Statistical Inference

Tullia Padellini

tulliapadellini.github.io 💻
tullia.padellini@uniroma1.it ✉

Inference vs Probability

from population to data and back

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

- **Inference** starts from a sample and describes the observed data with the aim of inferring relevant information on the population.

What is Inference?

a general introduction

- **Estimate**: recover some parameter explaining the phenomenon that generates the data

    point estimate: a *single number* that is our best guess for the parameter.

    interval estimate: an *interval of numbers* that is believed to contain the actual value of the parameter.

- **Hypothesis testing**: using data to validate certain statements or predictions

Random sample

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

- *independent*

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

- *identically distributed*

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

As a consequence

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

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

Toy Example

how to compute the sample distribution

Let $X_1, \dots, X_n$ i.i.d. (**i**ndependente **i**dentically **d**istributed) from a Poisson(λ).

The **sampling distribution** $f_{X_1,\dots,X_n}$ can be derived as follows:

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

Basic Concepts

short glossary of estimation tools

- **Parameter**: numerical characteristic of the population that we are trying to recover (hence typically unknown)

  Examples: λ in a Poisson

- **Statistics**: numerical 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 the population parameter

  Example: $T(X_1, \ldots, X_n) = \bar{X}$ is an estimator for μ

- **Estimate**: the value of an estimator corresponding to an *observed* sample:

  Example: $T(x_1, \ldots, x_n) = \bar{x}$ is an estimate corresponding to $\bar{X}$

Variability of Estimators

walking our way through it with an example

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

- 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: $T(x_1, \dots, x_n) = \bar{x} = 101.7$

CAVEAT: if we draw another sample from the same population, we will observe different results:

- 2-nd observed sample: $x' = (123, 119, 94, 116, 106, 91, 88, 107, 91, 103)$
- estimate: $T(x_1, \dots, x_n) = \bar{x'} = 103.8$

Since it is a function of a random object, an **estimator** is a *random variable*, and the **estimates** are its *realizations*.

Comments on estimators

not all estimators are good

There is no “universal estimator”, but it must be chosen according to:

- the distribution of the data

  we wouldn’t try to estimate the max of a discrete variable with a continuous value

- the parameter of interest

  we wouldn’t try to estimate the mean and the variance of a Normal distribution with the same estimator

Example:

- parameter of interest: mean of a Normal population
- estimator: $T(X_1, \dots, X_n) = X_{(n)}$

An easy case

one estimator you often find in practice

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

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Good Properties:

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

The sample mean

there is more to be happy about

$$\mathbb{E}[\bar{X}] = \mathbb{E}[X]$$

- *on average* it gives us the right value:

$$\mathbb{V}[\bar{X}] = \frac{\mathbb{V}[X]}{n}$$

- as n grows, we are increasingly confindent in our estimate

How do we define an estimator

when we have unconventional densities

The aim of the estimator is to try to recover the distribution that generated the data.

The are several *automatic* ways to derive an estimator, depending on how to use the data to recover the generating distribution.

- **Methods of Moments**:
  find a distribution that has some features of the observed sample
- **Maximum Likelihood**:
  find a distribution that maximises the probability of observing the sample at hand

Methods of Moments

for point estimation

The core idea is to *equate sample moments to population moments*, i.e.

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

Example:

Consider a random sample $X_1, \dots, X_n \sim$ Unif$(0, \theta)$, for which $\mathbb{E}[X] = \theta/2$.

The MOM estimator is found by equating $\mathbb{E}[X] = \theta/2$ with $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$:

$$\theta/2 = \bar{X} \qquad \Rightarrow \hat{\theta}_{MOM} = 2\bar{X}$$

Exercise

for you to spend your free time

Let $X_1, \dots, X_n \sim \text{Unif}(a, b)$, compute the MOM estimator for a and b.

Remember that

$$X \sim \text{Unif}(a, b) \qquad \Rightarrow \mathbb{E}[X] = \frac{b+a}{2} \qquad \mathbb{V}[X] = \frac{(b-a)^2}{12}$$

The Likelihood Function

the basic intuition

Let $X \sim$ Binomial(n, p), the probability mass function $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, gives us the probability of observing a value x.

Now assume that we know $n = 10$ and we observe $x = 8$

- if $p = 0.5$, $P(X = 8) = \binom{10}{8}(0.5)^8(0.5)^2 = 0.043$
- if $p = 0.7$, $P(X = 8) = \binom{10}{8}(0.7)^8(0.3)^2 = 0.233$

For $x = 8$, the parameter $p = 0.7$ seems to be more likely than $p = 0.5$.

