Point Estimation

Maura Mezzetti

Department of Economics and Finance
Università Tor Vergata

Outline

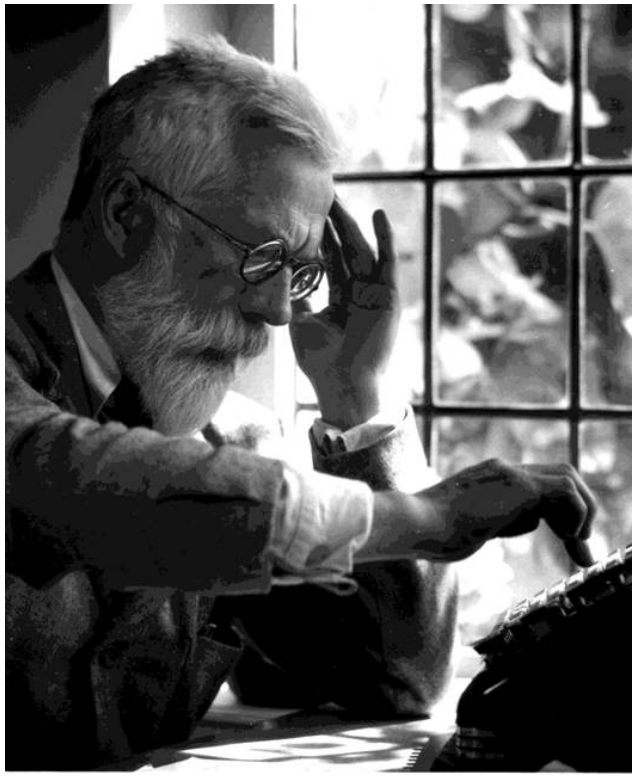
Point Estimation

"What! you have solved it already? "
"Well, that would be too much to say. I have discovered a suggestive fact, that is all."

Dr. Watson and Sherlock Holmes
The Sign of Four

History

The likelihood principle was first introduced by R.A. Fisher in 1922. The law of likelihood was identified by Ian Hacking.



History

"Modern statisticians are familiar with the notion that any finite body of data contains only a limited amount of information on any point under examination; that this limit is set by the nature of the data themselves the statistician's task, in fact, is limited to the extraction of the whole of the available information on any particular issue. " R. A. Fisher

The Likelihood

Definition The likelihood function is a function of the parameter with an observed sample, and is given by $L(\theta|x) = f(x|\theta)$.

Same expression, but now x is fixed and θ is variable.

Discrete Distributions

If $(X_1, \ldots, X_n)$ are discrete iid random variable with probability function $p(x|\theta)$, then, the likelihood function is given by

$$\begin{aligned} L(\theta|x) &= P(X_1 = x_1, \ldots, X_n = x_n|\theta) \\ &= \prod_{i=1}^{n} P(X_i = x_i|\theta) \\ &= \prod_{i=1}^{n} p(x_i|\theta) \end{aligned}$$

Likelihood

- For a given value of the parameter, the likelihood tells us how likely it is to see what we see, not viceversa.
- The notation $f(x|\theta)$ is only for convenience. It is *not* a conditional density. If $X_1, \ldots, X_n$ is an IID sample, then:

$$L(\theta|x) = f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Continuous Distributions

If $(X_1, \ldots, X_n)$ are continuous iid random variable with probability density function $f(x|\theta)$, then, the likelihood function is given by

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$

The Likelihood Principle

The Likelihood Principle: If x and y are two sample points such that $L(\theta|x)$ is proportional to $L(\theta|y)$, that is, there exists a constant $C(x, y)$ such that

$$L(\theta|x) = C(x, y)L(\theta|y) \qquad \forall\theta,$$

then the conclusions drawn from x and y should be identical.

Intuition

Suppose I tell you I have 100 cookies in my backpack. The cookies are of two types: chocolate chip cookies and fortune cookies. Moreover, I tell you that the number of fortune cookies is either 10 or 90. You draw a cookie out of my backpack at random and see that it is a fortune cookie.

Intuition

Based on this data, what is more likely: there are

1. 10 fortune cookies and 90 chocolate chip cookies, or
2. 90 fortune cookies and 10 chocolate chip cookies?

Based solely on one sample (fortune cookie), 2 is more likely.

Intuition

This is exactly the idea behind maximum likelihood estimation.

The method asks: what value of the a parameter is most consistent with the data?

Intuition

This is exactly the idea behind maximum likelihood estimation.

The method asks: what value of the a parameter is most consistent with the data?

In other words, what value of a parameter makes the data most likely?

Intuition

This is exactly the idea behind maximum likelihood estimation.

The method asks: what value of the a parameter is most consistent with the data?

In other words, what value of a parameter makes the data most likely?

Estimator

Definition (Casella Berger): A point estimator is any function $T(X_1, \ldots, X_n)$ of a sample. Any statistic is a point estimator.

Note: The definition makes no mention of any correspondence between the estimator and the parameter it is to estimate.

Definition: A point estimator or estimator of parameter θ is a statistic whose purpose is to estimate the value of the parameter θ

An estimator is a function of the sample, while an estimate is the realized value of an estimator (that is, a number).

