STATISTICS

Module 2 *of*
MATHEMATICS AND STATISTICS

Marina Santacroce

marina.santacroce@unibocconi.it
Department of Decision Sciences, Bocconi University

L15: Point parameter estimation maximum likelihood method.

References: S.Ross *Introduction to* Probability and Statistics for Engineers and Scientists: 7.1, 7.2, 7.7

Outline

Estimation of a population parameters

One of the main goals of inferential statistics is the estimation of the population parameters by observing the data of a random sample.

- $X_1, \ldots X_n$, are sampled from a population with distribution F_θ
- F_θ is specified up to a vector of *unknown parameters* θ

  (Poisson with unknown λ, normal with unknown μ and σ^2, ...)

$\Rightarrow$ use the sample to make inference about the unknown parameters.

- Point estimation method (this week)
  - maximum likelihood method
  - quality of a point estimation method evaluated by considering the mean square *error*
- Interval estimation method (after Easter break)
  - estimation through *interval* and related *confidence* level.

Point estimation

Def.: Any *statistic* $d = d(X_1, \ldots X_n)$ [1] *used to estimate* the value of an unknown *parameter* θ is called an (point) *estimator* of θ.
The *observed value of an estimator* is called the *estimate*.

- a popular estimator of the mean of a population μ is the sample mean $\overline{X}$
  if $X_1 = 1, X_2 = 2, X_3 = 3, X_4 = 2, X_5 = 2$ is the observed sample,
- the estimate of μ is the value of the estimator $\overline{X}$ computed on the sample $\Rightarrow$ $\overline{X} = 2$

Given a sample from a population with distribution specified by the p.d.f (or p.m.f. if discrete) denoted by f and depending on some unknown θ the joint p.d.f. (p.m.f.) is

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n|\theta) = \prod_{i:1}^{n} f_{X_i}(x_i|\theta) = f(x_1, \ldots, x_n|\theta) \quad x_i \in \mathbb{R}$$

In the case of a population with an exponential distribution

$$f(x_1, \ldots, x_n|\theta) = \prod_{i:1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i:1}^{n} x_i}, \quad x_i \in (0, \infty)$$

The objective would be to estimate θ from the observed data $X_1, \ldots X_n$.

[1] any random variable completely specified by the value of the random sample

Likelihood function

- Given a sample $X_1, \ldots, X_n$ and having observed $x_1, \ldots, x_n$
- the joint pdf (or pmf) associated to the sample, $f(x_1, \ldots, x_n|\theta)$, seen as a function of θ for fixed $x_1, \ldots, x_n$ is called likelihood function

$$L(\theta; x_1, \ldots, x_n) = L(\theta) = f(x_1, \ldots, x_n|\theta) \qquad \theta \in \Theta$$

- Considering $L(\theta; x_1, \ldots, x_n)$ is a change of perspective

$\Rightarrow$ $x_1, \ldots, x_n$ is fixed and $f(x_1, \ldots, x_n|\theta)$ is seen as a function of θ

- Note that $L(\theta; x_1, \ldots, x_n)$ is not necessarily a pdf or pmf

  $\int L(\theta; x_1, \ldots, x_n) d\theta \neq 1$

To give an intuition of the meaning of $L(\theta; x_1, \ldots, x_n)$, let us consider the discrete case:

$$L(\theta; x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n|\theta)$$

is the probability of observing $x_1, \ldots, x_n$ when θ is the true unknown parameter.