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

Maximum Likelihood Estimator

The **Maximum Likelihood Estimator (MLE)** is the value of the parameter that maximises the Likelihood:

$$\hat{\theta}_{MLE} = argmax L(\theta; x_1, \dots, x_n) = argmax l(\theta; x_1, \dots, x_n)$$

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

1. **Compute the derivative** of the log-likelihood and equate it to 0: $dl(\theta; x_1 \dots, x_n)/d\theta = 0$
2. **Isolate** θ to find the candidate for the **MLE** (i.e. the critical point)
3. **Check the sign** of $d^2 l(\theta; x_1 \dots, x_n)/d\theta^2$ in the candidate θ to verify that this is not a min or a saddle

Example

Maximum Likelihood for the parameter λ of a Poisson:

Remember that if $X_1, \dots, X_n$ random sample, with $X_i \sim$ Poisson(λ) then:

- joint distribution

$$p_{X_1,\dots,X_n}(x_1, \dots, x_n; \lambda) = \frac{1}{\prod_{i=1}^n x_i!} e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}$$

- Likelihood

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

- log-Likelihood

$$l(\lambda; x_1, \dots, x_n) = log\left(\frac{1}{\prod_{i=1}^n x_i!}\right) - n\lambda + \sum_{i=1}^n x_i log(\lambda)$$

Example

Maximum Likelihood for the parameter λ of a Poisson:

1. Compute the derivative of $l(\lambda; x_1, \ldots, x_n)$ and equate it to 0:

$$\frac{dl(\lambda; x_1, \ldots, x_n)}{d\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

2. Isolate λ to get the MLE estimate:

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

CAVEAT Even if $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \lambda)$ denotes a discrete distribution, it is **a continuous function in** λ, hence we can compute derivatives to find the max.

Core of the Likelihood

The multiplicative factor **depending on the data** but **not on the parameter** $\frac{1}{\prod_{i=1}^{n} x_i!}$ disappeared when we computed the derivative. This is always true:

- if $L(\lambda; x) = h(x)g(x, \theta)$, then $l(\lambda; x) = log(h(x)) + log(g(x, \theta))$
- the derivative of $log(h(x))$ does not depend on θ

$$\frac{dl(\theta; x)}{d\theta} = \frac{dlog(h(x))}{d\theta} + \frac{dlog(g(x, \theta))}{d\theta} = \frac{dlog(g(x, \theta))}{d\theta}$$

The function $g(x, \theta)$ is called the **core** of the likelihood and it contains all the information we need from the data.

Since **we can replace L with g without loss of information**, when we talk about *Likelihood* we actually talk about its *core*.

Exercise

Let $X_1, \dots, X_n$ be a random sample (i.i.d.), where each X_i has the following density function

$$f_X(x;\theta) = (\theta + 1)x^{\theta} \qquad\qquad x \in (0,1),\ \theta > -1$$

- Compute the joint distribution $f_{X_1,\dots,X_n}(x_1, \dots, x_n)$
- Find the likelihood distribution
- Determine the Maximum Likelihood estimator for θ

Evaluating Point estimators

- An estimator T for a parameter θ, is said to be **unbiased** if $\mathbb{E}[T] = \theta$.
  a “good” estimator is on average close to the real value of the parameter of interest

- An estimator T is **precise** if its variance $\mathbb{V}(T)$ is small.
  a “good” estimator is *always* on target

The **Mean Squared Error** (MSE) evaluates the performance of the estimator combining these two desiderata:

$$MSE(T) = \mathbb{V}(T) + \text{Bias}(T)^2$$

MSE

- if $\mathbb{E}[T] = \theta$ we say that the estimator is **unbiased** and the MSE reduces to its variance

Consistency

- the MSE can be alternatively defined as

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

- when

$$lim_{n \to \infty} MSE(T) = 0$$

  we have that as n grows T becomes closer and closer to real value of the parameter θ. This important property is called **consistency**, and reassures us that adding more observations improves the performances of the estimator

Exercise

comparing MSE

Let $X_1, ..., X_n$ be a random sample from a Normal distribution N(μ, μ^2). Consider the following estimators for the parameter μ:

$$T_1(X_1, ..., X_n) = \frac{X_1 + X_2 + ... + X_{n-1}}{n-1} - \frac{X_n}{n}$$

$$T_2(X_1, ..., X_n) = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Determine the bias of the two estimators.
- Determine the Mean Square Error of the two estimators
- Which of the two estimators is more efficient?

Interval Estimates

A **interval estimator** for a parameter θ is a random interval $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$, containing the most believable values for the parameter.

Intuitively, it is very difficult to predict the **exact** value of the unknown parameter (if T is a continuous random variable, this is even impossible, as by definition $P(T = \theta) = 0$), hence is more reasonable to ask for a range of possible parameters.

