Lecture 2: Parameter estimation

EMAT30007 Applied Statistics

Nikolai Bode

Department of Engineering Mathematics

Outline of the lecture

In this lecture you will learn:

- How to use the **maximum likelihood method** to estimate the parameters of a probability distribution.
- How to use the **method of moments** to estimate the parameters of a probability distribution.
- How to determine if a parameter estimator is biased.

Lecture 2, Topic 1 – Maximum Likelihood

- **Content covered**
  - Concept of parameter estimation in statistics
  - Likelihood function
  - Maximum Likelihood Estimation (MLE)

Statistical inference and parameter estimation

Statistical inference is the process of drawing conclusions about a population based on a sample or subset of observations.

Parameter estimation is a type of statistical inference where one assumes that the population is well described by a probability distribution and would like to infer the values of its parameters.

Problem statement

Assume that your data set of n observations $x_1, \ldots, x_n$ is generated by n identical and independent Random Variables $X_1, \ldots, X_n$, which all have the same Probability Density Function $P_X(x; \theta)$ with parameter(s) θ.

A parameter estimator $\hat{\theta}$ is a function or a method to calculate the parameter(s) θ of $P_X(x; \theta)$ using the data $x_1, \ldots, x_n$.

Example: *Coin tosses*

We want to estimate the probability of a head when tossing a given coin.
Assume that a coin toss can be described by a Bernoulli RV with distribution

$$X \sim Bernoulli(p) = P_X(x;p) = p^x(1-p)^{1-x} = \begin{cases} p & \text{if } x = 1, \text{ head} \\ 1-p & \text{if } x = 0, \text{ tail} \end{cases}$$

where p, the probability of a head, is the parameter we want to estimate.
The coin is tossed $n = 10$ times and we observe $x = [1, 1, 0, 1, 1, 1, 1, 0, 1, 0]$.
What is a good estimator $\hat{p}(x)$ of p given the data x?

Example: *Student heights*

We want to estimate the mean height of students in the University.
We assume that the height of a student is described by a Normal RV with PDF

$$X \sim Normal(\mu, \sigma^2) = P_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ and σ^2 are the mean and variance height.
We measure the height of $n = 10$ students chosen at random and we observe
$x = [176, 165, 189, 180, 172, 169, 162, 161, 183, 170]$ cm.
What is a good estimator $\hat{\mu}(x)$ for the mean height of *all* students μ given the data x?

Likelihood function

Intuition: the idea of the Maximum Likelihood method is to find the parameter values that maximise the probability to observe (or generate) the sample data.

The Likelihood function $L(\theta; x_1, \ldots, x_n)$ takes in input a given value of the parameters, θ, and returns the conditional probability $p(x_1, \ldots, x_n|\theta)$ of getting the data that were observed $x_1, \ldots, x_n$ if the parameter value was really θ.

If data $x_1, \ldots, x_n$ are drawn independently from the distribution $P_X(x; \theta)$, then the likelihood is

Product of probabilities

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} P_X(x_i; \theta) \tag{1}$$

↖ PMF

Example: *Likelihood of n coin tosses*

A coin toss is a Bernoulli RV with distribution $P_X(x; p) = p^x(1-p)^{1-x}$, with $x = 0$ or 1.
The coin is tossed $n = 10$ times and we observe $x = [1, 1, 0, 1, 1, 1, 1, 0, 1, 0]$.
The likelihood function is
$L(\theta; x) = \prod_{i=1}^{n} P_X(x_i; \theta) = p \cdot p \cdot (1-p) \cdot p \cdot p \cdot p \cdot p \cdot (1-p) \cdot p \cdot (1-p) = p^7(1-p)^3$

Maximum likelihood

The Maximum Likelihood Estimator (MLE), $\hat{\theta}_{MLE}$, is the parameter value that maximises the Likelihood function:

$$\hat{\theta}_{MLE}(x) = \arg\max_{\theta} L(\theta; x)$$

In practice, it is often more convenient to maximise the log-Likelihood function $logL(\theta; x) = \log(L(\theta; x))$. Note that $logL(\theta; x)$ and $L(\theta; x)$ have the same maximum because the logarithm is a strictly increasing function.

