Advanced Statistics: Theory and Methods

Dr. Nandini Kannan

email : nandini.kannan@plaksha.edu.in

Chapter 4

Estimation

4.1 Introduction

Let $X_1, \ldots, X_n$ be i.i.d. random variables drawn from a population that has a distribution characterized by a parameter θ that is unknown.

A **point estimator** of θ is some statistic (function of the random variables in a random sample) which we will denote by $T(X_1, \ldots, X_n) = \hat{\Theta}$. The value of the statistic for a sample of size n will be referred to as a **point estimate** and denoted by $\hat{\theta}$. [We will use the lower-case notation $\hat{\theta}$ interchangeably to denote an estimator or estimate for convenience.]

In Chapter 3, we determined that statistics are random variables. The distribution of the statistic is referred to as its' sampling distribution.

4.2 Bias and Consistency

Definition: $\hat{\theta}$ is said to be an **unbiased** estimator of θ if

$$\mu_{\hat{\theta}} = E(\hat{\theta}) = \theta.$$

If the equality does not hold, $\hat{\theta}$ is said to be biased.

Definition: The bias of an estimator $\hat{\theta}$ is

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

If the bias is 0, the estimator is unbiased.

Example Let $X_1, \ldots, X_n$ be a random sample of size n. Then $\bar{X}$ is an unbiased estimator of μ and S^2 is an unbiased estimator of σ^2.

Solution:

We have

$$\begin{aligned} E(\bar{X}) &= E\left\{\frac{1}{n}(X_1 + \ldots + X_n)\right\} \\ &= \frac{1}{n}[E(X_1) + \ldots + E(X_n)] \\ &= \frac{1}{n}[\mu + \ldots + \mu] \\ &= \mu \end{aligned}$$

$$\begin{aligned} E(S^2) &= E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right\} \\ &= \frac{1}{n-1}E[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2] \\ &= \frac{1}{n-1}[\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}^2)] \\ &= \frac{1}{n-1}[n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2)] \\ &= \sigma^2 \end{aligned}$$

Definition: $\hat{\theta}$ is a weakly consistent estimator of θ, if for every $\epsilon > 0$

$$P\{|\hat{\theta} - \theta| \geq \epsilon\} \longrightarrow 0 \qquad \text{as } n \to \infty$$

or equivalently

$$P\{|\hat{\theta} - \theta| < \epsilon\} \longrightarrow 1 \qquad \text{as } n \to \infty.$$

We write $\hat{\theta} \overset{P}{\longrightarrow} \theta$

Theorem 4.2.1. *A sufficient condition for $\hat{\theta}$ to be consistent for estimating θ is that both the bias and variance of $\hat{\theta}$ tend to 0 as $n \to \infty$.*

Theorem 4.2.2. *Weak Law of Large Numbers: Let $X_1, \ldots, X_n$ be iid random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Then*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{P}{\longrightarrow} \mu.$$

Proof: An application of Chebyshev's inequality yields the following:

$$\begin{aligned} P(|\bar{X}_n - \mu| \geq \epsilon) &= P[(\bar{X}_n - \mu)^2 \geq \epsilon^2] \\ &\leq \frac{E(\bar{X}_n - \mu)^2}{\epsilon^2} = \frac{Var(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}. \end{aligned}$$

The last term goes to 0 as $n \to \infty$.

■

Example: Consistency of sample moments.

Let $X_1, \ldots, X_n$ be iid random variables with $E|X_1|^k < \infty$ for some positive k. Then

$$\frac{1}{n}\sum_{j=1}^{n} X_j^k \xrightarrow{P} \mu_k^{'}$$

as $n \to \infty$ where $\mu_k^{'} = E(X^k)$.

The sample mean and variance are consistent estimators.

The estimators we have used so far have been fairly intuitive. We have used the sample analogs of the corresponding population parameters. For example, the sample average $\bar{X}$ was used as an estimator of μ; the sample variance S^2 was used as an estimator of σ^2.
Are there formal method of finding estimators of unknown parameters?