In addition a set of plausible values is more informative on the phenomenon than just a single guess.

The ingredients

A **confidence interval of level** $1-\alpha$ is a random interval $[L, U]$, where L and U are two *statistics*, such that

$$P(\theta \in [L, U]) = 1-\alpha$$

The **confidence level** ($1-\alpha$) is probability that the interval contains the true value of the parameter θ, *before the sample is observed*. Typically this value is chosen to be high (0.95 or 0.99).

Typically a confidence interval is built using the formula

$$T \pm err$$

where T is the point estimator for θ and err measures how accurate the point estimate is and depends on the level of confidence as well as $\mathbb{V}[T]$.

Confidence

a word of caution

BE CAREFUL: once we observe the sample, and we have an *estimate* of the confidence interval $[l, u]$, the probability that the parameter lies in this interval is either 0 or 1.

However, remembering the definition of probability as the limit of the relative frequency of an event, we can be **confident** that if we build a large number of confidence intervals, the parameter will be contained in the 95% of them.

Building a Confidence Interval

a practical example

Let $X_1, \dots, X_n$ be an iid random sample from a Norm(μ, σ^2) where σ^2 is known. Let us build a confidence interval of level $1 - \alpha$

- Take $\bar{X}$ as a point estimator of the parameter of interest
- Remember that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Norm}(0, 1)$$

so that

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Rearranging the terms we have the confidence interval $\bar{X} \pm \sigma/\sqrt{n} z_{\alpha/2}$

Toy example

a very boring one

When a shipment of coal is traded, a number of its properties should be known accurately, because the value of the shipment is determined by them. An important example is the so-called gross calorific value, which characterizes the heat content and is a numerical value in megajoules per kilogram (MJ/kg).

As there is uncertainty related to the measurement procedure, the measurement are random, and known to be normal, with a standard deviation of about 0.1 MJ/kg. For a shipment of coal, 23 measurements are given with sample mean xbar = 23.788

? compute the confifence level for μ at a confidence level $1 - \alpha = 0.95$

Toy example

a very boring one

When a shipment of coal is traded, a number of its properties should be known accurately, because the value of the shipment is determined by them. An important example is the so-called gross calorific value, which characterizes the heat content and is a numerical value in megajoules per kilogram (MJ/kg).

As there is uncertainty related to the measurement procedure, the measurement are random, and known to be normal, with a standard deviation of about 0.1 MJ/kg. For a shipment of coal, 23 measurements are given with sample mean xbar = 23.788

? compute the confifence level for μ at a confidence level $1 - \alpha = 0.95$

Hypothesis Testing

The main goal of **statistical testing** is to check whether the data support certain statements (**hypothesis**), usually expressed in terms of population parameters for variables measured in the study.

Usually, an *hypothesis* on the parameter θ is formalized as follows:

- $\theta = \theta_0$ *punctual* hypothesis
- $\theta \geq \theta_0$ or $\theta \leq \theta_0$ *one-sided* hypothesis
- $\theta \neq \theta_0$ *two-sided* hypothesis

Hypothesis

In a **hypothesis test** we compare two alternative hypothesis H_0 and H_1:

- The **Null Hypothesis** (H_0) is the hypothesis that is held to be true unless sufficient evidence to the contrary is obtained.
- The **Alternative Hypothesis** (H_1) represent the new theory we would like to test.

Example: We want to test whether an astrologer can correctly predict which of 3 personalities charts applies to a person.

- $H_0 : p = 1/3$
  the astrologer doesn't have any predictive power (the probability of guessing the personality is $1/3$)
- $H_1 : p \geq 1/3$
  the astrologer does have predictive power

Test logic

Innocent until proven guilty

	H_0 is true	H_0 is false
Accept H_0	✌	Type II Error
Reject H_0	Type I Error	✌

- If we want to completely avoid Type II Error we should **always Reject** H_0
- If we want to completely avoid Type I Error we should **always Accept** H_0

It is impossible to simultaneously avoid both: which one is more important?

As H_0 represent the current condition, we would like to subvert it only when the data provide strong evidence against it

Testing procedure:

How to solve a test $H_0 = \theta \leq \theta_0$ versus $H_1 = \theta > \theta_0$:

1. Choose a level α of significance (i.e. the probability of Type I Error), typically $\alpha = 0.05$
2. Choose a test statistic T, i.e. a statistic that describes how far that point estimate falls from the parameter value given in the null hypothesis
3. Given an observed sample $(x_1, \dots, x_n)$, compute the $t = T(x_1, \dots, x_n)$
4. Compute the p-value, $P(T > t | H_0) = p$, a measure of how compatibles the data are with H_0
5. If $p \leq \alpha$, reject H_0, otherwise do not reject it