# Statistics Pre-course

# Part 3

# Basics of Inferential Statistics

Alfonso Russo

Department of Economics and Finance

*Tor Vergata* University of Rome

September 2024

# Part III syllabus

1. Basics of Inference
2. Confidence Intervals
3. Central Limit Theorem
4. Point Estimation
5. Statistical Models
6. Main Estimators

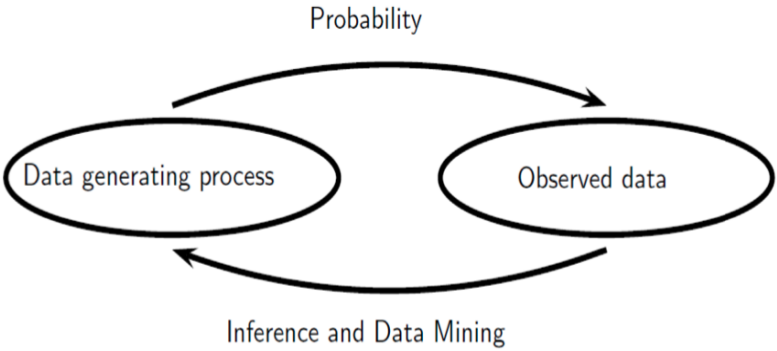
# Probability and Inference

## from population to data and back

- **Probability** starts from population, which is described by means of probability statements and functions, and predicts what happens in a sample extracted from it.

- **Inference** starts from a sample and describes the observed data in order to recover the data generating process.

# Inference in a figure



- **Estimation:** recover some *parameters* explaining the phenomena that generates the data
  - **point estimate**: a single number which is our best guess for a parameter
  - **interval estimate**: an interval of values which is believed to contain the true value of a parameter
- **Hypothesis testing:** using data to validate certain statements or predictions

- Suppose that you are interested in the (observable) random variable $X$ having support $\mathcal{X}$. We denote its probability law by $f_X^*(x)$. Assume we do not know the exact probability law of $X$ but that we know it actually belongs to a family $\mathcal{F}$ of probability laws.
- Typically, $\mathcal{F}$ is a parametric family:

$$\mathcal{F} = \{f_X(x \mid \theta),\ \theta \in \Theta\} \tag{1}$$

- Each member of the family depends on an unknown quantity $\theta$ take can take values in the *parameter space* $\Theta$
- In Inference we want to determine, through an observed sample, what member of $\mathcal{F}$ (or what value $\theta^*$ of $\theta$) identifies the probability law of $X$.
- In other words, we assume that we know the functional form of the probability law (Bernoulli, Gamma, Normal, etc) but we do not know the value of the unknown parameter(s) to characterise it.

## The elements of a Statistical Model

To characterise a random variable $X$, we need three basic elements:

- $\mathcal{X}$
- $f_X(x \mid \theta)$
- $\Theta$

A statistical model is then the triple

$$\{\mathcal{X} \ ; \ f_X(x \mid \theta) \ ; \ \Theta\} \tag{2}$$

Note that

- $X$ might be multidimensional
- $\Theta \in \mathbb{R}$; $\Theta \in \mathbb{R} \times \mathbb{R}$

Formulate the following statistical models:

- Bernoulli
- Geometric
- Normal
- Uniform

A **Random sample** is a collection of random variables $X_1, \ldots, X_n \sim f_{X_1,\ldots,X_n}$ that are

- independent

$$f_{X_1,\ldots,X_n} = \prod_{i=1}^{n} f_{X_i}(x_i) \tag{3}$$

- identically distributed

$$f_{X_i}(x_i) = f_X(x_i) \quad \forall\, i \tag{4}$$

As a consequence

$$f_{X_1,\ldots,X_n} = \prod_{i=1}^{n} f_X(x_i) \tag{5}$$

An **observed sample** $(x_1, \ldots, x_n)$ is a specific realisation of the random sample

## How populated is the Population?

Find the difference!