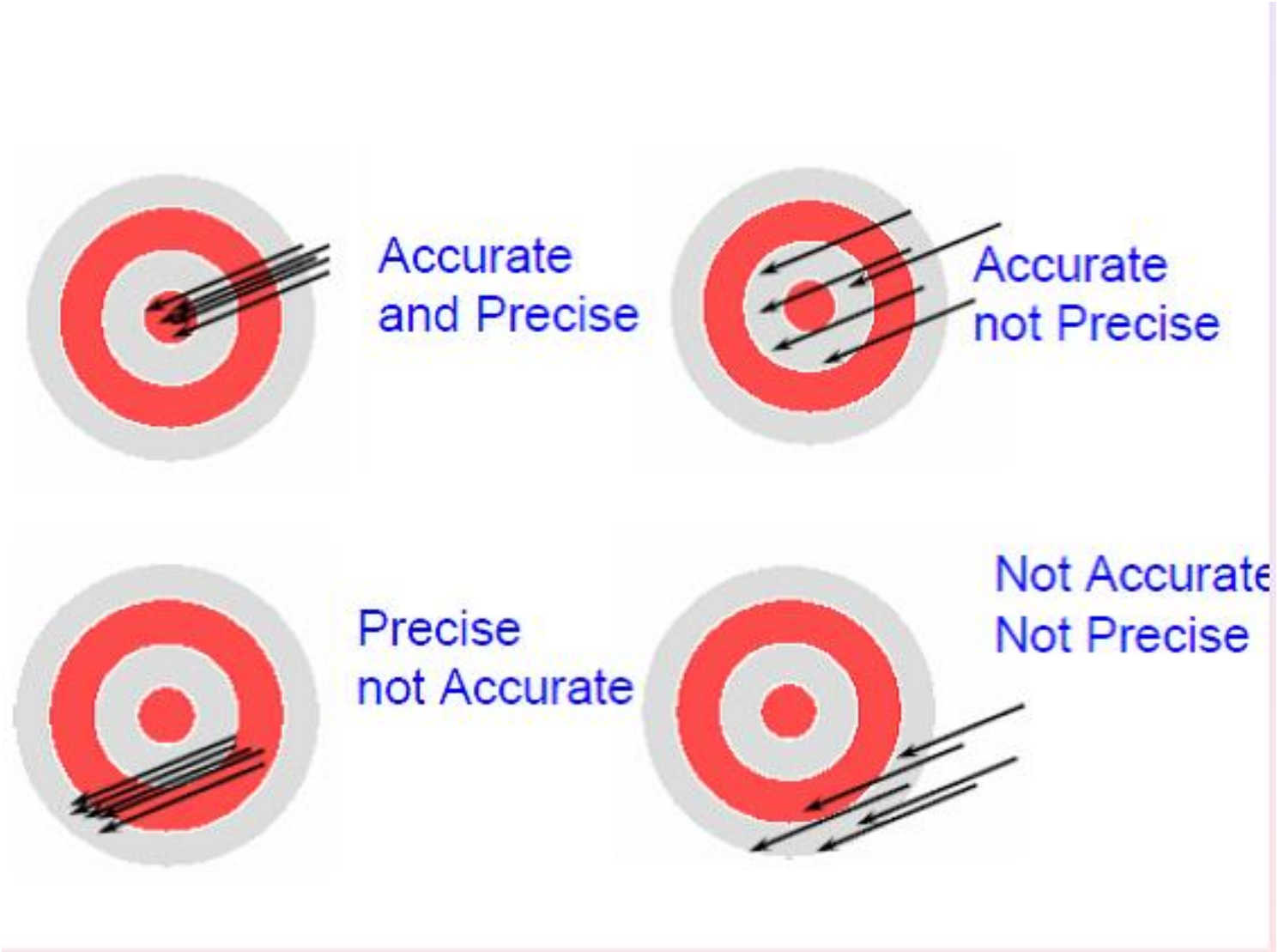
Estimation Property

What should we require to a good estimator? We will be considering various qualities that a "good" estimator should possess, but firstly, it should be KEPT IN MIND that, by virtue of it being a function of the sample values, an estimator is itself a random variable. So its behavior for different random samples will be described by a probability distribution.

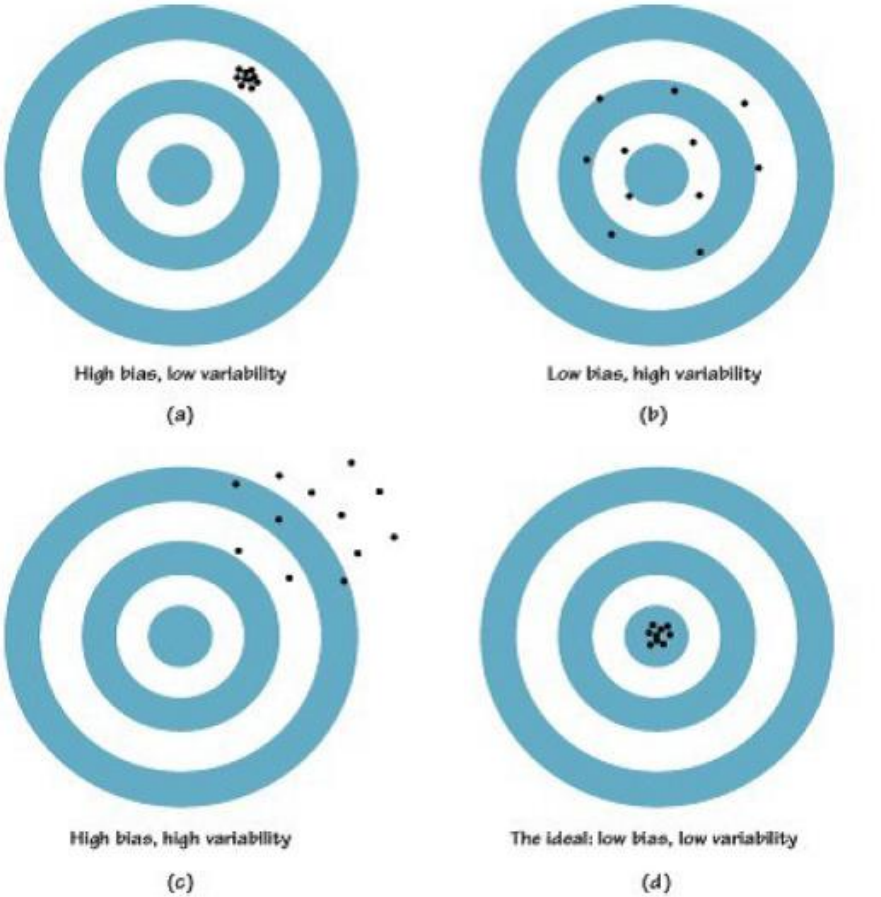
Estimation Property

It seems reasonable to require that the distribution of the estimator be somehow centered with respect to the parameter θ (accuracy). If it is not, the estimator will tend either to under-estimate or over-estimate θ. A further property that a good estimator should possess is precision, that is, the dispersion of the distribution should be small.