In some cases it is possible to find an analytical expression for $\hat{\theta}_{MLE}$ as follows:

- Differentiate the (log-)Likelihood function with respect to its parameter θ and find $\hat{\theta}$ that solves the equation

$$\frac{\partial}{\partial\theta} L(\theta; x) = 0 \tag{2}$$

- Verify that the solution of Eq. (2) $\hat{\theta}$ is a maximum of $L(\theta)$: $\frac{\partial^2}{\partial\theta^2} L(\theta; x)\Big|_{\hat{\theta}} < 0$

Example: *Maximum Likelihood for a Bernoulli distribution*

Given the n independent observations $x_1, \ldots, x_n$ from Bernoulli RVs with distribution $P_X(x;p) = p^x(1-p)^{1-x}$, the Likelihood can be computed using Eq. (1):

$$\begin{aligned} L(p;x) &= \prod_{i=1}^{n} P_X(x_i;p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = \\ &= p^{x_1}(1-p)^{1-x_1} \cdot \ldots \cdot p^{x_n}(1-p)^{1-x_n} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} \end{aligned}$$

The log-Likelihood is

$$logL(p) = \log(L(p)) = (\textstyle\sum_i x_i)\log p + (n - \sum_i x_i)\log(1-p).$$

To find the $\hat{p}_{MLE}$ let's take the derivative of $logL$ with respect to p and set it to 0:

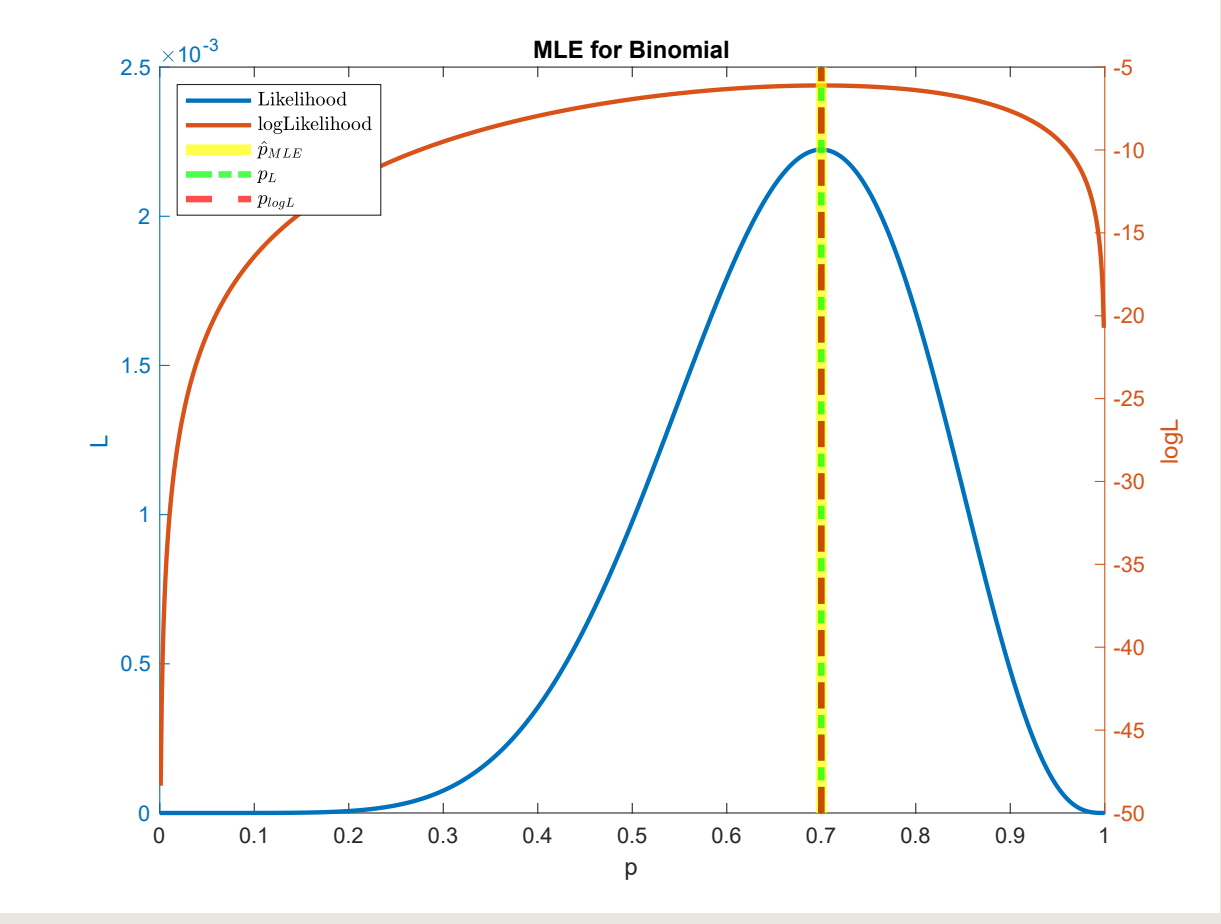
$$\frac{d}{dp} logL(p) = \frac{d}{dp}[(\textstyle\sum_i x_i)\log p + (n - \sum_i x_i)\log(1-p)] =$$
$$(\textstyle\sum_i x_i)/p - (n - \sum_i x_i)/(1-p) = 0$$

