}) \\ &= \frac{\partial}{\partial \theta} \log \mathcal{L}(\theta; \mathbf{X}) \end{aligned}$$

Score Function: Expectation

Key result: $\mathbb{E}[U(\theta; \mathbf{X})] = 0$.

Score Function: Expectation

Score Function: Variance

The variance of the score function is given by

$$\begin{aligned}\text{Var}(U(\theta;\mathbf{X})) &= \mathbb{E}\left[(U(\theta;\mathbf{X}) - \mathbb{E}[U(\theta;\mathbf{X})])^2\right] \\ &= \mathbb{E}\left[(U(\theta;\mathbf{X}))^2\right] \quad \text{because } \mathbb{E}[U(\theta;\mathbf{X})] = 0 \\ &= \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\mathcal{L}(\theta;\mathbf{X})\right)^2\right]\end{aligned}$$

The variance of the score function is known as the **Fisher information** and is usually denoted $\mathcal{I}(\theta)$.

Fisher Information

Under certain *regularity conditions*, we can show that the Fisher information is also given by

$$\mathcal{I}(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\log\mathcal{L}(\theta;\mathbf{X})\right]$$

which is often easier to evaluate. Two important regularity conditions are:

- Likelihood function is continuous in θ.
- Domain of the density function of the data does not depend on θ.

Fisher Information

Key result: $\mathcal{I}(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\log\mathcal{L}(\theta;\mathbf{X})\right]$

Fisher Information

Fisher Information

The Fisher information, $\mathcal{I}(\theta)$ is a measure of the information contained within the dataset $\mathbf{X}$ about the unknown parameter θ.

Suppose $X_1, \ldots, X_n$ is an iid sample of size n. Then if $i(\theta)$ denotes the Fisher information from the observation X_i and $\mathcal{I}_n(\theta)$ denotes the Fisher information in the entire sample of size n, the following result holds

$$\mathcal{I}_n(\theta) = ni(\theta).$$

Fisher Information – Transformations

Suppose that we are interested in the parameter $\phi = g(\theta)$. Then, given data $\mathbf{X}$, the likelihood function $\mathcal{L}^*$ for ϕ satisfies

$$\mathcal{L}^*(\phi; \mathbf{X}) = \mathcal{L}(\theta; \mathbf{X}).$$

Differentiating with respect to ϕ, we obtain

$$\frac{\partial}{\partial \phi} \log \mathcal{L}^*(\phi; \mathbf{X}) = \frac{\partial}{\partial \theta} \log \mathcal{L}(\theta; \mathbf{X}) \frac{\partial \theta}{\partial \phi}$$

where $\frac{\partial \theta}{\partial \phi} = \{\frac{\partial \phi}{\partial \theta}\}^{-1} = \{g'(\theta)\}^{-1}$.

Thus we see that the information $\mathcal{I}^*(\phi) = \mathbb{E}[\{\frac{\partial}{\partial \phi} \log \mathcal{L}^*(\phi; \mathbf{X})\}^2]$ about ϕ in the sample is given by

$$\mathcal{I}^*(\phi) = \frac{\mathcal{I}(\theta)}{\{g'(\theta)\}^2}.$$

The Cramér-Rao Lower Bound

Now, we shall derive an important inequality concerning the variance of an estimator, known as the *Cramér-Rao lower bound*.

Suppose that T an unbiased estimator for $m(\theta)$ (i.e. $\mathbb{E}(T) = m(\theta)$). Here $m(\theta)$ is just 'some function of' a parameter of interest, θ, and we note that we could have $m(\theta) = \theta$ (i.e. T unbiased for θ).

Additionally, we assume that we have a sample of data $\mathbf{X}$ with joint density $f(\mathbf{x}; \theta)$. We shall assume a continuous probability distribution for $\mathbf{X}$ and work under standard regularity conditions.

Similar results may be obtained in the discrete case by replacing integral signs with summations.