Estimation Property



Estimation Property



Unbiased Estimator

Definition The *bias* of a point estimator T, of a parameter θ, is the difference between the expected value of T and θ. $B_\theta(T) = E(T) - \theta$. An estimator whose bias is identically equal to zero (in θ) is called *unbiased estimator* of θ and satisfies: $E_\theta(T) = \theta$.

Example If $X_1, \ldots, X_n$ are i.i.d. Bernoulli with parameter p, the (silly) estimator $T = 1/2$ is biased, since $E_p[T] = 1/2 \neq p$. The sample mean $T = (X_1 + \ldots + X_n)/n$ is unbiased, since $E_p[T] = p$.

Possible undesirability of unbiasedness

- It is easy to see graphically and intuitively that unbiasedness may not be desirable if it comes at the cost of a higher estimation error.
- Biased but more precise estimators may be preferable to unbiased estimators
- Moreover, within the class of unbiased estimators we need to define other criteria to choose which estimator we prefer
- We now turn to the property of Mean Squared Error, which allows us to rank the desirability of a set of unbiased estimators.

Definition of Efficiency

- Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators of θ. If

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$$

  then $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$.
- The relative efficiency or relative precision of $\hat{\theta}_1$ with respect to θ_2

$$\frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$$

- Within the set of unbiased estimators we clearly prefer the most efficient ones

Mean Squared Error

- The most well-known criteria to evaluate an estimator is based on the **mean squared error** (MSE) which is a measure of the performance of an estimator $T = t(\mathbf{X})$ defined as the expected value of the (squared) estimation error, i.e.

$$MSE_\theta(T) = E_\theta\left[(T-\theta)^2\right]$$

- According to this criterion, the estimator $T^\star$ of θ is better than another estimator T if the MSE of $T^\star$ is uniformly smaller than that of T , i.e.

$$MSE_\theta(T^\star) \leq MSE_\theta(T), \quad \forall\theta$$

with at least one value of θ such that:

$$MSE_\theta(T^\star) < MSE_\theta(T).$$

Mean Squared Error

- The MSE is easier to treat analytically with respect to other measures of goodness of an estimator as, since it may be expressed as

$$MSE_\theta(T) = Var_\theta(T) + (B_\theta(T))^2,$$

  where $B_\theta(T) = E(T) - \theta$ is the Bias of the estimator.
- MSE incorporates two components, one measuring the variability of the estimator (precision), and the other measuring its bias (accuracy). An estimator that has good MSE properties has small combined variance and bias

Example

If $X_1, \ldots, X_3$ are i.i.d. as X, with expected value $E(X) = \mu$ and $Var(X) = \sigma^2$, consider the following estimators of μ

$$\begin{aligned} T_1 &= X_1 + X_2 - X_3 \\ T_2 &= \frac{X1+X_2+X_3}{4} \\ T_3 &= 0 \end{aligned}$$

Which estimator would you prefer?

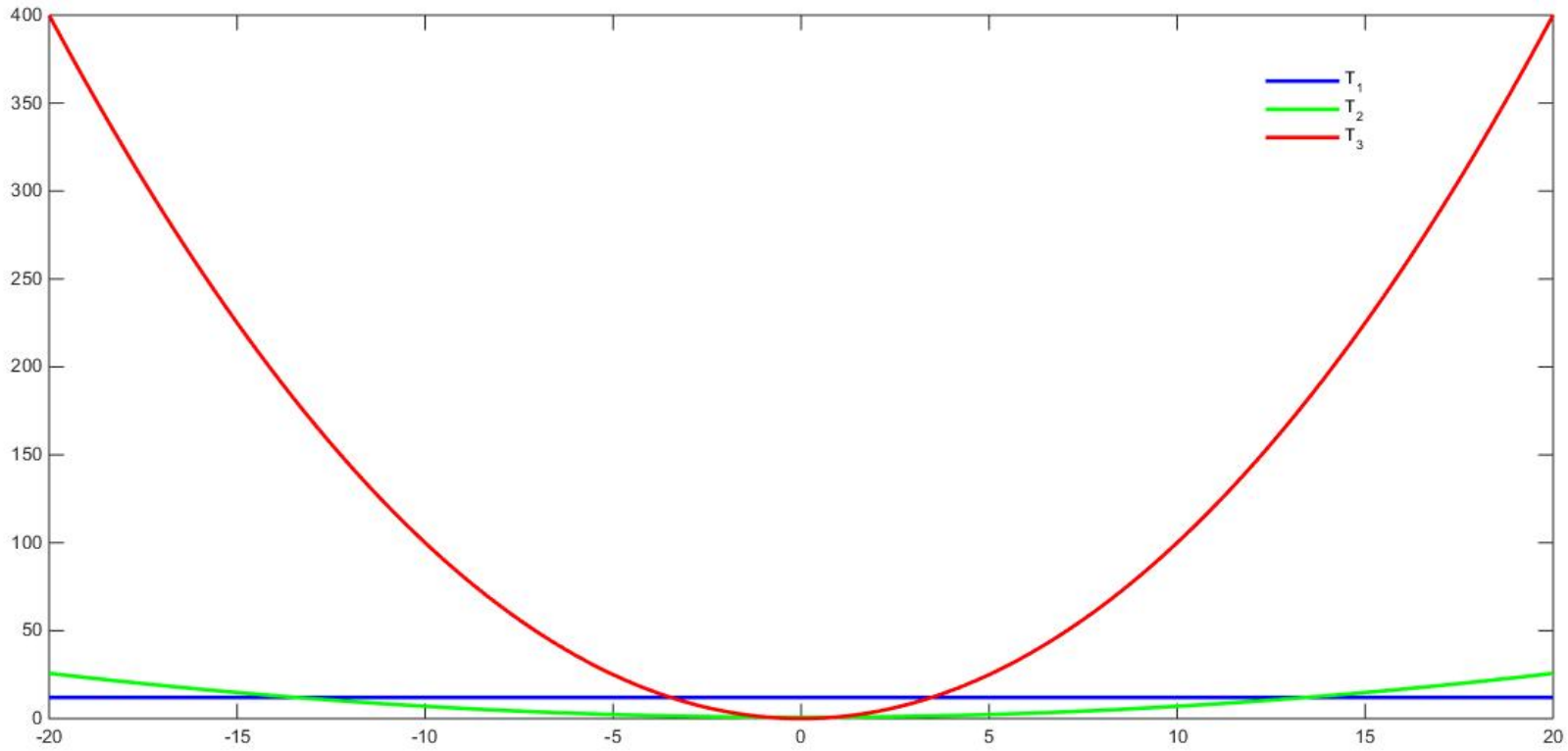
$$E(T_1) = E(X_1 + X_2 - X_3) = \mu \qquad Var(T_1) = Var(X_1 + X_2 - X_3) = 3\sigma^2$$
$$E(T_2) = \frac{E(X1+X_2+X_3)}{4} = 3\frac{\mu}{4} \qquad Var(T_2) = \frac{Var(X1+X_2+X_3)}{16} = \frac{3}{16}\sigma^2$$
$$E(T_3) = 0 \qquad Var(T_3) = 0$$