Multiplying both sides by $p(1-p)$ we get $(1-p)\sum_i x_i - np + p\sum_i x_i = 0$ and

$$\hat{p}_{MLE} = \frac{1}{n}\textstyle\sum_i x_i = \bar{x}$$

← Mean

Example: *Maximum Likelihood for a Bernoulli distribution*



Likelihood and log-Likelihood functions computed on data $x = [1, 1, 0, 1, 1, 1, 1, 0, 1, 0]$, with their maxima and the MLE estimate $\hat{p}_{MLE}$ at 0.7.

Example: *Maximum Likelihood for a Normal distribution*

Given the n independent observations $x_1, \ldots, x_n$ from Normal RVs with distribution $P_X(x; \mu, v) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-\mu)^2}{2v}\right]$, with variance $v = \sigma^2$, the log-Likelihood is:

大 PDF

$$logL(\mu, v; x) = \log\left(\prod_{i=1}^{n} P_X(x_i; \mu, v)\right) = -n \log\sqrt{2\pi v} - \frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2$$

To find $\hat{\mu}_{MLE}$, let's take the derivative of $logL$ with respect to μ and set it to 0:

$$\frac{\partial}{\partial \mu} logL = \frac{1}{2v}\sum_i 2(x_i - \mu)(-1) = 0 \quad \Rightarrow \quad \hat{\mu}_{MLE} = \frac{1}{n}\sum_i x_i = \bar{x}$$

