

Point Estimation: definition of estimators



Point estimator: any function $W(X_1, \dots, X_n)$ of a data sample.

The exercise of point estimation is to use particular functions of the data in order to estimate certain unknown population parameters.

Examples: Assume that X_1, \dots, X_n are drawn i.i.d. from some distribution with unknown mean μ and unknown variance σ^2 .

Potential point estimators for μ include: sample mean $\bar{X}_n = \frac{1}{n} \sum_i X_i$; sample median $\text{med}(X_1, \dots, X_n)$.

Potential point estimators for σ^2 include: the sample variance $\frac{1}{n} \sum_i (X_i - \bar{X}_n)^2$.

Any point estimator is a random variable, whose distribution is that induced by the distribution of X_1, \dots, X_n .

Example: $X_1, \dots, X_n \sim \text{i.i.d. } N(\mu, \sigma^2)$.

Then sample mean $\bar{X}_n \sim N(\mu_n, \sigma_n^2)$ where $\mu_n = \mu, \forall n$ and $\sigma_n^2 = \sigma^2/n$.

For a particular realization of the random variables x_1, \dots, x_n , the corresponding point estimator evaluated at x_1, \dots, x_n , i.e., $W(x_1, \dots, x_n)$, is called the *point estimate*.



In these lecture notes, we will consider three types of estimators:

1. Method of moments
2. Maximum likelihood
3. Bayesian estimation



Method of moments: very intuitive idea

Assume: $X_1, \dots, X_n \sim \text{i.i.d. } f(x|\theta_1, \dots, \theta_K)$

Here the unknown parameters are $\theta_1, \dots, \theta_K$ ($K \leq N$).

Idea is to find values of the parameters such that the population moments are as close as possible to their “sample analogs”. This involves finding values of the parameters to solve the following K -system of equations:

$$\begin{aligned}
m_1 &\equiv \frac{1}{n} \sum_i X_i = EX = \int x f(x|\theta_1, \dots, \theta_K) \\
m_2 &\equiv \frac{1}{n} \sum_i X_i^2 = EX^2 = \int x^2 f(x|\theta_1, \dots, \theta_K) \\
&\vdots \\
m_K &\equiv \frac{1}{n} \sum_i X_i^K = EX^K = \int x^K f(x|\theta_1, \dots, \theta_K).
\end{aligned}$$

Example: $X_1, \dots, X_n \sim$ i.i.d. $N(\theta, \sigma^2)$. Parameters are θ, σ^2 .

Moment equations are:

$$\begin{aligned}
\frac{1}{n} \sum_i X_i &= EX = \theta \\
\frac{1}{n} \sum_i X_i^2 &= EX^2 = VX + (EX)^2 = \sigma^2 + \theta^2.
\end{aligned}$$

Hence, the MOM estimators are $\theta^{MOM} = \bar{X}_n$ and $\sigma^{2MOM} = \frac{1}{n} \sum_i X_i^2 - (\bar{X}_n)^2$.

Example: $X_1, \dots, X_n \sim$ i.i.d. $U[0, \theta]$. Parameter is θ .

MOM: $\bar{X}_n = \frac{\theta}{2} \implies \theta^{MOM} = 2 \cdot \bar{X}_n$.

■■■

Remarks:

- Apart from these special cases above, for general density functions $f(\cdot|\vec{\theta})$, the MOM estimator is often difficult to calculate, because the “population moments” involve difficult integrals. (In Pearson’s original paper, the density was a mixture of two normal density functions:

$$f(x|\vec{\theta}) = \lambda \cdot \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + (1-\lambda) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)$$

with unknown parameters $\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2$.)

- The model assumption that $X_1, \dots, X_n \sim$ i.i.d. $f(\cdot|\vec{\theta})$ implies a number of moment equations equal to the number of moments, which can be $\gg K$. This leaves room for evaluating the model specification.

For example, in the uniform distribution example above, another moment condition which should be satisfied is that

$$\frac{1}{n} \sum_i X_i^2 = EX^2 = VX + (EX)^2 = \frac{\theta^2}{3} + \frac{\theta}{2}. \quad (1)$$

At the MOM estimator θ^{MOM} , one can see whether

$$\frac{1}{n} \sum_i X_i^2 = \frac{\theta^{MOM2}}{3} + \frac{\theta^{MOM}}{2}.$$

(Later, you will learn how this can be tested more formally.) If this does not hold, then that might be cause for you to conclude that the original specification that $X_1, \dots, X_n \sim \text{i.i.d. } U[0, \theta]$ is inadequate. Eq. (1) is an example is an **overidentifying restriction**.