Example

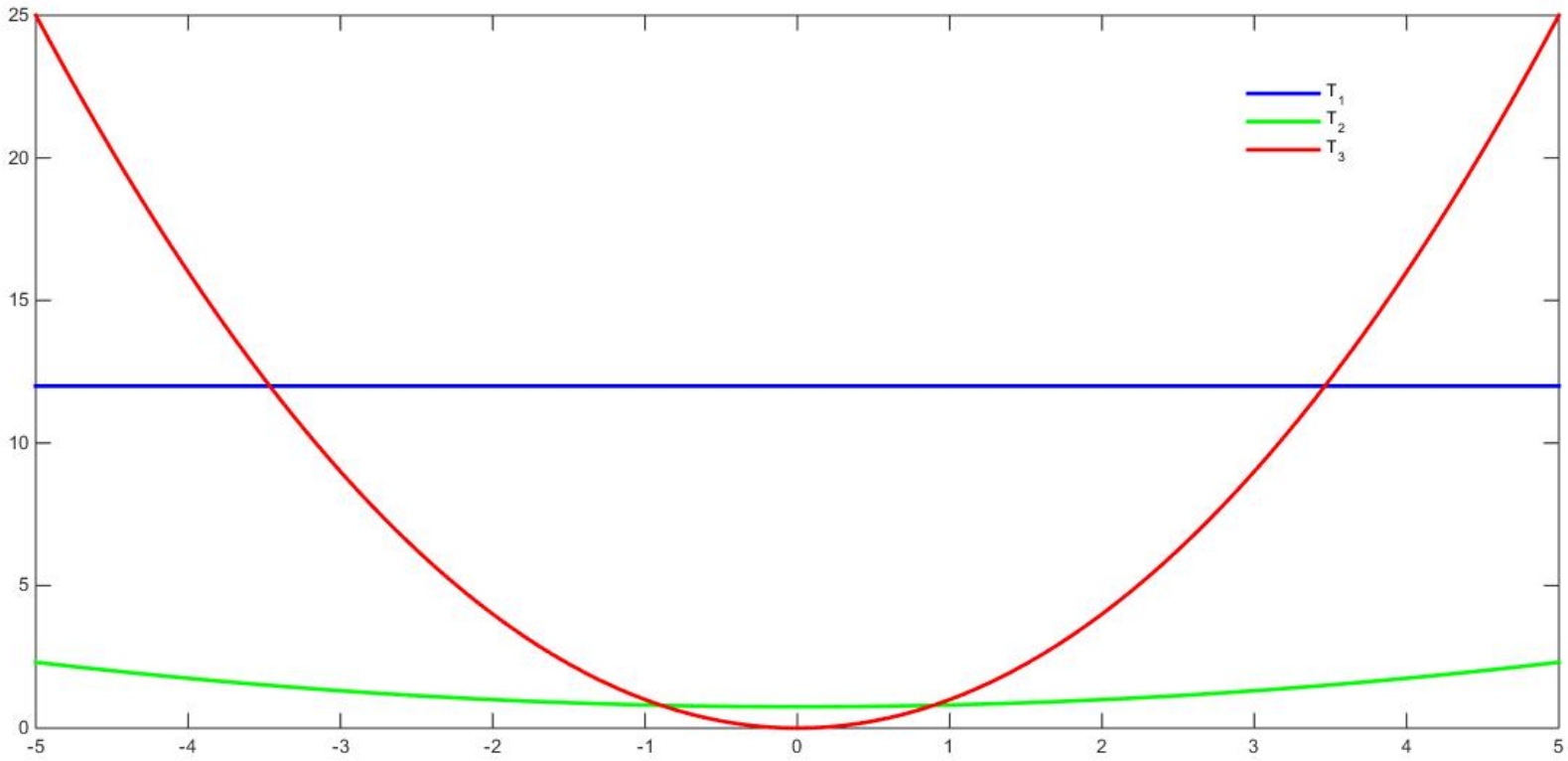
T_1 is the only unbiased estimator, T_3 is biased but it is the one with the lowest variance, let's compare the MSE.