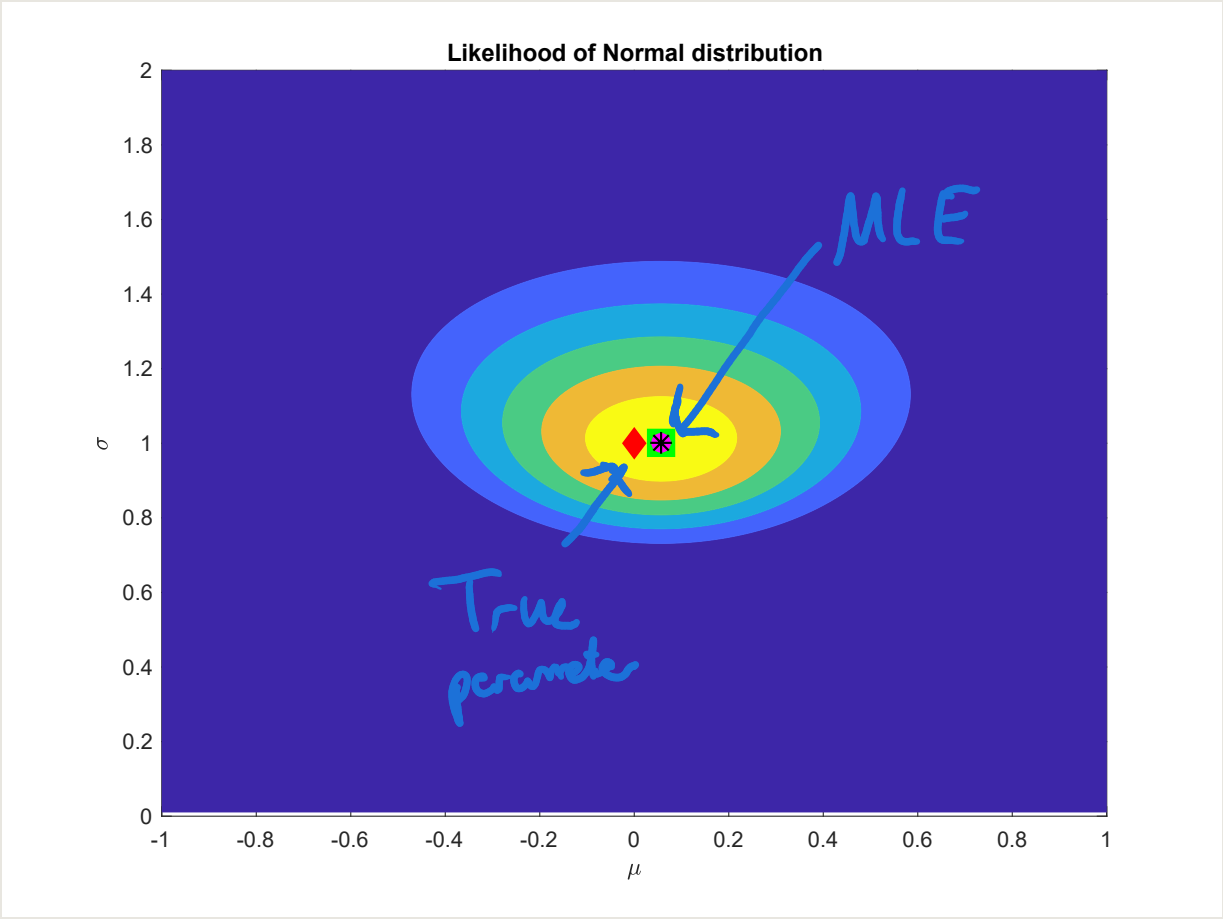
| Sample Mean

To find $\hat{v}_{MLE}$, let's take the derivative of $logL$ with respect to v and set it to 0:

$$\frac{\partial}{\partial v} logL = -\frac{n}{2v} + \frac{\sum_i (x_i - \mu)^2}{2v^2} = 0 \quad \Rightarrow \quad \hat{v}_{MLE} = \frac{1}{n}\sum_i (x_i - \hat{\mu}_{MLE})^2$$

| Sample variance

Example: *Maximum Likelihood for a Normal distribution*



Likelihood function of a Normal distribution computed on data $x = [-1.16, -0.04, -0.36, 1.53, 0.31, 0.37, 1.15, 1.24, -1.25, -1.22]$. The maxima of Likelihood and log-Likelihood functions coincide with the MLE estimate at $(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE}) = (0.06, 1.00)$. The red diamond denotes the true parameters at $(0, 1)$.

Lecture 2, Topic 2 – Method of Moments

Content covered

- Moments of probability distributions
- Method of Moments for parameter estimation

Moments of a probability distribution

The k^{th} theoretical moment of the RV X with distribution P_X depends on its parameters, θ, and is defined as the *expectation value* of X^k, for $k = 1, 2, ...$:

$$E(X^k; \theta) \equiv \int P_X(x; \theta) x^k dx \quad (3)$$

integral becomes a sum for discrete distribution.

For example, the first theoretical moment $E(X^1) = \mu$ is the mean.

The k^{th} empirical or sample moment of the sample observations $x_1, \ldots, x_n$ is defined as

$$M_k \equiv \frac{1}{n} \sum_{i=1}^{n} x_i^k \quad (4)$$

Sum over observations

For example, the first sample moment $M_1 = \bar{x}$ is the sample average.

about mean

Moments about the mean

For $k > 1$, the k^{th} theoretical moment about the mean is defined as $E[(X - \mu)^k]$ and the k^{th} sample moment about the mean is defined as $M_k^* = \frac{1}{n} \sum_i (x_i - \bar{x})^k$.