There are two methods that are used extensively in Estimation Theory. The first is called the **Method of Moments**; the second

the **Method of Maximum likelihood**.

4.3 Method of Moments:

Definition: The $r-th$ *moment* of a random variable X is

$$\mu_r^{'} = E(X^r).$$

The $r-th$ *central moment* is

$$\mu_r = E[(X - \mu)]^r,$$

where $\mu = \mu_1^{'} = E(X)$.

The moments of a random variable will involve the unknown parameters of the underlying distribution. The method of moments consisting of equating the first few moments of the population to the corresponding sample moments and solving for the unknown parameters. We need as many equation as the number of unknown parameters.

Definition: The rth **sample moment** of a set of observations $x_1, \ldots, x_n$ is the mean of their rth powers and is denoted by $m_r^{'}$. We

have

$$m_r^{'} = \frac{1}{n}\sum_{i=1}^{n} x_i^r.$$

The Method of Moments consists of solving the system of equations

$$m_r^{'} = \mu_r^{'}, \qquad r = 1, \ldots, k$$

for the k parameters.

Example: Consider $X_1, \ldots, X_n$ iid Bernoulli random variables with probability of success given by p. We have

$$\mu_1^{'} = E(X) = p$$

The distribution has only one unknown parameter, so we need only one equation

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

which yields the estimator

$$\tilde{p} = \frac{X}{n}$$

where X is the total number of successes observed; i.e. the estimator of p is the sample proportion of successes.

Example: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ with both μ and σ^2 unknown.

We know that

$$\mu_1^{'} = \mu \qquad \text{and} \qquad \mu_2^{'} = E(X^2) = Var(X) + \mu^2 = \sigma^2 + \mu^2$$

The first two sample moments are

$$m_1^{'} = \bar{X}, \qquad m_2^{'} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

We equate the two sets of moments:

$$\bar{X} = \mu \tag{4.1}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2 \tag{4.2}$$

Substituting $\tilde{\mu} = \bar{X}$ from (1) into (2), we have

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

which yields

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 \\ &= \frac{1}{n}[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2] \end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Example: Let $X_1, \ldots, X_n$ be iid Poisson random variables with parameter λ.

We know that

$$\mu_1^{'} = \lambda$$

The first sample moment is

$$m_1^{'} = \bar{X}$$

Equating the two, we get the method of moments estimator $\tilde{\lambda} = \bar{X}$.

Example: Let $X_1, \ldots, X_n$ be iid $Gamma(\alpha, \beta)$ random variables.

We have

$$\mu_1^{'} = \mu = \alpha\beta \qquad \text{and} \qquad \mu_2^{'} = E(X^2) = \alpha\beta^2 + \mu^2 = \alpha\beta^2 + (\alpha\beta)^2$$

The first two sample moments are

$$m_1^{'} = \bar{X}, \qquad m_2^{'} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Equating the two sets of moments, we have

$$\bar{X} = \alpha\beta \tag{4.3}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \alpha\beta^2 + (\alpha\beta)^2 = \alpha(\alpha+1)\beta^2 \tag{4.4}$$

From (3), substitute $\alpha\beta = \bar{X}$ into (4) to get

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

which yields

$$\tilde{\beta} = \frac{1}{\bar{X}}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right]$$

and

$$\tilde{\alpha} = \frac{\bar{X}^2}{\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right]}$$

Method of moment estimators are simple to compute and have several nice properties. They are consistent and do not require the joint distribution of the data. However, they are not necessarily unique and often not the most efficient. We also see from some of the examples above that they are not always unbiased.

4.4 Maximum Likelihood

To help us understand the principle of maximum likelihood, let us start with an example.

I have a coin that is either (1) Fair with $P(H) = P(T) = 1/2$ or (2) Biased with $P(H) = 0.9, P(T) = 0.1$. Based on observed data I have to decide which coin I have in my possession.

I toss the coin once and observe a head: based on a single observation what decision could I make?

We know that $P(H) = 0.9$ for coin 2 and $P(H) = 0.5$ for coin 1. So based on a single toss, I would decide that the chances of observing a head are higher with coin 2 than with coin 1. Hence I choose coin 2.