- While the MOM estimator focuses on using the *sample uncentered moments* to construct estimators, there are other sample quantities which could be useful, such as the sample median (or other sample percentiles), as well as sample minimum or maximum. (Indeed, for the uniform case above, the sample maximum would be a very reasonable estimator for θ .) All these estimators are lumped under the rubric of “generalized method of moments” (GMM).



Maximum Likelihood Estimation

Let $X_1, \dots, X_n \sim \text{i.i.d.}$ with density $f(\cdot | \theta_1, \dots, \theta_K)$.

Define: the likelihood function, for a continuous random variable, is the joint density of the sample observations:

$$L(\vec{\theta} | x_1, \dots, x_n) = \prod_{i=1}^n f(x_i | \vec{\theta}).$$

View $L(\vec{\theta} | \vec{x})$ as a function of the parameters $\vec{\theta}$, for the data observations \vec{x} .

From “classical” point of view, the likelihood function $L(\vec{\theta} | \vec{x})$ is a random variable due to the randomness in the data \vec{x} . (In the “Bayesian” point of view, which we talk about later, the likelihood function is also random because the parameters $\vec{\theta}$ are also treated as random variables.)

The **maximum likelihood estimator (MLE)** are the parameter values $\vec{\theta}^{ML}$ which maximize the likelihood function:

$$\vec{\theta}^{ML} = \operatorname{argmax}_{\vec{\theta}} L(\vec{\theta}|\vec{x}).$$

Usually, in practice, to avoid numerical overflow problems, maximize the log of the likelihood function:

$$\vec{\theta}^{ML} = \operatorname{argmax}_{\vec{\theta}} \log L(\vec{\theta}|\vec{x}) = \sum_i \log f(x_i|\vec{\theta}).$$

Analogously, for discrete random variables, the likelihood function is the joint probability mass function:

$$L(\vec{\theta}|\vec{x}) = \prod_{i=1}^n P(X = x_i|\vec{\theta}).$$



Example: $X_1, \dots, X_n \sim \text{i.i.d. } N(\theta, 1)$.

- $\log L(\theta|\vec{x}) = \log(n/\sqrt{2\pi}) - \frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2$
- $\max_{\theta} \log L(\theta|\vec{x}) = \min_{\theta} \frac{1}{2} \sum_i (x_i - \theta)^2$
- FOC: $\frac{\partial \log L}{\partial \theta} = \sum_i (x_i - \theta) = 0 \Rightarrow \theta^{ML} = \frac{1}{n} \sum_i x_i$ (sample mean)

Also should check second order condition: $\frac{\partial^2 \log L}{\partial \theta^2} = -n < 0$: so satisfied.

Example: $X_1, \dots, X_n \sim \text{i.i.d. Bernoulli with prob. } p$. Unknown parameter is p .

- $L(p|\vec{x}) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$
-

$$\begin{aligned} \log L(p|\vec{x}) &= \sum_{i=1}^n [x_i \cdot \log p + (1-x_i) \cdot \log(1-p)] \\ &= y \log p + (n-y) \log(1-p) : \quad y \text{ is number of 1's} \end{aligned}$$

- FOC: $\frac{\partial \log L}{\partial p} = \frac{y}{p} - \frac{n-y}{1-p} \implies p^{ML} = \frac{y}{n}$

For $y = 0$ or $y = n$, the p^{ML} is (respectively) 0 and 1: corner solutions.

- SOC: $\left. \frac{\partial \log L}{\partial p^2} \right|_{p=p^{ML}} = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2} < 0$ for $0 < y < n$.

When parameter is multidimensional: check that the *Hessian matrix* $\frac{\partial^2 \log L}{\partial \theta \partial \theta'}$ is negative definite.