Example: *Moments of a Bernoulli distribution*

Let's use Eq. (3) to compute the first two theoretical moments of a Bernoulli distribution $P_X(x;p) = p^x(1-p)^{1-x}$:

- The first theoretical moment is $E(X) \equiv \sum_{x=0,1} p^x(1-p)^{1-x} \cdot x = p$.
- The second theoretical moment is $E(X^2) \equiv \sum_{x=0,1} p^x(1-p)^{1-x} \cdot x^2 = p$.

Example: *Moments of a Normal distribution*

The first two moments of a Normal distribution $P_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ are:

- $E(X) = \mu + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{z^2}{2\sigma^2}\right] z dz = \mu + 0$, using the change of variable $z = x - \mu$ and noting that $P_X(x;0,\sigma^2)$ is symmetric about 0.
- First note the relationship between the second moment and the second moment about the mean (the variance): $E(X^2) = E[(X-\mu)^2] + \mu^2$. The variance is $E[(X-\mu)^2] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} (x-\mu)^2 dx = \sigma^2 \int_{-\infty}^{\infty} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} z^2 dz = \sigma^2$ with the change of variable $z = \frac{(x-\mu)}{\sigma}$ and solving the last integral by parts.

Method of Moments to estimate parameters

Intuition: the idea of the Method of Moments is to find the parameter values for which the theoretical moments are equal to the sample moments.

The Method of Moments (MoM) works as follows:

1. Compute the first m **theoretical moments** (Eq. 3), $E(X^k; \theta)$ for $k = 1, ..., m$, where m is the number of parameters of P_X.
2. Compute the first m **sample moments** (Eq. 4), M_k for $k = 1, ..., m$, using the n observations $x_1, \ldots, x_n$.
3. Write the system of m equations obtained by equating each k^{th} theoretical moment with the corresponding sample moment for $k = 1, ..., m$, and solve for the m unknown parameters:

$$\begin{cases} E(X^1; \theta_1, ..., \theta_m) = M_1 \\ \quad \ldots \\ E(X^m; \theta_1, ..., \theta_m) = M_m \end{cases} \quad \text{or} \quad \begin{cases} E[X^1; \theta_1, ..., \theta_m] = M_1 \\ \quad \ldots \\ E[(X - \mu)^m; \theta_1, ..., \theta_m] = M_m^* \end{cases}$$

Example: *Method of Moments for a Bernoulli distribution*

The Bernoulli distribution has one parameter, p, so we only need one equation for the first moment, where we use the result $E(X;p) = p$ derived previously:

$$E(X;p) = \bar{x} \quad \Rightarrow \quad \hat{p}_{MM} = \frac{1}{n}\sum_i x_i$$

Example: *Method of Moments for a Normal distribution*

The Normal distribution has two parameters, μ and σ^2, so we need the two equations for the first two moments. Using the results for $E(X)$ and $E(X^2)$ derived previously:

$$\begin{cases} E(X) = \mu = \frac{1}{n}\sum_i x_i = \bar{x} \quad \Rightarrow \quad \mu = \bar{x} \\ E(X^2) = \sigma^2 + \mu^2 = \frac{1}{n}\sum_i x_i^2 \quad \Rightarrow \quad \sigma^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 \end{cases}$$

where in the last line we substituted the RHS of the first equation for μ. We get:

$$\hat{\mu}_{MM} = \frac{1}{n}\sum_i x_i \qquad \text{and} \qquad \hat{\sigma^2}_{MM} = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

Note that these estimates are equal to those of the Maximum Likelihood method.

Example: *Estimate the parameters of a Gamma distribution*

The Gamma distribution has PDF $P_X(x;\alpha,\theta) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\theta^\alpha}e^{-x/\theta}$, which depends on two parameters, α and θ.
Using the Maximum Likelihood method would be difficult because we should take the derivative of the Gamma function $\Gamma(\alpha)$.
To apply the MoM we need two equations for the first two moments. The first two theoretical moments about the mean of a Gamma distribution are $E(X) = \alpha\theta$ and $E[(X-\mu)^2] = \alpha\theta^2$. The system of equations of the MoM is:

$$\begin{cases} E(X) = \alpha\theta = \frac{1}{n}\sum_i x_i = \bar{x} \quad \Rightarrow \quad \alpha = \bar{x}/\theta \\ E[(X-\mu)^2] = \alpha\theta^2 = \frac{1}{n}\sum_i (x_i-\bar{x})^2 \quad \Rightarrow \quad \theta = \frac{1}{n\bar{x}}\sum_i (x_i-\bar{x})^2 \end{cases}$$