- **Performance analysis of Statistics students this year**: Some of your tests are collected to assess your statistics knowledge.
- **Apple New Product**: Apple has developed a new device and some units are tested to estimate average duration.
- **How rich could Temple lawyers get?**: Lawyers in Temple are surveyed to estimate how remunerative are careers in Law.
- **Humidity in desertic areas**: Deserts worldwide are sampled to study humidity levels.

$$\left\{ \mathcal{X}^{\mathbf{n}} \quad ; \quad f_n(x_n \mid \theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta) \quad ; \quad \theta \in \mathbf{\Theta} \right\} \tag{6}$$

## Example

Let $X_1, \ldots, X_n$ be i.i.d (**i**ndependent and **i**dentically **d**istributed) from a $Poisson(\lambda)$

The joint distribution takes form

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= \prod_{i=1}^{n} f_X(x_i) \\ &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \frac{1}{\prod_{i=1}^{n} x_i!}\, e^{-n\lambda}\, \lambda^{\sum_{i=1}^{n} x_i} \end{aligned} \tag{7}$$

- **Parameter:** numerical characteristics of the population that we are able to recover (typically unknown)
  - Example: $\lambda$ in a Poisson
- **Statistic:** a function of the sample that does not directly depend on any unknown parameter
  - Example: $S(X_1, \ldots, X_n) = X_{(n)} - X_{(1)}$
- **Estimator:** a statistic used to estimate a population parameter
  - Example: $T(X_1, \ldots, X_n) = \bar{X}$ is an estimator for $\mu$
- **Estimate:** the value of an estimator corresponding to an observed sample
  - Example: $T(x_1, \ldots, x_n) = \bar{x}$ is the value of $\bar{X}$ coming from $(x_1, \ldots, x_n)$

In order to assess the IQ of Tor Vergata students, we interview 10 people and use tha sample mean $\bar{X}$ as an estimator for the population mean $\mu$.

- observed sample: $x = (x_1 = 95, x_2 = 104, x_3 = 104, x_4 = 95, x_5 = 88, x_6 = 126, x_7 = 77, x_8 = 112, x_9 = 111, x_10 = 105)$
- estimate: $T(x_1, \ldots, x_10) = \bar{x} = 101.7$

## Variability of Estimators

If we collected another sample, we would obtain different results

- 2nd observed sample: $x' = (123, 119, 94, 116, 106, 91, 88, 107, 91, 103)$
- estimate: $T(x'_1, \ldots, x'_1 0) = \bar{x'} = 103.8$

Since it is a function of random objects, an estimator is a *random variable*, and the estimates are its *realisations*

## Comments on estimators

There is no "universal estimator". We must choose it according to

- the distribution of the data
  - we would not try to estimate the maximum of a discrete variable with a continuous value
- the parameter of interest
  - we would not try to estimate the mean and the variance of a Normal distribution using the same estimator

If the parameter of interest is the expected value of the population $\mathbb{E}[X]$, then the obvious candidate is the **sample mean**

$$\bar{X} = n^{-1} \sum_{i=1}^{n} X_i \tag{8}$$

**Good Properties:**

- from the *Law of Large Numbers* we know that $\bar{X} \to \mathbb{E}[X]$ as $n \to \infty$
- the *Central Limit Theorem* provides us with an approximate distribution of $\bar{X}$

## How do we define an estimator

The aim of an estimator is to recover and characterise the unknown distribution that generated the data

There are several *automatic* ways to derive an estimator, depending on how we use the data to estimate the generating distribution

- **Method of Moments:**
  - Find a distribution that has some features in common with the observed sample
- **Maximum Likelihood:**
  - find a distribution that maximises the probability of observing the sample at hand

The core idea behind the MM is to equate population and sample moments

$$\begin{cases} \mathbb{E}[X] = n^{-1} \sum_{i=1}^{n} X_i \\ \mathbb{E}[X^2] = n^{-1} \sum_{i=1}^{n} X_i^2 \\ \mathbb{E}[X^3] = n^{-1} \sum_{i=1}^{n} X_i^3 \\ \dots \\ \dots \end{cases} \tag{9}$$

The MOM gives *consistent* estimators for population parameters. Week assumptions needed. However, these will be biased!

Suppose you want to estimate $k$ unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ that characterise the distribution $f_Y(y; \boldsymbol{\theta})$ of the random variable $Y$. Suppose further that *population moments* (the moments of the "true" distribution) can be expressed as functions of the unknown parameters

$$\begin{aligned} \phi_1 &:= \mathbb{E}[Y] = h_1(\theta_1, \ldots, \theta_k) \\ \phi_2 &:= \mathbb{E}[Y^2] = h_2(\theta_1, \ldots, \theta_k) \\ &\vdots \qquad \vdots \qquad \vdots \qquad \vdots \\ \phi_k &:= \mathbb{E}[Y^k] = h_k(\theta_1, \ldots, \theta_k) \end{aligned} \tag{10}$$

Working with a sample $(y_1, \ldots, y_n)$ we could use the $j$th sample moment as an estimator $\hat{\phi}_j$ for $\phi_j$.

