Lecture 04: Introduction to Point Estimation

Mathematical Statistics II, MATH 60062/70062

Tuesday February 1, 2022

Reference: Casella & Berger, 7.1-7.2.2

Point estimation problem

Given a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ and a parametric model for $\boldsymbol{X}$,

$$\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta}),$$

where the model parameter $\boldsymbol{\theta}$ is fixed and unknown, the **point estimation** problem seeks to find methods for

- Estimating $\boldsymbol{\theta}$
- Estimating some function of $\boldsymbol{\theta}$, say $\tau(\boldsymbol{\theta})$

Point estimator

A **point estimator**, $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$, is any function of the sample $\boldsymbol{X}$. Therefore, any statistic is a point estimator.

- The only restriction is that $W(\boldsymbol{X})$ cannot depend on $\boldsymbol{\theta}$.
- There is no mention of the range of $W(\boldsymbol{X})$.

Point estimator

A **point estimator**, $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$, is any function of the sample $\boldsymbol{X}$. Therefore, any statistic is a point estimator.

- The only restriction is that $W(\boldsymbol{X})$ cannot depend on $\boldsymbol{\theta}$.
- There is no mention of the range of $W(\boldsymbol{X})$.

An **estimator** is a function of the sample (i.e., a function of random variables), so an estimator is a *random variable*; an **estimate** is the realized value of an estimator (that is, a *number*), obtained when a sample is actually taken.

Point estimation methods

Methods of finding point estimators

- Method of moments
- Maximum likelihood estimation
- Bayesian estimation

Methods of evaluating point estimators

- Mean squared error
- Bias and variance
- Best unbiased estimators
- Sufficiency and unbiasedness

Often the methods of evaluating estimators will suggest new ones.

Method of moments

Let $X_1, \ldots, X_n$ be a sample from a population with PDF or PMF $f(x \mid \theta_1, \ldots, \theta_k)$. A **method of moments estimator** are found by equating the first k **sample moments** to the corresponding k **population moments** (typically functions of $\boldsymbol{\theta}$),

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \mu_1' = E(X)$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu_2' = E(X^2)$$

$$\vdots$$

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k = \mu_k' = E(X^k)$$

and solving for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

Poisson method of moments

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Pois}(\theta)$, where $\theta > 0$. Find the method of moments estimator for θ.

Poisson method of moments

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Pois}(\theta)$, where $\theta > 0$. Find the method of moments estimator for θ.

The first sample moment is

$$m_1 = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.$$

The first population moment is

$$\mu_1' = E(X) = \theta.$$

Therefore, the method of moments estimator for θ is

$$\hat{\theta} = \bar{X}.$$

Normal method of moments

Suppose $X_1, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. The first two population moments are $E(X) = \mu$ and $E(X^2) = \sigma^2 + \mu^2$.

Therefore, the method of moments estimators for μ and σ^2 can be found by solving

$$\bar{X} = \mu$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2.$$

The estimators are

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Binomial method of moments

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Bin}(k, p)$. The first two population moments are $E(X) = kp$ and $E(X^2) = kp(1-p) + k^2p^2$.

Therefore, the method of moments estimators for k and p can be found by solving

$$\bar{X} = kp$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = kp(1-p) + k^2p^2.$$

The estimators are

$$\hat{k} = \frac{\bar{X}^2}{\bar{X} - (1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad \hat{p} = \frac{\bar{X}}{\hat{k}}.$$

Binomial method of moments

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Bin}(k, p)$. The first two population moments are $E(X) = kp$ and $E(X^2) = kp(1-p) + k^2p^2$.

Therefore, the method of moments estimators for k and p can be found by solving

$$\bar{X} = kp$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = kp(1-p) + k^2p^2.$$

The estimators are

$$\hat{k} = \frac{\bar{X}^2}{\bar{X} - (1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad \hat{p} = \frac{\bar{X}}{\hat{k}}.$$

It's possible to get *negative* estimates of k and p. The range of the estimator does not coincide with the parameter it is estimating.

The likelihood function

Suppose $X_1, \ldots, X_n$ are an iid sample from a population with PDF or PMF $f_X(x \mid \boldsymbol{\theta})$. Given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f_X(x_i \mid \boldsymbol{\theta})$$

is called the **likelihood function**.

The likelihood function is the same function as the joint PDF/PMF $f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})$, but their interpretations are different.

The likelihood function

Suppose $X_1, \ldots, X_n$ are an iid sample from a population with PDF or PMF $f_X(x \mid \boldsymbol{\theta})$. Given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f_X(x_i \mid \boldsymbol{\theta})$$

is called the **likelihood function**.

The likelihood function is the same function as the joint PDF/PMF $f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})$, but their interpretations are different.

- The joint PDF/PMF $f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})$ describes the random behavior of sample $\boldsymbol{X}$ with a specified population ($\boldsymbol{\theta}$ is fixed).

The likelihood function

Suppose $X_1, \ldots, X_n$ are an iid sample from a population with PDF or PMF $f_X(x \mid \boldsymbol{\theta})$. Given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f_X(x_i \mid \boldsymbol{\theta})$$

is called the **likelihood function**.

The likelihood function is the same function as the joint PDF/PMF $f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})$, but their interpretations are different.