$$MSE_\mu(T_1) = 3 \times \sigma^2$$
$$MSE_\mu(T_2) = \frac{3}{16}\sigma^2 + \frac{\mu^2}{16}$$
$$MSE_\mu(T_3) = \mu^2$$

Mean Squared Error



Mean Squared Error



Best Unbiased Estimator

- Comparing two estimators on the basis of the MSE is not always possible since it may happen that, between T_1 and T_2, the first estimator is better for certain values of θ, while the second is better for other values of θ.
- In general, it is not possible to find an estimator with uniformly minimum MSE and so the *best* estimator for a given inferential problem usually does not exist.
- The reason why there is not one *best* estimator of θ according to the MSE criterion is that the class of all the possible estimators is too large. To make the problem of finding the *best* estimator tractable, we can restrict the class of possible estimators, the class of unbiased estimators.

UMVUE Estimator

Definition: within the class of the unbiased estimators of θ, the estimator with minimum MSE corresponds to the *uniform minimum variance unbiased estimator* (UMVUE), i.e. the estimator $T^\star$ such that:

$$Var_\theta(T^\star) \leq Var_\theta(T), \quad ,\forall \theta \in \Theta$$

for any other unbiased estimator T of θ.

Consistency

- Some criteria of evaluating an estimator take into account the so-called *asymptotic properties*, i.e. the behavior of the estimator when the sample size becomes infinite. In contrast, the criteria previously illustrated (unbiasdness, efficiency) are referred to as *finite-sample criteria*.
- To introduce the main asymptotic criteria, we have to consider a *sequence of estimators*:

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

with n growing indefinitely.

Consistency

- Weak consistency: A sequence of estimators $\{T_n\}$ is a *consistent sequence of estimators* of θ if, for every $\varepsilon > 0$ and every $\theta \in \Theta$:

$$lim_{n\to\infty} P_\theta \left(|T_n - \theta| < \varepsilon \right) = 1$$

- In practice, this means that when the sample size becomes infinite, and so the information increases indefinitely, the estimator will be arbitrarily close to the parameter with high probability.
- Strong consistency: A sequence of estimators $\{T_n\}$ is a *consistent sequence of estimators* of θ if, for every $\theta \in \Theta$:

$$lim_{n\to\infty} MSE_\theta(T_n) = 0$$

Example

Let $(X_1, \ldots, X_n)$ be a random sample of i.i.d. random variables distributed as $N(\mu, \sigma^2)$. Consider the following estimator of μ:

$$T_n = \frac{1}{2}(X_1 + \bar{X}_n)$$

where X_1 is the first observed random variable and $\bar{X}_n$ is the sample mean based on n observations.

1. Find MSE of the estimator T_n and compare with MSE of $\bar{X}_n$
2. Is T_n a consistent estimator of μ?

Why are properties of estimators interesting?

We choose between estimators comparing their properties.
It is useful to distinguish between:

- Finite sample properties that hold for a given sample size n.
  - Unbiasedness
  - Efficiency
  - (Sufficiency)
- Asymptotic properties that hold when sample size goes to ∞:
  - Consistency
  - Asymptotic unbiasedness
  - Asymptotic efficiency
  - Asymptotic normality
- Other properties (e.g. Invariance)

Why are asymptotic properties important

We obviously always work with finite samples, but:

- we would feel unconfortable in using an estimator that had undesirable properties in the hypothetical case in which the sample size could go to ∞ (we reach the whole population).
- A finite sample may be sufficiently large for asymptotic results to hold with a very good approximation, even if its actual size is effectively not so large.
- Small sample properties are often difficult to characterize and less attractive than asymptotic properties.
- Asymptotic hypothesis testing is easy to define and perform, while it may be more problematic in a small sample.

Method of Finding Estimators

In some cases it is an easy task to decide how to estimate a parameter, and often intuition alone can lead us to a very good estimators. For example, estimating a parameter with its sample analogue is usually reasonable. In more complicated models, it is needed a more methodological way of estimating parameters. The most popular method of finding estimators:

- Method of Moments
- Least Squares Estimation
- Maximum Likelihood Estimation

Method of Moments

- This is one of the oldest and simplest methods of finding estimators. It dates back to K. Pearson in the late 1800s.
- Suppose that the random sample, $\mathbf{X} = (X_1, \ldots, X_n)$, is drawn from a statistical model in which the parameter θ has dimension k,
$$\theta = (\theta_1, \ldots, \theta_k)$$
- The **method of moments** consists in equating the first k population moments to the corresponding sample moments and solving the resulting system of equations with respect to θ.

Method of Moments

- The population and the sample moments of order r are defined, respectively, as:

$$\mu_r(\theta) = E_\theta(X^r) = \int_\infty^\infty x^r f(x;\theta) dx$$

$$M_r = \frac{1}{n}\sum_i X_i^r$$

- The system of equations to be solved with respect to θ is then:

$$\mu_i(\theta) = M_i$$

Method of Moments

Step 1 : Identify how many parameters the distribution has. (Let's say m).

Step 2 : Find the first m population moments.

Step 3 : Equalize each of the population moments to the corresponding sample moment.

Step 4 : Solve the system to find solutions to the parameters.

Step 5 : The solutions are the MoM estimators.