More formally,

$$\hat{\phi}_j = n^{-1} \sum_{i=1}^{n} y_i^j \qquad \text{for } j = 1, \ldots, k \tag{11}$$

The Method of Moments estimator of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ is the solution to the following system

$$\begin{aligned} \hat{\phi}_1 &= h_1(\hat{\theta}_1, \ldots, \hat{\theta}_k) \\ \hat{\phi}_2 &= h_2(\hat{\theta}_1, \ldots, \hat{\theta}_k) \\ \vdots \quad &\quad \vdots \quad \vdots \quad \vdots \\ \hat{\phi}_k &= h_k(\hat{\theta}_1, \ldots, \hat{\theta}_k) \end{aligned} \tag{12}$$

- Let $X_1, \ldots, X_n$ be a random sample from a population with probability law

$$f_X(x \mid \theta) = \frac{1}{\theta} x^{\frac{1-\theta}{\theta}} \quad 0 < x < 1. \tag{13}$$

  Find the MoM estimator for $\theta$.

- Consider a sample $X_1, \ldots, X_n \sim \mathcal{N}(1, \sigma^2)$. Find the MOM estimator for $\sigma^2$.

Let $X \sim Binomial(n, p)$. The probability mass function

$$\mathbb{P}(X = x) = \left( \begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \tag{14}$$

gives us the probability of observing a specific value $x$.

Assume that we know $n = 10$ and we suddenly observe $x = 8$.

- **With $p = 0.5$:** $\mathbb{P}(X = 8) = \left( \begin{array}{c} 10 \\ 8 \end{array} \right) (0.5)^8 (0.5)^2 = 0.043$
- **With $p = 0.7$:** $\mathbb{P}(X = 8) = \left( \begin{array}{c} 10 \\ 8 \end{array} \right) (0.7)^8 (0.3)^2 = 0.233$

For $x = 8$ a value of the parameter $p = 0.7$ seems more likely than $p = 0.5$.

When we fix a realisation $x$ and we consider it to be a function of the unknown parameter $p$, the p.m.f. $\binom{n}{x} p^x (1-p)^{n-x}$ gives us a measure of how compatible is $x$ is $p$.

This is called the **Likelihood** of $p$.

**NB** The Likelihood tells us how **plausible** a value of the parameter is, but it does not measure its **probability**

- Given a statistical model $\{\mathcal{X}^n \ ; \ f_n(\boldsymbol{x}_n \mid \theta) \ ; \ \theta \in \Theta\}$, the function $\mathcal{L} : \Theta \to \mathbb{R}^+$ defined as

$$\mathcal{L}(\theta \mid \boldsymbol{x}_n) \propto f_n(\boldsymbol{x}_n \mid \theta) \quad , \quad \theta \in \Theta \tag{15}$$

  is called the likelihood function associated to the observed sample $\boldsymbol{x}_n$.

- Since $\boldsymbol{x}_n = (x_1, \ldots, x_n) \in \mathbb{R}$, $\mathcal{L}(\cdot)$ is a function of $\theta$.
- Values of $\theta$ for which $\mathcal{L}(\theta \mid \boldsymbol{x}_n)$ is higher, are the most "compatible" with the observed sample.

The **Maximum Likelihood Estimator (MLE)** for a parameter $\theta$, is the solution to the following maximisation problem

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max_{\theta} \mathcal{L}(\theta \mid x_1, \ldots, x_n) \\ &= \log\Big(\mathcal{L}(\theta \mid x_1, \ldots, x_n)\Big) \\ &= \ell(\theta \mid x_1, \ldots, x_n)\end{aligned} \tag{16}$$

Operationally, the steps to find the **MLE** are:

1. Compute the derivative of the log-likelihood and equate it to 0

$$\frac{d\ell(\theta \mid x_1, \ldots, x_n)}{d\theta} = 0 \tag{17}$$

2. Isolate $\theta$ to find a candidate for the MLE

3. Check second derivative

- Consider a Bernoulli model for a generic sample of dimension $n$.