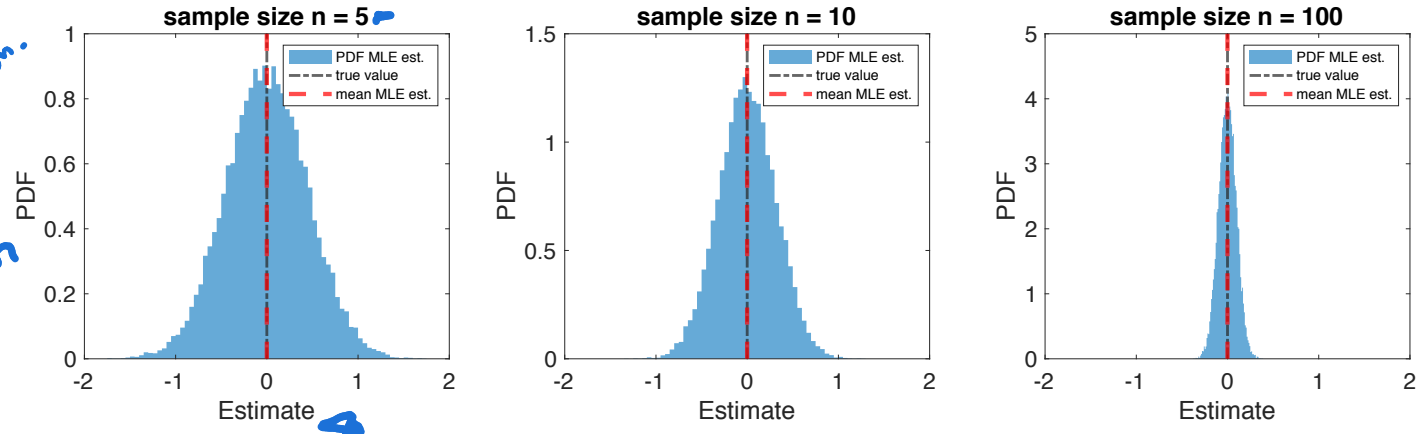
We get:

$$\hat{\alpha}_{MM} = \frac{\bar{x}}{\hat{\theta}_{MM}} = \frac{n\bar{x}^2}{\sum_i (x_i-\bar{x})^2} \qquad \text{and} \qquad \hat{\theta}_{MM} = \frac{1}{n\bar{x}}\sum_i (x_i-\bar{x})^2$$

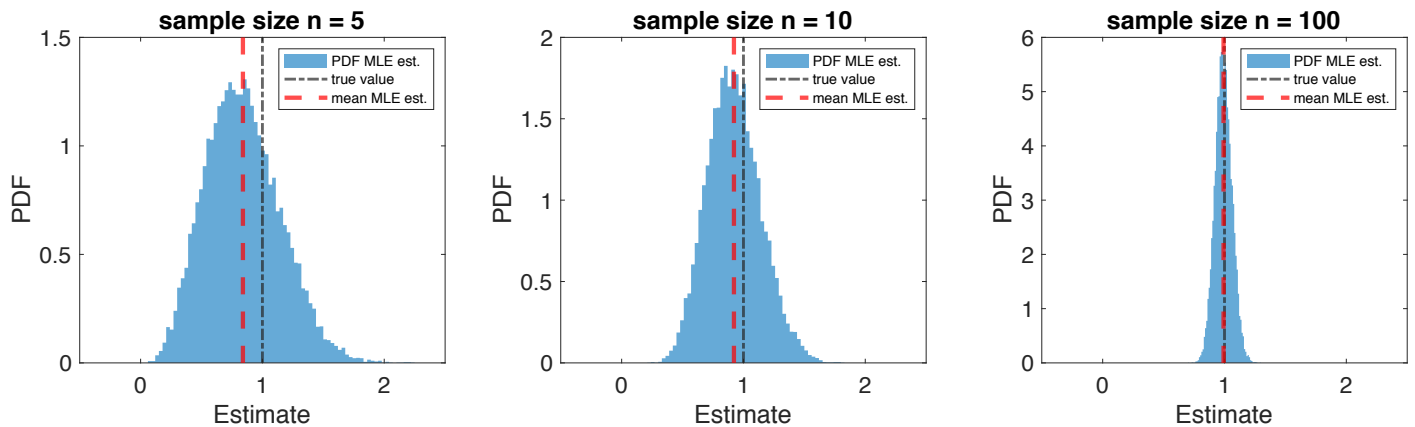
Lecture 2, Topic 3 – Biased and unbiased estimators