Example of Method of Moments

- Example: for uniform distribution $U(a, b)$, two unknown parameters. Equate the first 2 population moments to the corresponding sample moments

$$E(X) = \frac{a+b}{2} \qquad Var(X) = \frac{(b-a)^2}{12}$$

- Solve for a and b (can be a bit messy)

$$\bar{x} = \frac{a+b}{2} \qquad s^2 = \frac{(b-a)^2}{12}$$

Least Square Estimation

The method of least squares is about estimating parameters by minimizing the squared discrepancies between observed data, on the one hand, and their expected values on the other.
Application: linear regression *(see other slide)*

The Likelihood Function

Recall the definition of likelihood function:
Suppose that $(X_1, X_2, \ldots, X_n)$ are random variables with joint density or frequency function $f(x|\theta)$ where $\theta \in \Theta$. Given outcomes $X = x$,

$$L(\theta|x) = f(x|\theta)$$

for each possible sample $x = (x_1, \ldots, x_n)$, the likelihood function $L(\theta|x)$ is a real-valued function defined on the parameter space Θ. Since we are under the hypothesis of independent variables the likelihood function is

$$L(\theta|X) = L(\theta|X_1, \ldots, X_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

Maximum Likelihood Estimation (MLE)

- By far the most popular estimation method
- MLE is the parameter point for which observed data is most likely under the assumed probability model
- In practice, $L(\theta|x)$ is the probability (or density) of the observed data as a function of θ. So, it provides the evidence of the data in favor of any single value of θ in Θ: if for two values of θ, say $\theta^{(1)}$ and $\theta^{(2)}$, we have that $L(\theta^{(1)}) > L(\theta^{(2)})$, the probability of the observed sample is larger under $\theta^{(1)}$ and so more evidence exists in favor of this value of θ.

Maximum Likelihood Estimation (MLE)

Definition The maximum likelihood estimator (MLE) $\hat{\theta}$ of θ is the location at which $L(\theta|x)$ attains its maximum as a function of θ. Its numerical value is often called the maximum likelihood estimate.

Maximum Likelihood Estimation

- For an observed sample x, it is natural to estimate θ as the value of the parameter that maximizes $L(\theta)$. This leads to the maximum likelihood estimate (MLE) of θ that, formally, may be defined as the value:

$$\hat{\theta} = \hat{\theta}(x) \qquad such \quad that \qquad L(\hat{\theta}) = sup_{\theta \in \Theta} L(\theta)$$

- Consequently, the maximum likelihood estimator (MLE) is $\hat{\theta}(X)$
- In practice, the *mle* is the parameter value for which the observed sample is most likely.

How to find Maximum Likelihood Estimation (MLE)

There are essentially two distinct methods for finding MLEs:

- Direct maximization: Examine $L(\theta|x)$ directly to determine which value of θ maximizes $L(\theta|x)$ for a given sample $x_1, \ldots, x_n$. This method is particularly useful when the range (or support) of the data depends on the parameter.
- Likelihood equations: If the range of the data does not depend on the data, the parameter space Θ is an open set, and the likelihood function is differentiable with respect to $\theta = (\theta_1, , \theta_p)$ over Θ, then the maximum likelihood estimate $\hat{\theta}$ satisfies the equations

$$\frac{\partial}{\partial \theta_k} L(\hat{\theta}|x) = 0 \qquad \text{for } k = 1, \ldots, p$$

These equations are called the likelihood equations.

How to find MLE

- Most of the times, to find *MLE* require to solve an optimization problem making use of differential calculus. It is usually simpler to maximize the *log-likelihood*, since the logarithm is a monotonic increasing transformation, the two problems are equivalent.
- In the uniparametric case, to maximize likelihood we have first to solve the likelihood equation:

$$\frac{d}{d\theta} l(\theta) = \sum_i \frac{f'(x_i;\theta)}{f(x_i;\theta)} = 0$$

How to find MLE

- Equating to zero the likelihood equation, we find one or more candidates for the mle since the first derivative being zero is only a necessary condition for a maximum.
- A way to verify if a root of the equation above is indeed a local maximum, we can check if the second derivative is negative.
- Although, even if we have found only one local maximum, to be sure that it is also a global maximum we have to check the value of $l(\theta)$ at the boundary of the parameter space.

Example Maximum Likelihood Estimation

- Suppose that $X \sim N(\mu, 1)$, with μ unknown. The likelihood function for the parameter $\theta = \mu$ is:

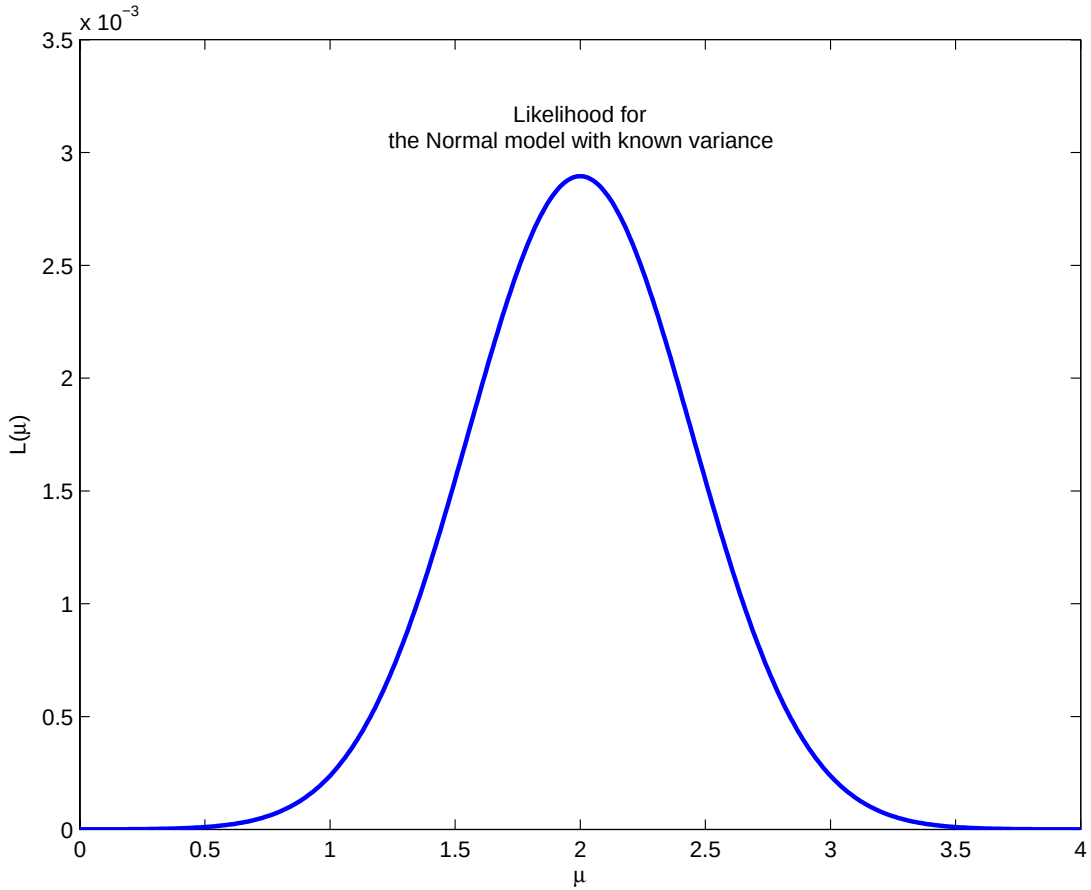
$$\begin{aligned} L(\mu) &= \prod_i \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right) = \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_i (x_i - \mu)^2\right) \end{aligned}$$

