Method of Maximum Likelihood

15·11. Method of Maximum Likelihood Estimation. From theoretical point of view, the most general method of estimation known is the method of *Maximum Likelihood Estimators* (M.L.E.) which was initially formulated by C.F. Gauss but as a general method of estimation was first introduced by Prof. R.A. Fisher and later on developed by him in a series of papers. Before introducing the method we will first define *Likelihood Function.*

Likelihood Function. *Definition.* Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a population with density function $f(x, \theta)$. Then the likelihood function of the sample values $x_1, x_2, \ldots, x_n$, usually denoted by $L = L(\theta)$ is their joint density function, given by

$$L = f(x_1, \theta)\, f(x_2, \theta) \ldots f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta). \quad \ldots(15{\cdot}53)$$

L gives the relative likelihood that the random variables assume a particular set of values $x_1, x_2, \ldots, x_n$. For a given sample $x_1, x_2, \ldots, x_n$, L becomes a function of the variable θ, the parameter.

The principle of maximum likelihood consists in finding an estimator for the unknown parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, say, which maximises the likelihood function $L(\theta)$ for variations in parameter *i.e.*, we wish to find $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ so that

$$L(\hat{\theta}) > L(\theta) \qquad \forall\ \theta \in \Theta$$

i.e.,
$$L(\hat{\theta}) = \text{Sup}\, L(\theta)\ \forall\ \theta \in \Theta.$$

Thus if there exists a function $\hat{\theta} = \hat{\theta}\,(x_1, x_2, \ldots, x_n)$ of the sample values which maximises L for variations in θ, then $\hat{\theta}$ is to be taken as an estimator of θ. $\hat{\theta}$ is usually called *Maximum Likelihood Estimator* (*M.L.E.*). Thus $\hat{\theta}$ is the solution, if any, of

$$\frac{\partial L}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial^2 L}{\partial \theta^2} < 0 \qquad \ldots(15{\cdot}54)$$

Since $L > 0$, and $\log L$ is a non-decreasing function of L; L and $\log L$ attain their extreme values (maxima or minima) at the same value of $\hat{\theta}$. The first of the two equations in (15·54) can be rewritten as

$$\frac{1}{L} \cdot \frac{\partial L}{\partial \theta} = 0 \quad \Rightarrow \quad \frac{\partial \log L}{\partial \theta} = 0, \qquad \ldots(15{\cdot}54a)$$

a form which is much more convenient from practical point of view.

If θ is vector valued parameter, then $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$, is given by the solution of simultaneous equations :

$$\frac{\partial}{\partial \theta_i} \log L = \frac{\partial}{\partial \theta_i} \log L\,(\theta_1, \theta_2, \ldots, \theta_k) = 0\ ;\ i = 1, 2, \ldots, k \qquad \ldots(15{\cdot}54b)$$

Equations (15·54*a*) and (15·54*b*) are usually referred to as the *Likelihood Equations* for estimating the parameters.

Remark. For the solution $\hat{\theta}$ of the likelihood equations, we have to see that the second derivative of L w.r. to θ is negative. If θ is vector valued, then for L to be maximum, the matrix of derivatives

$$\left(\frac{\partial^2 \log L}{\partial \theta_i\, \partial \theta_j}\right)_{\theta = \hat{\theta}} \text{ should be negative definite.}$$

15·11·1. Properties of Maximum Likelihood Estimators.

We make the following assumptions, known as the *Regularity Conditions* :

(*i*) The first and second order derivatives, *viz.*, $\dfrac{\partial \log L}{\partial \theta}$ and $\dfrac{\partial^2 \log L}{\partial \theta^2}$ exist and are continuous functions of θ in a range R (including the true value θ_0 of the parameter) for almost all x. For every θ in R

$$\left|\frac{\partial}{\partial \theta} \log L\right| < F_1(x) \quad \text{and} \quad \left|\frac{\partial^2}{\partial \theta^2} \log L\right| < F_2(x)$$

where $F_1(x)$ and $F_2(x)$ are integrable functions over $(-\infty, \infty)$.

(*ii*) The third order derivative $\dfrac{\partial^2}{\partial \theta^3} \log L$ exists such that

$$\left|\frac{\partial^3}{\partial \theta^3} \cdot \log L\right| < M(x)$$

where $E[M(x)] < K$, a positive quantity.