The Cramér-Rao Lower Bound

Recall that the score function is written

$$U(\theta; \mathbf{X}) = \frac{\partial}{\partial \theta} \log \mathcal{L}(\theta; \mathbf{X}) = \frac{\partial}{\partial \theta} \log f(\mathbf{x}; \theta).$$

Remember here that the likelihood function is simply the joint density but expressed as a function of θ rather than $\mathbf{x}$. We consider the correlation between T and $U(\theta; \mathbf{X})$

$$\text{Corr}(T, U(\theta; \mathbf{X})) = \frac{\text{Cov}(T, U(\theta; \mathbf{X}))}{\sqrt{\text{Var}(T)\text{Var}(U(\theta; \mathbf{X}))}}.$$

We know that $|\text{Corr}(T, U(\theta; \mathbf{X}))| \leq 1$. Therefore, squaring both sides of the above

$$\frac{[\text{Cov}(T, U(\theta; \mathbf{X}))]^2}{\text{Var}(T)\text{Var}(U(\theta; \mathbf{X}))} \leq 1.$$

The Cramér-Rao Lower Bound

Recall that $\text{Var}(U(\theta; \mathbf{X})) = \mathcal{I}(\theta)$. Then, we obtain

$$\frac{[\text{Cov}(T, U(\theta; \mathbf{X}))]^2}{\text{Var}(T)\mathcal{I}(\theta)} \leq 1$$

$$\implies \text{Var}(T) \geq \frac{[\text{Cov}(T, U(\theta; \mathbf{X}))]^2}{\mathcal{I}(\theta)} \tag{1}$$

Now, we consider $\text{Cov}(T, U(\theta; \mathbf{X}))$.

$$\begin{aligned}
\text{Cov}(T, U(\theta; \mathbf{X})) &= \mathbb{E}\left[(T - \mathbb{E}(T))\left(U(\theta; \mathbf{X}) - \mathbb{E}(U(\theta; \mathbf{X}))\right)\right] \\
&= \mathbb{E}\left[(T - m(\theta))\left(U(\theta; \mathbf{X}) - 0\right)\right] \\
&= \mathbb{E}\left[TU(\theta; \mathbf{X})\right] - m(\theta)\mathbb{E}\left[U(\theta; \mathbf{X})\right] \\
&= \mathbb{E}\left[TU(\theta; \mathbf{X})\right].
\end{aligned}$$

The Cramér-Rao Lower Bound

$$
\begin{aligned}
\mathbb{E}\left[TU(\theta;\mathbf{X})\right] &= \int_{\mathcal{X}} t(\mathbf{x}) \left\{ \frac{\partial}{\partial\theta} \left[\log f(\mathbf{x};\theta)\right] \right\} f(\mathbf{x};\theta)\mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}} t(\mathbf{x}) \left\{ \frac{\frac{\partial}{\partial\theta} f(\mathbf{x};\theta)}{f(\mathbf{x};\theta)} \right\} f(\mathbf{x};\theta)\mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}} t(\mathbf{x}) \frac{\partial}{\partial\theta} f(\mathbf{x};\theta)\mathrm{d}\mathbf{x} \\
&= \frac{\partial}{\partial\theta} \int_{\mathcal{X}} t(\mathbf{x}) f(\mathbf{x};\theta)\mathrm{d}\mathbf{x} \\
&= \frac{\partial}{\partial\theta} \mathbb{E}(T) \\
&= \frac{\partial}{\partial\theta} m(\theta) = m'(\theta)
\end{aligned}
$$

So $\mathrm{Cov}(T, U(\theta;\mathbf{X})) = m'(\theta)$ and substitution into (1) yields the result...

The Cramér-Rao Lower Bound

If T is an unbiased estimator for $m(\theta)$ then

$$\text{Var}(T) \geq \frac{[m'(\theta)]^2}{\mathcal{I}(\theta)}.$$

The term $\frac{[m'(\theta)]^2}{\mathcal{I}(\theta)}$ is known as the **Cramér-Rao lower bound**.

Notice that when T is **unbiased** for θ, $m'(\theta) = 1$ and the inequality becomes

$$\text{Var}(T) \geq \frac{1}{\mathcal{I}(\theta)}.$$

The Cramér-Rao lower bound tells us the minimum possible variance for a given estimator $T(\mathbf{X})$.

Learning Outcomes

- Understand the concept of **unbiasedness** with regard to estimators and be able assess whether or not an estimator is unbiased.
- The definition and importance of **mean squared error** and its relationship with bias and variance of an estimator.
- The property of **consistency** and how to assess whether or not an estimator is consistent for a given parameter.
- Definition of the **score function**, its properties and its relationship to the **Fisher information**.
- Derivation and definition of the **Cramér-Rao lower bound** for the variance of an estimator.