- Let, for example, the observed sample be:

$$\mathbf{x} = (2, 2.5, 1.5, 3, 1)$$

Example Maximum Likelihood Estimation: continue

As it is evidenced by the figure likelihood reaches its maximum when μ is equal to 2, that is the sample mean.



Example Maximum Likelihood Estimation: continue

$$\frac{dlogL(\mu)}{d\mu} = \sum_i (x_i - \mu)$$

$$\frac{dlogL(\mu)}{d\mu} = 0 \qquad when \qquad \mu = \bar{x}$$

$$\frac{d^2 logL(\mu)}{d^2\mu} = -n$$

Example Maximum Likelihood Estimation: Bernoulli

- For a Bernoulli distribution, the probability distribution is:

$$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

- Given a sample of N observations, the joint distribution of

$$p(x_1, \ldots, x_N|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

- Suppose we have observed the sample $x = (0, 0, 0, 1, 0, 0, 1, 1)$.

Example Maximum Likelihood Estimation: Bernoulli

- The likelihood for the Bernoulli distribution

$$L(\theta|x_1,\ldots,x_N) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$L(\theta|x_1,\ldots,x_N) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{1-\sum_{i=1}^{n} x_i}$$

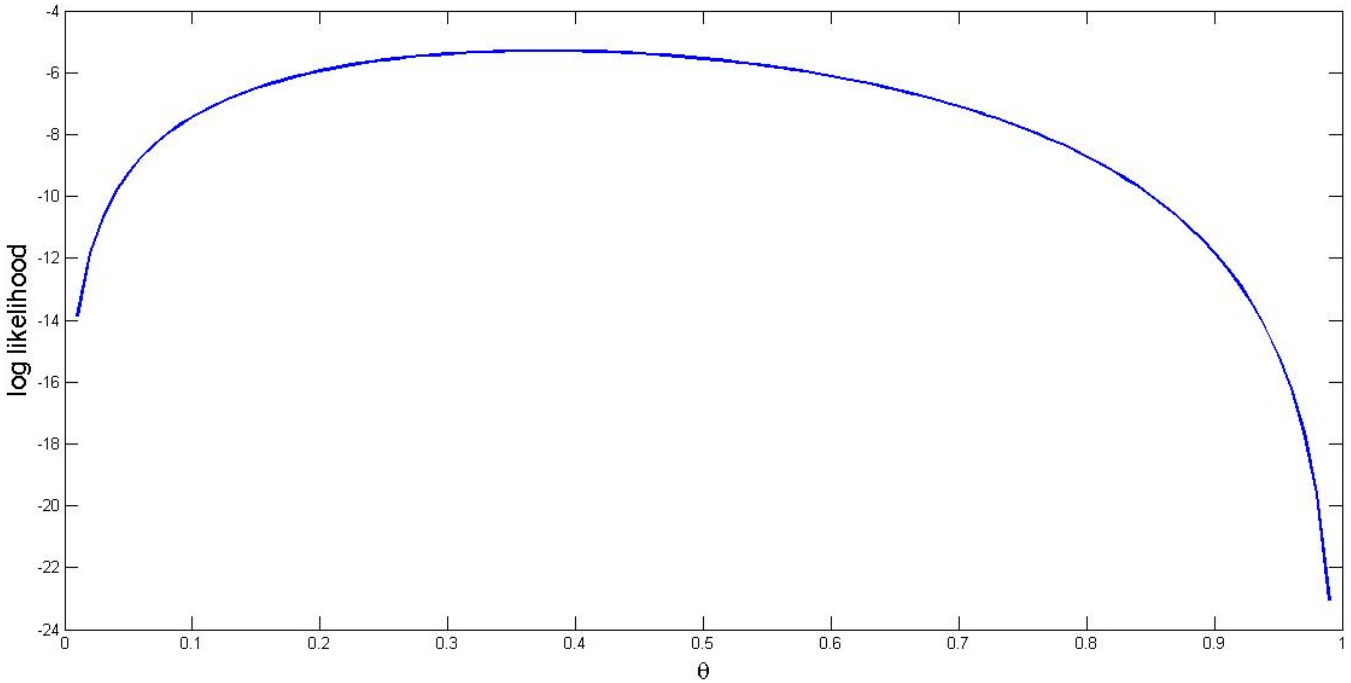
- Given the observed sample, the likelihood can be written as:

$$L(\theta|x_1,\ldots,x_N) = \theta^3(1-\theta)^5$$

- Sometimes is easier to deal with logarithm of likelihood

$$log\ L(\theta|x_1,\ldots,x_N) = 3 \times log\,(\theta) + 5 \times log\,(1-\theta)$$