- **Content covered**
  - Biased and unbiased estimators
  - Pros/Cons of parameters estimation methods

Biased estimators – intuition



An estimator is unbiased if its mean over many samples is **centred around the true value, irrespective of the sample size** (top figure).



The estimator in the bottom figure is biased because for small samples its mean is skewed to the left of the true value (dotted line).

Unbiased estimators

A parameter estimator $\hat{\theta}(X_1, ..., X_n)$ is unbiased if its expectation value is equal to the true value θ:

$$E[\hat{\theta}(X_1, ..., X_n)] = \theta \tag{5}$$

where the expectation is computed over the distribution of n independent observations; it is biased otherwise. Intuitively, an estimator is unbiased if its average over infinitely many samples of n observations is equal to the true value.

It can be shown that the MLE is asymptotically unbiased: $E[\hat{\theta}_{MLE}] \xrightarrow{n\to\infty} \theta$.

Example: $\bar{X}$ *is an unbiased estimator of the mean of* ***any*** *RV*

$\bar{X}(X_1, ..., X_n) = \frac{1}{n}\sum_i X_i$. Let's use Eq. (5) to see if it's biased:

$$E[\bar{X}(X_1, ..., X_n)] = E\left[\frac{1}{n}\sum_i X_i\right] = \frac{1}{n}\sum_i E[X_i] = \frac{1}{n}(n\mu) = \mu$$

where μ is the true population mean and we used the property of the linearity of the expectation. As $E[\bar{X}] = \mu$ the estimator is unbiased.
Note that to derive this result we didn't make any assumption on the distribution of RV X.

Example: *The ML and MoM estimator of σ^2 of a Normal RV is biased*

Using the estimator found for the variance of a Normal RV, $\hat{\sigma}^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2$:

$$E[\hat{\sigma^2}(X_1, ..., X_n)] = E\left[\frac{1}{n}\sum_i X_i^2 - \bar{X}^2\right] = E\left[\frac{1}{n}\sum_i X_i^2 - \left(\frac{1}{n}\sum_i X_i\right)^2\right] =$$

$$\frac{1}{n}\sum_i E[X_i^2] - \frac{1}{n^2}\sum_i E[(\sum_i X_i)(\sum_j X_j)] =$$

$$\frac{1}{n}\sum_i E[X_i^2] - \frac{1}{n^2} E[\sum_i X_i^2 + \sum_{i\neq j} X_i X_j] =$$

$$\frac{n-1}{n^2}\sum_i E[X_i^2] - \frac{1}{n^2}\sum_{i\neq j} E(X_i)E(X_j) =$$

$$\frac{n-1}{n^2} n(\sigma^2 + \mu^2) - \frac{1}{n^2} n(n-1)\mu^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

where we used the property of the expectation of the product of two *independent* RVs, $E(X_i X_j) = E(X_i)E(X_j)$.

Hence the estimator of the variance of a RV $\hat{\sigma}^2$ is biased but asymptotically unbiased:

$$E(\hat{\sigma^2}) = \frac{n-1}{n}\sigma^2 \xrightarrow{n\to\infty} \sigma^2$$

In general, an unbiased estimator for the variance of any RV is:

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

Recap

- Maximum Likelihood
  - **Pros**:
    - asymptotically unbiased
    - easy to calculate analytically when the derivative of the PDF is easy to compute.
  - **Cons**:
    - finding the maximum can be difficult and computationally demanding.
- Method of Moments
  - **Pros**:
    - asymptotically unbiased
    - easier to calculate than MLE in some cases.
  - **Cons**:
    - sometimes for small samples gives estimates outside of parameter space
    - requires to solve integrals and systems of equations, which is not always possible.
- The two methods are not equivalent and can yield different estimates.