- The joint PDF/PMF $f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})$ describes the random behavior of sample $\boldsymbol{X}$ with a specified population ($\boldsymbol{\theta}$ is fixed).
- The likelihood function $L(\boldsymbol{\theta} \mid \boldsymbol{x})$ is viewed as a function of $\boldsymbol{\theta}$ with a fixed sample point $\boldsymbol{X} = \boldsymbol{x}$.

The likelihood function

Suppose $X_1, \ldots, X_n$ are an iid sample from a population with PDF or PMF $f_X(x \mid \boldsymbol{\theta})$. Given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f_X(x_i \mid \boldsymbol{\theta})$$

is called the **likelihood function**.

The likelihood function is the same function as the joint PDF/PMF $f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})$, but their interpretations are different.

- The joint PDF/PMF $f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})$ describes the random behavior of sample $\boldsymbol{X}$ with a specified population ($\boldsymbol{\theta}$ is fixed).
- The likelihood function $L(\boldsymbol{\theta} \mid \boldsymbol{x})$ is viewed as a function of $\boldsymbol{\theta}$ with a fixed sample point $\boldsymbol{X} = \boldsymbol{x}$.
- The likelihood function is *not* the PDF/PMF of the parameter, because $\int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta}$ is not necessarily one.

Maximum likelihood estimator

For each sample $\boldsymbol{x}$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ be a parameter value at which $L(\boldsymbol{\theta} \mid \boldsymbol{x})$ attains its maximum as a function of $\boldsymbol{\theta}$, with $\boldsymbol{x}$ held fixed. A **maximum likelihood estimator (MLE)** of the parameter $\boldsymbol{\theta}$ based on a sample $\boldsymbol{X}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$.

- The range of the MLE coincides with the range of the parameter.
- The MLE is the parameter point for which the observed sample is most likely.
- In general, the MLE is a good point estimator, possessing some of the optimality properties.

Finding the MLE

Finding the MLE is essentially a maximization problem,

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\, L(\boldsymbol{\theta} \mid \boldsymbol{x}).$$

If the likelihood is differentiable as a function of $\boldsymbol{\theta}$, *possible candidates* for the MLE are the values of $(\theta_1, \dots, \theta_k)$ that are solutions to

$$\frac{\partial}{\partial \theta_j} L(\boldsymbol{\theta} \mid \boldsymbol{x}) = 0, \quad j = 1, \dots, k.$$

- There may be multiple extreme points in the interior of the domain of the function; second-order conditions must be verified to ensure that $\hat{\boldsymbol{\theta}}$ is a maximizer.
- The boundary must be checked separately for extrema.

Normal MLE

Suppose that $X_1, \dots, X_n$ are iid $\mathcal{N}(\theta, 1)$, where $-\infty < \theta < \infty$. The likelihood function is

$$L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2} = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}.$$

Setting the first derivative

$$\frac{\partial}{\partial\theta} L(\theta \mid \boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} \sum_{i=1}^{n}(x_i - \theta)$$

equal to zero gives the solution $\hat{\theta} = \bar{x}$.

Normal MLE

Suppose that $X_1, \dots, X_n$ are iid $\mathcal{N}(\theta, 1)$, where $-\infty < \theta < \infty$. The likelihood function is

$$L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2} = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}.$$

Setting the first derivative

$$\frac{\partial}{\partial\theta} L(\theta \mid \boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} \sum_{i=1}^{n}(x_i - \theta)$$

equal to zero gives the solution $\hat{\theta} = \bar{x}$. The solution is a maximizer because $L(\theta \mid \boldsymbol{x})$ is concave when $\theta = \bar{x}$,

$$\frac{\partial^2}{\partial\theta^2} L(\theta \mid \boldsymbol{x})\Big|_{\theta=\bar{x}} < 0.$$

So, $\hat{\theta} = \bar{X}$ is the MLE for θ.

Log-likelihood function

When differentiation is used to find the maximum likelihood estimator, it is often easier to use the natural logarithm of the likelihood function, $\log L(\theta \mid \boldsymbol{x})$, called the **log-likelihood function**.

Note that

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\boldsymbol{x}) &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta} \mid \boldsymbol{x}) \\ &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \log L(\boldsymbol{\theta} \mid \boldsymbol{x}).\end{aligned}$$

The equations

$$\frac{\partial}{\partial\theta_j} \log L(\boldsymbol{\theta} \mid \boldsymbol{x}) = 0, \quad j = 1, \ldots, k,$$

are called the **score equations**.

Normal MLE

Suppose that $X_1, \ldots, X_n$ are iid $\mathcal{N}(\theta, 1)$, where $-\infty < \theta < \infty$. The log-likelihood function is

$$\log L(\theta \mid \boldsymbol{x}) = \log \left(\frac{1}{\sqrt{2\pi}} \right)^n - \frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2.$$

Setting the first derivative

$$\frac{\partial}{\partial \theta} \log L(\theta \mid \boldsymbol{x}) = \sum_{i=1}^{n} (x_i - \theta)$$

equal to zero gives the solution $\hat{\theta} = \bar{x}$. Compared to the likelihood function, it is much easier to verify the concavity of $\log L(\theta \mid \boldsymbol{x})$,

$$\frac{\partial^2}{\partial \theta^2} \log L(\theta \mid \boldsymbol{x}) \Big|_{\theta = \bar{x}} = -n < 0.$$