What if I tossed the coin twice and observed 1 head and 1 tail? Let $p_1(x)$ denote the probability of outcome x with coin 1, and $p_2(x)$ denote the same probability with coin 2. We have

Outcome	$p_1(x)$	$p_2(x)$
2 Heads	0.25	0.81
1 Head, 1 Tail	0.50	0.18
2 Tails	0.25	0.01

Based on the probabilities and the observed data (1H, 1T), I will choose coin 1 because the chances of observing the particular outcome are higher for coin 1.

This is the principle of **maximum likelihood**: we look at the sample values and then choose as our estimates of the unknown parameters the values for which the probability of getting the sample values is maximum.

In the case of discrete random variables, if the observed sample values of a random sample are $x_1, \ldots, x_n$, the joint probability of observing these values is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1|\theta)f(x_2|\theta)...f(x_n|\theta)$$

by independence. This is the **joint distribution** of the n random

variables. Since the sample values have been observed and are fixed numbers, we regard $f(x_1|\theta)f(x_2|\theta)...f(x_n|\theta)$ as a function of θ and refer to this function as the **likelihood function** $L(\theta)$.

The interpretation of the likelihood function as a joint probability applies only in the case of discrete random variables. An analogous definition applies when the random sample comes from a continuous population. In that case $L(\theta)$ is the value of the joint probability density function of the random variable $X_1, X_2, \ldots, X_n$ at $x_1, \ldots, x_n$.

Definition: If $x_1, \ldots, x_n$ are the values of a random sample from a population with the parameter θ, the **likelihood function** of the sample is

$$L(\theta) = f(x_1|\theta)f(x_2|\theta)...f(x_n|\theta).$$

The **maximum likelihood estimator** (MLE) of θ, denoted by $\hat{\theta}$ is the value of θ that maximizes the likelihood function (or log-likelihood).

To maximize the likelihood, we take derivatives with respect to

the unknown parameters, set the derivatives equal to 0 and solve.

The derivative of the log-likelihood function is called the **score function**.

Example: Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ. Find the MLE of λ.

Solution: The likelihood function based on $x_1, \ldots, x_n$ is

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.$$

Maximizing $L(\lambda)$ wrt λ is equivalent to maximizing $\ln L(\lambda)$. We have

$$\ln L(\lambda) = -n\lambda + \sum_{i=1}^{n} x_i \ln(\lambda) - \ln(\prod_{i=1}^{n} x_i!)$$

Then

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0$$

Solving for λ, we get

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

which is a reasonable estimator.

To check if the solution is indeed a maximum, we take the second derivative

$$\frac{\partial^2 \ln L(\lambda)}{\partial \lambda^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2}$$

which is negative.

Example: Let $X_1, \ldots, X_n$ be a random sample from an Exponential distribution with parameter β. Find the MLE of β.

Solution: The likelihood function based on $x_1, \ldots, x_n$ is

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-\sum_{i=1}^{n} x_i/\beta}.$$

Maximizing $L(\beta)$ wrt β is equivalent to maximizing $\ln L(\beta)$. We have

$$\ln L(\beta) = -n \ln \beta - \frac{\sum_{i=1}^{n} x_i}{\beta}$$

Then

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \frac{-n}{\beta} + \frac{\sum_{i=1}^{n} x_i}{\beta^2} = 0$$

Solving for β, we get

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

which is a reasonable estimator.

To check if the solution is indeed a maximum, we take the second derivative and evaluate it at the value $\hat{\beta}$. We have

$$\frac{\partial^2 \ln L(\beta)}{\partial \beta^2} = \frac{n}{\beta^2} - 2\frac{\sum_{i=1}^{n} x_i}{\beta^3} = \frac{n}{\bar{x}^2} - 2\frac{n\bar{x}}{\bar{x}^3} = -\frac{n}{\bar{x}^2}$$

which is negative.

Example: Let $X_1, \ldots, X_n$ be a random sample from a Normal distribution with parameters μ and σ^2. Find the MLE's of μ and σ^2.