Keeping in mind that

$$L(\theta; x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | \theta)$$

is the probability of observing $x_1, \ldots, x_n$ when θ is the true unknown parameter, the following inequality

$$\frac{L(\theta_1; x_1, \ldots, x_n)}{L(\theta_2; x_1, \ldots, x_n)} > 1,$$

means that

- the probability of observing the sample outcome $x_1, \ldots, x_n$ is larger if $\theta = \theta_1$ rather than if $\theta = \theta_2$

$\Rightarrow$ θ_1 *is more likely than* θ_2 (for a given $x_1, \ldots, x_n$)

Example: We tossed a coin with unknown $p = \mathbb{P}(\text{head})$ 5 times and we observed $(1, 1, 1, 1, 0)$, where $X_i = 1$ if the i-th toss is head and zero otherwise.

$$L(p; 1, 1, 1, 1, 0) = \mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 0; p) = p^4(1 - p)$$

Comparing the likelihoods for $p = \frac{1}{5}$ and $p = \frac{4}{5}$

$$\frac{L(p = \frac{4}{5}; 1, 1, 1, 1, 0)}{L(p = \frac{1}{5}; 1, 1, 1, 1, 0)} = \frac{4^4/5^5}{4/5^5} = 4^3 > 1$$

given we observed in the first four tosses head and in the last tail,

$$p = \tfrac{4}{5} \text{ is } (4^3 \text{ times}) \text{ more likely than } p = \tfrac{1}{5}.$$

Maximum likelihood estimators

The idea, underlying the maximum likelihood method, is to estimate θ with the value $\hat{\theta}$ which makes the sample outcome $x_1, x_2, \ldots, x_n$ most likely. Recall, we have written

- the joint p.d.f. (or p.m.f.) of the sample, given the sample outcome is $x_1, \ldots, x_n$, as a function of θ

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta) = L(\theta; x_1, \ldots, x_n).$$

$\Rightarrow$ the maximum likelihood estimate is the value $\hat{\theta}$ of the parameter θ which maximizes the likelihood of the observed values $x_1, \ldots, x_n$

$$L(\hat{\theta}; x_1, \ldots, x_n) = \max_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n)$$

- It is often useful to maximize $\log[L(\theta; x_1, x_2, \ldots, x_n)] = l(\theta; x_1, x_2, \ldots, x_n)$ rather than $L(\theta; x_1, x_2, \ldots, x_n)$.

  It is equivalent, since log is increasing monotone, thus the two functions have the maximum at the same value $\hat{\theta}$.

MLE of an exponential

1. MLE of an exponential parameter θ ($\theta > 0$).

$$L(\theta; x_1, x_2, \ldots, x_n) = \theta^n e^{-\theta \sum_{i:1}^{n} x_i},$$

thus, considering the $\log L$, we have

$$l(\theta; x_1, x_2, \ldots, x_n) = n \log(\theta) - \theta \sum_{i:1}^{n} x_i$$

$$\hat{\theta} = \frac{n}{\sum_{i:1}^{n} x_i} = \frac{1}{\overline{x}}$$

- The maximum likelihood estimator $d(X_1, \ldots, X_n)$ for the parameter λ of an exponential population is

$$\Rightarrow \qquad d(X_1, \ldots, X_n) = \frac{1}{\overline{X}}$$

MLE of a Bernoulli

2. MLE of a Bernoulli parameter p.

Let us consider a sample $X_1, \ldots, X_n$ from a Bernoulli population with unknown parameter $p \in (0, 1)$.

- $X_i \sim \text{Be}(p)$, the m.p.f. is $f(x_i|p) = (1-p)^{1-x_i} p^{x_i}, \quad x_i = 0, 1.$
- The related likelihood function is

$$L(p; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|p) = \prod_{i:1}^{n} f(x_i|p) = (1-p)^{n-\sum_{i:1}^{n} x_i} p^{\sum_{i:1}^{n} x_i}$$

passing to the $\log L$ we find

$$l(p; x_1, x_2, \ldots, x_n) = (n - \sum_{i:1}^{n} x_i) \log(1-p) + \sum_{i:1}^{n} x_i \log(p), \quad \text{if } p \neq \{0, 1\}.$$

We find the ML estimate as the value $\hat{p}$ which maximizes $l(p; x_1, \ldots, x_n)$:

$$\hat{p} = \frac{\sum_{i:1}^{n} x_i}{n}$$

- The maximum likelihood estimator of the unknown parameter p of a Bernoulli distribution is given by

$$d(X_1, \ldots, X_n) = \overline{X}.$$

MLE for the Poisson

3. MLE of a Poisson parameter λ.

Let $X_1, \ldots, X_n$ be a sample from a Poisson population with unknown parameter $\lambda > 0$.