You can think of ML as a MOM estimator: for X_1, \dots, X_n i.i.d., and K -dimensional parameter vector θ , the MLE solves the FOCs:

$$\begin{aligned} \frac{1}{n} \sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta_1} &= 0 \\ \frac{1}{n} \sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta_2} &= 0 \\ &\vdots \\ \frac{1}{n} \sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta_K} &= 0. \end{aligned}$$

Under LLN: $\frac{1}{n} \sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta_k} \xrightarrow{p} E_{\theta_0} \frac{\partial \log f(X|\theta)}{\partial \theta_k}$, for $k = 1, \dots, K$, where the notation E_{θ_0} denote the expectation over the distribution of X at the true parameter vector θ_0 .

Hence, MLE is equivalent to MOM with the moment conditions

$$E_{\theta_0} \frac{\partial \log f(X|\theta)}{\partial \theta_k} = 0, \quad k = 1, \dots, K.$$



Bayes estimators

Fundamentally different view of the world. Model the unknown parameters $\vec{\theta}$ as random variables, and assume that researcher's beliefs about θ are summarized in a *prior distribution* $f(\theta)$.

In this sense, Bayesian approach is "subjective", because researcher's beliefs about θ are accommodated in inferential approach.

$X_1, \dots, X_n \sim$ i.i.d. $f(x|\theta)$: the Bayesian views the density of each data observation as a *conditional density*, which is conditional on a realization of the random variable θ .

Given data X_1, \dots, X_n , we can update our beliefs about the parameter θ by computing the posterior density (using Bayes Rule):

$$\begin{aligned} f(\theta|\vec{x}) &= \frac{f(\vec{x}|\theta) \cdot f(\theta)}{f(\vec{x})} \\ &= \frac{f(\vec{x}|\theta) \cdot f(\theta)}{\int f(\vec{x}|\theta)f(\theta)d\theta}. \end{aligned}$$

A Bayesian point estimate of θ is some feature of this posterior density. Common point estimators are:

- Posterior mean:

$$E[\theta|\vec{x}] = \int \theta f(\theta|\vec{x})d\theta.$$

- Posterior median: $F_{\theta|\vec{x}}^{-1}(0.5)$, where $F_{\theta|\vec{x}}$ is CDF corresponding to the posterior density: i.e., $F_{\theta|\vec{x}}(\tilde{\theta}) = \int_{-\infty}^{\tilde{\theta}} f(\theta|\vec{x})d\theta$.
- Posterior mode: $\max_{\theta} f(\theta|\vec{x})$. This is the point at which the density is highest.

Note that $f(\vec{x}|\theta)$ is just the likelihood function, so that the posterior density $f(\theta|\vec{x})$ can be written as:

$$f(\theta|\vec{x}) = \frac{L(\theta|\vec{x}) \cdot f(\theta)}{\int L(\theta|\vec{x})f(\theta)d\theta}.$$

But there is a difference in interpretation: in Bayesian world, the likelihood function is random due to both \vec{x} and θ , whereas in classical world, only \vec{x} is random.

Example: $X_1, \dots, X_n \sim$ i.i.d. $N(\theta, 1)$, with prior density $f(\theta)$.

$$\text{Posterior density } f(\theta|\vec{x}) = \frac{\exp(-\frac{1}{2} \sum_i (x_i - \theta)^2) f(\theta)}{\int \exp(-\frac{1}{2} \sum_i (x_i - \theta)^2) f(\theta) d\theta}.$$

Integral in denominator can be difficult to calculate: computational difficulties can hamper computation of posterior densities.

However, note that the denominator is not a function of θ . Thus

$$f(\theta|\vec{x}) \propto L(\theta|\vec{x}).$$

Hence, if we assume that $f(\theta)$ is constant (ie. uniform), for all possible values of θ , then the posterior mode $\operatorname{argmax}_{\theta} f(\theta|\vec{x}) = \operatorname{argmax}_{\theta} L(\theta|\vec{x}) = \theta^{ML}$.



Example: Bayesian updating for normal distribution, with normal priors

$X \sim N(\theta, \sigma^2)$, assume σ^2 is known.

Prior: $\theta \sim N(\mu, \tau^2)$, assume τ is known.

Then posterior distribution

$$\theta|X \sim N(E(\theta|X), V(\theta|X)),$$

where

$$E(\theta|X) = \frac{\tau^2}{\tau^2 + \sigma^2}X + \frac{\sigma^2}{\sigma^2 + \tau^2}\mu$$
$$V(\theta|X) = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

This is an example of a *conjugate* prior and conjugate distribution, where the posterior distribution comes from the same family as the prior distribution.