Solution: The likelihood function based on $x_1, \ldots, x_n$ is

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \\ & \quad exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}\right]. \end{aligned}$$

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}$$

We need to take partial derivatives wrt both parameters:

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = -\frac{2}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)(-1)}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2}$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Setting both derivatives to 0, we obtain

$$\begin{aligned} \sum_{i=1}^{n} x_i - n\mu &= 0 \\ \sum_{i=1}^{n}(x_i-\mu)^2 - n\sigma^2 &= 0 \end{aligned}$$

The MLE of μ is

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

and the MLE of σ^2 is

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Example: Let $X_1, \ldots, X_n$ be iid $Gamma(\alpha, \beta)$ random variables.

We have

$$L(\alpha, \beta) = \frac{1}{[\Gamma(\alpha)\beta^{\alpha}]^n}\prod_{i=1}^{n} x_i^{\alpha-1} e^{-\sum_{i=1}^{n} x_i/\beta}.$$

The log-likelihood is

$$lnL(\alpha, \beta) = -nln\Gamma(\alpha) - n\alpha ln\beta + (\alpha - 1)\sum_{i=1}^{n} lnx_i - \sum_{i=1}^{n} x_i/\beta.$$

The partial derivatives are

$$\frac{\partial \ln L(\alpha, \beta)}{\partial \alpha} = -nln\beta + \sum_{i=1}^{n} lnx_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \tag{4.5}$$

$$\frac{\partial \ln L(\alpha, \beta)}{\partial \beta} = -\frac{n\alpha}{\beta} + \frac{\sum_{i=1}^{n} x_i}{\beta^2}. \tag{4.6}$$

Setting the second partial derivative equal to 0, we obtain

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i}{n\hat{\alpha}} = \frac{\bar{X}}{\hat{\alpha}}.$$

Substituting this into the first equation, we get

$$-nln\bar{X} + nln\hat{\alpha} + \sum_{i=1}^{n} lnx_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0,$$

which is a nonlinear equation and requires an iterative method for its' solution.

4.4.1 Properties of MLE's

Theorem 4.4.1. *Invariance of the MLE: If $\hat{\theta}$ is the MLE of θ, then the MLE of $g(\theta)$ is $g(\hat{\theta})$.*

Theorem 4.4.2. *Let $X_1, \ldots, X_n$ be iid from a population with a distribution that satisfies certain regularity conditions. Then the MLE is consistent, i.e.*

$$\hat{\theta} \overset{P}{\longrightarrow} \theta$$

Define

$$I(\theta) = E\left[\frac{\partial}{\partial\theta} log f(X|\theta)\right]^2.$$

This is the Fisher Information.

Theorem 4.4.3. *Let $X_1, \ldots, X_n$ be iid from a population with a distribution that satisfies certain regularity conditions. Then $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ tends to a standard normal distribution.*

4.5 Accuracy and Precision

What are properties a good estimator must possess? We would clearly like estimators to be consistent and perhaps unbiased.

Since estimators are random variables, their values will vary from sample to sample. We would like this variability to be small. One such measure is the Mean Squared Error (MSE):

Definition Let T be a statistic and θ the parameter of interest. Then the MSE of T as an estimator of θ is

$$MSE[T;\theta] = E[(T-\theta)^2].$$

The MSE can be expressed as

$$MSE[T;\theta] = Var(T) + [Bias(T)]^2.$$

- The variance of an estimator is referred to as its **precision,** and the bias its **accuracy**.
- If T is an unbiased estimator, then its MSE is simply its variance.

Definition: The **relative efficiency** of T_1 compared to T_2 as an

estimator of θ is

$$RE(T_1, T_2) = \frac{MSE[T_1; \theta]}{MSE[T_2; \theta]}.$$

If $RE(T_1, T_2) > 1$, then T_1 is more efficient than T_2.

If T_1 and T_2 are both unbiased, then the relative efficiency is just the ratio of the variances.

Definition: If we consider all possible unbiased estimators of a parameter θ, the one with the smallest variance is called the **most efficient estimator** of θ.