- $X_i \sim$ Poisson(λ), the m.p.f. is $f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \quad x_i = 0, 1, \ldots$
- The related likelihood function is

$$L(\lambda; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\lambda) = \prod_{i:1}^{n} f(x_i|\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{i:1}^{n} x_i}}{\prod_{i:1}^{n} x_i!}$$

passing to the log L we find

$$l(\lambda; x_1, x_2 \ldots, x_n) = -n\lambda + \sum_{i:1}^{n} x_i \log(\lambda) - \sum_{i:1}^{n} \log(x_i!).$$

We find the ML estimate as the value $\hat{\lambda}$ which maximizes $l(\lambda; x_1, \ldots, x_n)$:

$$\hat{\lambda} = \frac{\sum_{i:1}^{n} x_i}{n}$$

- The maximum likelihood estimator of the unknown parameter λ of a Poisson distribution is $d(X_1, \ldots, X_n) = \overline{X}$.

MLE for the normal

4. MLE of a normal with parameters μ and σ^2.

Let $X_1, \ldots, X_n$ be a sample from a normal population with unknown parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

- $X_i \sim \mathrm{N}(\mu, \sigma^2)$, the m.p.f. is $f(x_i|\mu, \sigma^2) = \frac{e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}, \quad x_i \in \mathbb{R}$
- The related likelihood function is

$$L(\mu, \sigma^2; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\mu, \sigma^2) = \prod_{i:1}^{n} f(x_i|\mu, \sigma^2) = \frac{e^{-\sum_{i:1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}}{(2\pi\sigma^2)^{\frac{n}{2}}},$$

passing to the $\log L$ we find

$$l(\mu, \sigma^2; x_1, x_2 \ldots, x_n) = -\sum_{i:1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(2\pi).$$

We find the ML estimates as the values $\hat{\mu}$ and $\hat{\sigma}^2$ which maximizes $l(\mu, \sigma^2; x_1, \ldots, x_n)$:

$$\hat{\mu} = \frac{\sum_{i:1}^{n} x_i}{n} = \overline{x} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i:1}^{n} (x_i - \overline{x})^2}{n}.$$

- The maximum likelihood estimators of the unknown parameters of a normal are $\overline{X}$ and $\frac{\sum_{i:1}^{n} (X_i - \overline{X})^2}{n}$.

MLE for the uniform

5. MLE of a uniform on $(0, \theta)$.

Let $X_1, \ldots, X_n$ be a sample from a uniform on $(0, \theta)$ with unknown parameter $\theta > 0$.

- $X_i \sim$ U[(0,θ)], the m.p.f. is $f(x_i|\theta) = \frac{1}{\theta}, \quad x_i \in (0, \theta)$.
- The related likelihood function is

$$L(\theta; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i:1}^{n} f(x_i|\theta) = \frac{1}{\theta^n};\ x_i \in (0, \theta),\ \forall i.$$

Note that to maximize $L(\theta; x_1, x_2, \ldots, x_n)$ we should choose $\hat{\theta}$ as small as possible!

We find the ML estimate as the value $\hat{\theta}$ which maximizes $L(\theta; x_1, \ldots, x_n)$:

$$\hat{\theta} = \max(x_1, \ldots, x_n)$$

- The maximum likelihood estimator of the unknown parameter θ of a uniform distribution on $(0, \theta)$ is

$$d(X_1, \ldots, X_n) = \max(X_1, \ldots, X_n).$$

How good is an estimator?

Let $X_1, \dots X_n$ be a sample from a population whose distribution is known up to an unknown parameter θ, and let $d = d(X_1, \dots X_n)$ be an estimator of θ.

- How to understand if d is a good estimator of θ?