(*iii*) For every θ in R,

$$E\left(-\frac{\partial^2}{\partial \theta^2} \log L\right) = \int_{-\infty}^{\infty} \left(-\frac{\partial^2}{\partial \theta^2} \log L\right) L dx$$
$$= I(\theta),$$

is finite and non-zero.

(*iv*) The range of integration is independent of θ. But if the range of integration depends on θ, then $f(x, \theta)$ vanishes at the extremes depending on θ.

This assumption is to make the differentiation under the integral sign valid.

Under the above assumptions M.L.E. possesses a number of important properties, which will be stated in the form of theorems.

Theorem 15·11. (*Cramer-Rao Theorem*). "*With probability approaching unity as $n \to \infty$, the likelihood equation $\frac{\partial}{\partial \theta} \log L = 0$, has a solution which converges in probability to the true value θ_0". In other words M.L.E.'s are consistent.*

Remark. *MLE's* are always consistent estimators but need not be unbiased. For example in sampling from $N(\mu, \sigma^2)$ population, [c.f. Example 15·31],

MLE(μ) = $\bar{x}$ (sample mean), which is both unbiased and consistent estimator of μ.

MLE(σ^2) = s^2 (sample variance), which is consistent but not unbiased estimator of σ^2.

Theorem 15·12. (*Hazoor Bazar's Theorem*). *Any consistent solution of the likelihood equation provides a maximum of the likelihood with probability tending to unity as the sample size (n) tends to infinity.*

Theorem 15·13. (Asymptotic Normality of MLE's). *A consistent solution of the likelihood equation is asymptotically normally distributed about the true value* θ_0. Thus, $\hat{\theta}$ is asymptotically $N\left(\theta_0, \dfrac{I}{I(\theta_0)}\right)$, as $n \rightarrow \infty$.

Remark. Variance of M.L.E. is given by

$$V(\hat{\theta}) = \frac{1}{I(\theta)} = \frac{1}{\left[E\left(-\dfrac{\partial^2}{\partial\theta^2}\log L\right)\right]} \qquad \ldots(15{\cdot}55)$$

Theorem 15·14. *If M.L.E. exists, it is the most efficient in the class of such estimators.*

Theorem 15·15. *If a sufficient estimator exists, it is a function of the Maximum Likelihood Estimator.*

Proof. If $t = t(x_1, x_2, \ldots, x_n)$ is a sufficient estimator of θ, then Likelihood Function can be written as (*c.f.* Theorem 15·7)

$$L = g(t, \theta)\, h(x_1, x_2, x_3, \ldots, x_n \mid t)$$

where $g(t, \theta)$ is the density function of t and $h(x_1, x_2, \ldots, x_n \mid t)$ is the density function of the sample, given t, and is independent of θ.

$$\therefore \qquad \log L = \log g(t, \theta) + \log h(x_1, x_2, \ldots, x_n \mid t)$$

Differentiating w.r.t. θ, we get

$$\frac{\partial \log L}{\partial \theta} = \frac{\partial}{\partial \theta} \log g(t, \theta) = \psi(t, \theta), \text{ (say)}, \qquad \ldots(15{\cdot}56)$$

which is a function of t and θ only.

M.L.E. is given by

$$\frac{\partial \log L}{\partial \theta} = 0 \quad \Rightarrow \quad \psi(t, \theta) = 0$$

$$\therefore \qquad \hat{\theta} = \eta(t) = \text{Some function of sufficient statistic.}$$

$$\Rightarrow \qquad \hat{t} = \psi(\theta) = \text{Some function of M.L.E.}$$

Hence the theorem.

Theorem 15·17. (Invariance Property of MLE). *If T is the MLE of θ and $\psi(\theta)$ is one to one function of θ, then $\psi(T)$ is the MLE of $\psi(\theta)$.*

Example 15·31. *In random sampling from normal population $N(\mu; \sigma^2)$, find the maximum likelihood estimators for*

(i) μ *when* σ^2 *is known,*

(ii) σ^2 *when* μ *is known, and*

(iii) the simultaneous estimation of μ *and* σ^2.