Posterior mean $E(\theta|X)$ is weighted average of X and prior mean μ .

In this case, as $\tau \rightarrow \infty$ (so that prior information gets worse and worse): then $E(\theta|X) \rightarrow X$ (a.s.). These are just the MLE (for just one data observation).

When you observe an i.i.d. sample $\vec{X}_n \equiv (X_1, \dots, X_n)$, with sample mean \bar{X}_n :

$$E(\theta|\vec{X}_n) = \frac{n\tau^2}{n\tau^2 + \sigma^2}\bar{X}_n + \frac{\sigma^2}{\sigma^2 + n\tau^2}\mu$$
$$V(\theta|\vec{X}_n) = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}.$$

In this case, as the number of observations $n \rightarrow \infty$, the posterior mean $E(\theta|\vec{X}_n) \rightarrow \bar{X}_n$. So as $n \rightarrow \infty$, the posterior mean converges to the MLE: when your sample becomes arbitrarily large, you place no weight on your prior information.

“Data augmentation” The important philosophical distinction of the Bayesian approach is that data and model parameters are treated on an “equal footing”. Hence, just as we make posterior inference on model parameters, we can also make posterior inference on unobserved variables in “latent variable” models, which are models where not all the model variables are observed.

Consider a simple example (the “binary probit” model):

$$\begin{aligned} z &= \beta x + \epsilon, \quad \epsilon \sim N(0, 1) \\ y &= \begin{cases} 0 & \text{if } z \geq 0 \\ 1 & \text{if } z < 0. \end{cases} \end{aligned} \quad (2)$$

The researcher observes (x, y) , but not (z, β) . He wishes to form the posterior of $z, \beta|x, y$.

We do all inference conditional on x . Therefore the relevant prior is

$$\begin{aligned} (z, \beta|x) &= (z|\beta, x) \cdot \beta|x \\ &= N(\beta x, 1) \cdot \underbrace{N(\bar{\beta}, a^2)}_{\equiv f(\beta)}. \end{aligned} \quad (3)$$

In the above, we assume the marginal prior on β is normal (and doesn’t depend on x). The conditional prior density of $z|\beta, x$ is derived from the model specification (2).

The posterior can also be factored into two parts:

$$\begin{aligned} (z, \beta|y, x) &= (z|\beta, y, x) \cdot (\beta|y, x) \\ &\propto (z|\beta, y, x) \cdot L(y|\beta, x) \cdot f(\beta) \\ &= \begin{cases} \left[\frac{\phi(z-\beta x)}{\Phi(\beta x)} \right] \cdot \Phi(\beta x) \cdot \frac{1}{a} \phi\left(\frac{\beta-\bar{\beta}}{a}\right) & \text{with support } z \geq 0, \text{ if } y = 1 \\ \left[\frac{\phi(z-\beta x)}{1-\Phi(\beta x)} \right] \cdot (1 - \Phi(\beta x)) \cdot \frac{1}{a} \phi\left(\frac{\beta-\bar{\beta}}{a}\right) & \text{with support } z < 0, \text{ if } y = 0 \end{cases} \\ &= \begin{cases} \phi(z - \beta x) \cdot \frac{1}{a} \phi\left(\frac{\beta-\bar{\beta}}{a}\right) & \text{with support } z \geq 0, \text{ if } y = 1 \\ \phi(z - \beta x) \cdot \frac{1}{a} \phi\left(\frac{\beta-\bar{\beta}}{a}\right) & \text{with support } z < 0, \text{ if } y = 0 \end{cases} \end{aligned} \quad (4)$$

In the above, Φ and ϕ denote the CDF and density functions for the $N(0, 1)$ distribution. In the second line, note that the proportionality constant (ie. the denominator in Baye’s rule) does not depend on (β, z) . In the third equation above, note that $(z|\beta, y, x)$ is a truncated standard normal distribution (with the direction of truncation depending on whether $y = 0$ or $y = 1$).

Accordingly, this can be marginalized over β to obtain the posterior of $z|y, x$. Using Bayesian procedure to do posterior inference on latent data variables is sometimes called “data augmentation”.

In non-Bayesian context, obtaining values for missing data values is usually done by some sort of imputation procedure. Thus, data augmentation can be viewed as a sort of Bayesian imputation procedure. One attractive feature of the Bayesian approach is that it follows easily and naturally from the usual Bayesian logic.