$\Rightarrow$ the idea is to consider the estimation *error* i.e. $d(X_1, \dots, X_n) - \theta$ and evaluating the quality of the estimator from the *mean square error*

$$r(d, \theta) = \mathbb{E}\left(d(X_1, \dots, X_n) - \theta\right)^2.$$

Unfortunately in general it does not exist an estimator d which minimizes $r(d, \theta)$ for all the possible values of θ.

Example: the constant estimator $d^* = 4$ has $r(d^*, \theta) = 0$ if $\theta = 4$

$$\Rightarrow \qquad r(d, \theta) \geq r(d^*, \theta) = 0 \quad \text{if } \theta = 4.$$

Since

$$r(d, \theta) = \mathrm{Var}(d(X_1, \dots, X_n)) + \left(\underbrace{\mathbb{E}(d(X_1, \dots, X_n)) - \theta}_{b_\theta(d)} \right)^2$$

where $b_\theta(d) = \mathbb{E}(d(X_1, \dots, X_n)) - \theta$ is the *bias* of d as an estimator of θ.

Unbiased estimators

We define an unbiased estimator of θ (or correct) as an estimator $d = d(X_1, \ldots X_n)$ such that

$$b_\theta(d) = \mathbb{E}(d(X_1, \ldots, X_n)) - \theta = 0, \quad \text{for all } \theta.$$

If d is an unbiased estimator of θ then its mean square error is

$$r(d, \theta) = \mathbb{E}\left(d(X_1, \ldots, X_n) - \theta\right)^2 = \mathrm{Var}(d(X_1, \ldots, X_n)).$$

Thus we can hope to find an unbiased estimator with uniformly minimum variance (among the unbiased estimators).

Example: Let $X_1, \ldots X_n$ be a sample from a population having unknown mean μ and known variance σ^2, and consider the two estimators of μ:

- $d_1 = d_1(X_1, \ldots X_n) = X_1$ and $d_2 = d_2(X_1, \ldots X_n) = \overline{X}$
- d_1 and d_2 are unbiased estimators and $\mathrm{Var}(d_1) = \sigma^2 \geq \mathrm{Var}(d_2) = \frac{\sigma^2}{n}$.

$\Rightarrow$ Thus $\overline{X}$ is preferable to X_1 as an estimator of the mean μ of the population.

Example: uniform

Let $X_1, \ldots, X_n$ be a sample from a uniform on $(0, \theta)$ with unknown parameter $\theta > 0$.

- $X_i \sim \mathrm{U}[(0,\theta)]$ and $\mathbb{E}(X_i) = \frac{\theta}{2}$.

$d_1 = d_1(X_1, \ldots, X_n) = 2\overline{X}$ would be a natural choice of estimator for θ.

$$\mathbb{E}\left(d_1(X_1, \ldots, X_n)\right) = \mathbb{E}\left(2\overline{X}\right) = \theta \quad \text{and} \quad r(d_1, \theta) = \mathrm{Var}(2\overline{X}) = 4\frac{\theta^2}{12n} = \frac{\theta^2}{3n}$$

- d_1 is unbiased and its mean square error is $r(d_1, \theta) = \theta^2/3n$.

The MLE $d_2(X_1, \ldots, X_n) = \max_i X_i$ represents another possible choice.

$$\mathbb{E}(d_2(X_1, \ldots, X_n)) = \mathbb{E}(\max_i X_i) = \frac{n}{n+1}\theta$$

$$b_\theta(d_2) = \frac{n}{n+1}\theta - \theta = -\frac{\theta}{n+1}$$

- MLE is biased although it is *asymptotically unbiased* (correct)!

$$r(d_2, \theta) = \mathrm{Var}\left(\max_i X_i\right) + (b_\theta(d_2))^2 = \frac{2\theta^2}{(n+1)(n+2)} \leq \frac{\theta^2}{3n} = r(d_1, \theta)$$

$\Rightarrow$ Comparing the two estimators of the parameter θ of a uniform on $(0, \theta)$ we conclude that the MLE is preferable to d_1.