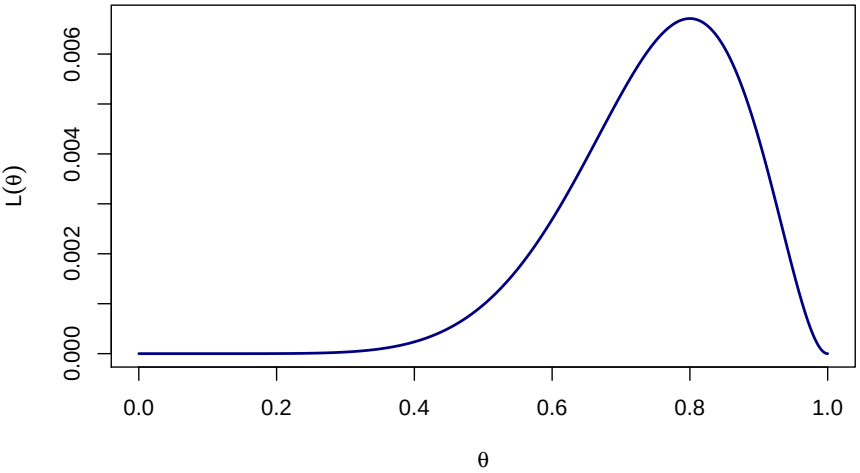
$$\mathcal{L}(\theta \mid \boldsymbol{x}_n) = f_n(\boldsymbol{x}_n \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i}\,(1-\theta)^{n-\sum_{i=1}^{n} x_i} \tag{18}$$

- I run the experiment $10$ times and we observe $\sum_{i=1}^{n} = 8$.
- The likelihood function is therefore

$$\mathcal{L}(\theta \mid \boldsymbol{x}_n) = \theta^8(1-\theta)^2 \quad , \quad \theta \in [0,1] \tag{19}$$

## Bernoulli maximum likelihood

- From visual inspection of the likelihood function



## Example

Let $X_1, \ldots, X_n$ be a random sample with $X_i \sim Poisson(\lambda)$. Then:

- likelihood function:

$$\mathcal{L}(\lambda \mid x_1, \ldots, x_n) = \frac{1}{\prod_{i=1}^{n} x_i!} e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i} \tag{20}$$

- log-likelihood:

$$\ell(\lambda \mid x_1, \ldots, x_n) = \log\left(\frac{1}{\prod_{i=1}^{n} x_i!}\right) - n\lambda + \log(\lambda) \sum_{i=1}^{n} x_i \tag{21}$$

## Example

- Compute the derivative of $\ell(\lambda \mid x_1, \ldots, x_n)$ and equate it to 0

$$\frac{d\ell(\lambda \mid x_1, \ldots, x_n)}{d\lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i = 0 \tag{22}$$

- Isolate $\lambda$ to get the MLE estimate

$$-n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i = 0 \implies \hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n \tag{23}$$

**NB** Even if $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n \mid \lambda)$ denotes a discrete distribution, the Likelihood is still a continuous function in $\lambda$, hence it can be differentiated and maximised.

## Exercises

- Consider a sample $X_1, \ldots, X_n$ of discrete random variables where $X_i \sim Geometric(\theta)$ with probability mass function given by

$$f_X(x|\theta) = \theta(1-\theta)^{x-1} \quad \forall\, x \in \mathbb{N}_+ \text{ and } 0 < \theta < 1 \tag{24}$$

  Find the MLE for $\theta$.

- Let $X_1, \ldots, X_n$ be a random sample with pdf

$$f(x \mid x_0, \theta) = \theta x_0^{\theta} x^{-\theta-1} \tag{25}$$

  Suppose $x_0 > 0$ is known and given. Find the MLE for $\theta$.

## Evaluating point estimators

Consider an estimator $T$ for an unknown population parameter $\theta$.

The estimator $T$ is said to be **unbiased** if

$$\mathbb{E}[T] = \theta \tag{26}$$

The estimator $T$ is **precise** if its variance $\mathbb{V}[T]$ is *small*

A "good" estimator is, on average, close to the real value of the parameter it is trying to recover and is always "on target".
The **Mean Squared Error** (MSE) evaluates the performance of an estimator $T$ combining the concepts of Bias and Variance

$$MSE(T) = \mathbb{V}[T] + [Bias(T)]^2 \tag{27}$$

If $\mathbb{E}[T] = \theta$ the estimator $T$ is unbiased and its MSE reduces to its variance.

**Consistency**

- the MSE can be alternatively defined as

$$MSE(T) = \mathbb{E}[(T - \theta)^2] \tag{28}$$

- when we get that

$$\lim_{n \to \infty} MSE[T] = 0 \tag{29}$$

the estimator $T$ becomes closer and closer to the true parameter value $\theta$ as $n$ grows. This important property is called **consistency** and it reassures us that increasing the sample size will improve the performance of our estimator