Example Maximum Likelihood Estimation: Bernoulli



Example Maximum Likelihood Estimation: Bernoulli

Maximise the function for θ differentiating it with respect to θ:

$$
\begin{aligned}
log\, L(\theta|x_1, \ldots, x_N|\theta) &= 3 \times log\,(\theta) + 5 \times log\,(1-\theta) \\
\frac{\partial log\, L(\theta|x_1, \ldots, x_N|\theta)}{\partial \theta} &= \frac{3}{\theta} - \frac{5}{1-\theta} \\
0 &= \frac{3}{\theta} - \frac{5}{1-\theta} \\
\hat{\theta} &= 3/8 = 0.375
\end{aligned}
$$

How to find MLE: multiparametric case

- When we have a vector of parameters we have to maximize a function of several variables and so the problem is more complex.
- In simple cases, the maximization at issue may be performed by solving the system of linear equations:

$$\frac{\partial}{\partial \theta_k} l(\theta) = 0$$

- Then, it is necessary to check that the matrix of the second derivatives is negative definite.
- In this case, verifying that we have really found a global maximum may be a difficult task.

How to find MLE: multiparametric case

To use two variate case, to verify $log\, L(\theta_1, \theta_2)$ has a local maximum, it must ne shown the following three conditions hold:

1. The first-order partial derivatives are 0,

$$\frac{\partial}{\partial \theta_1} logL(\theta_1, \theta_2)\Big|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} = 0 \quad \frac{\partial}{\partial \theta_2} logL(\theta_1, \theta_2)\Big|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} = 0$$

2. At least one second-order partial derivative is negative,

$$\frac{\partial^2}{\partial^2 \theta_1} logL(\theta_1, \theta_2)\Big|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} < 0 \; \frac{\partial^2}{\partial^2 \theta_2} logL(\theta_1, \theta_2)\Big|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} < 0$$

How to find MLE: multiparametric case

3. The Determinant of Hessian matrix of the second-order partial derivative is positive,

$$\left| \begin{array}{cc} \frac{\partial^2}{\partial^2 \theta_1} logL(\theta_1, \theta_2) & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} logL(\theta_1, \theta_2) \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_2} logL(\theta_1, \theta_2) & \frac{\partial^2}{\partial^2 \theta_2} logL(\theta_1, \theta_2) \end{array} \right|_{\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2} > 0$$

Properties of Maximum Likelihood

- Maximum likelihood estimators have many interesting properties (see Casella Berger)
- MLE of θ depends on the sample only through *the sufficient statistic* for this parameter (recall Factorization Theorem).
- Invariance property (see next slide): is that if $\hat{\theta}$ is the MLE of θ, then, for any function $\tau(\theta)$, the MLE of $\lambda = \tau(\theta)$, is $\hat{\lambda} = \tau(\hat{\theta})$

Induced Likelihood

If $\eta = \tau(\theta)$ is a parametric function, then the likelihood for θ is defined by

$$L^\star(\eta|x) = \sup_{\theta:\tau(\theta)=\eta} L(\theta|x)$$

Theorem (Invariance Principle) If $(\hat{\theta})$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Properties of Maximum Likelihood Estimator

- Maximum likelihood estimation is:
  - Asymptotically unbiased
  - Consistent
  - Asymptotically normally distributed
- When the sample is large enough, we can approximate the variance covariance matrix V with inverse of information matrix (the expected value of the matrix of the second derivatives of $l(\theta)$ with negative sign):

$$V = [I_n]^{-1}$$

- If exists an estimator which is Best Unbiased Estimator (BUE) this is the maximum likelihood estimator (see Cramer Rao inequality)

Example of Maximum Likelihood Estimator

In a experiment on *DNA sequencing* we know that we can obtain four possible sequences with probabilities $p_1 = 1 - \theta$, $p_2 = \theta - \theta^2$, $p_3 = \theta^2 - \theta^3$ and $p_4 = \theta^3$ respectively, where $0 \leq \theta \leq 1$.
Let's consider n independent realizations of this experiment, where n_1, n_2, n_3, n_4, with $\sum_{i=1}^{4} n_i = n$ are observed occurrences for each of the 4 possible sequences.

Example of Maximum Likelihood Estimator

- Find the joint distribution of (n_1, n_2, n_3, n_4).
- Show that the Maximum Likelihood Estimator of θ is given by

$$\hat{\theta} = \frac{n_2 + 2n_3 + 3n_4}{n_1 + 2n_2 + 3n_3 + 3n_4}$$

Example of Maximum Likelihood Estimator: Solution

$$
\begin{aligned}
L(\theta) &= (1-\theta)^{n_1} \times (\theta - \theta^2)^{n_2} (\theta^2 - \theta^3)^{n_3} \theta_3^{n_4} \\
l(\theta) &= n_1 \times log(1-\theta) + n_2 \times log(\theta - \theta^2) + n_3 \times log(\theta^2 - \theta^3) + n_4 \times log(\theta^3) \\
l(\theta) &= n_1 \times log(1-\theta) + n_2 \times log(\theta) + n_2 \times log(1-\theta) + 2n_3 \times log(\theta) + \\
&\quad + n_3 \times log(1-\theta) + 3n_4 \times log(\theta) \\
l(\theta) &= (n_1 + n_2 + n_3) \times log(1-\theta) + (n_2 + 2n_3 + 3n_4) \times log(\theta) \\
\frac{\partial l(\theta)}{\partial \theta} &= -\frac{(n_1 + n_2 + n_3)}{1-\theta} + \frac{(n_2 + 2n_3 + 3n_4)}{\theta} \\
\frac{\partial l(\theta)}{\partial \theta} &= 0 \\
\hat{\theta} &= \frac{(n_2 + 2n_3 + 3n_4)}{(n_1 + 2n_2 + 3n_3 + 3n_4)}
\end{aligned}
$$