Solution. $X \sim N(\mu,\ \sigma^2)$ then

$$L = \prod_{i=1}^{n}\left[\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}\right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right\}$$

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

Case (*i*). When σ^2 is known, the likelihood equation for estimating μ is

$$\frac{\partial}{\partial\mu}\log L = 0 \;\Rightarrow\; -\frac{1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-\mu)(-1) = 0$$

or $$\sum_{i=1}^{n}(x_i-\mu) = 0 \;\Rightarrow\; \sum_{i=1}^{n}x_i - n\mu = 0$$

$\Rightarrow$ $$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x} \qquad \ldots(*)$$

Hence M.L.E. for μ is the sample mean $\bar{x}$.

Case (*ii*). When μ is known, the likelihood equation for estimating σ^2 is

$$\frac{\partial}{\partial \sigma^2} \log L = 0 \Rightarrow -\frac{n}{2} \times \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

$$\Rightarrow \quad n - \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0, \quad i.e., \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \qquad \ldots(**)$$

Case (*iii*). The likelihood equations for simultaneous estimation of μ and σ^2 are

$$\frac{\partial}{\partial \mu} \log L = 0 \quad \text{and} \quad \frac{\partial}{\partial \sigma^2} \log L = 0, \text{ thus giving}$$

$$\hat{\mu} = \bar{x} \qquad \text{[From (*)]}$$

and
$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \qquad \text{[From (**)]}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2, \text{ the sample variance.}$$

Remark. Since M.L.E. is the most efficient, we conclude that in sampling from a normal population, the sample mean $\bar{x}$ is the most efficient estimator of the population mean μ.

Example 15·32. *Prove that the maximum likelihood estimate of the parameter* α *of a population having density function :*

$$\frac{2}{\alpha^2}(\alpha - x),\ 0 < x < \alpha$$

for a sample of unit size is $2x$, x *being the sample value. Show also that the estimate is biased.*

Solution. For a random sample of unit size ($n = 1$), the likelihood function is :

$$L(\alpha) = f(x, \alpha) = \frac{2}{\alpha^2}(\alpha - x) \; ; \; 0 < x < \alpha$$

Likelihood equation gives :

$$\frac{d}{d\alpha} \log L = \frac{d}{d\alpha}\left[\log 2 - 2 \log \alpha + \log (\alpha - x)\right] = 0$$

$$\Rightarrow \quad -\frac{2}{\alpha} + \frac{1}{\alpha - x} = 0 \quad \Rightarrow \quad 2(\alpha - x) - \alpha = 0 \quad \Rightarrow \quad \alpha = 2x$$

Hence MLE of α is given by $\hat{\alpha} = 2x$.

$$E(\hat{\alpha}) = E(2X) = 2 \int_0^{\alpha} x . f(x, \alpha)\, dx$$

$$= \frac{4}{\alpha^2} \int_0^{\alpha} x(\alpha - x)\, dx = \frac{4}{\alpha^2} \left| \frac{\alpha x^2}{2} - \frac{x^3}{3} \right|_0^{\alpha} = \frac{2}{3} \alpha$$

Since $E(\hat{\alpha}) \neq \alpha$, $\hat{\alpha} = 2x$ is not an unbiased estimate of α.

Example 15·33. (*a*) *Find the maximum likelihood estimate for the parameter* λ *of a Poisson distribution on the basis of a sample of size* n. *Also find its variance.*

Solution. The probability function of the Poisson distribution with parameter λ is given by

$$P(X = x) = f(x, \lambda) = \frac{e^{-\lambda} \lambda^x}{x\,!} ; x = 0, 1, 2, \ldots$$

Likelihood function of random sample $x_1, x_2, \ldots, x_n$ of n observations from this population is

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

$$\therefore \quad \log L = -n\lambda + \left(\sum_{i=1}^{n} x_i\right) \log \lambda - \sum_{i=1}^{n} \log (x_i\,!)$$

$$= -n\lambda + n\bar{x} \log \lambda - \sum_{i=1}^{n} \log (x_i\,!)$$

The likelihood equation for estimating λ is

$$\frac{\partial}{\partial \lambda} \log L = 0 \quad \Rightarrow \quad -n + \frac{n\bar{x}}{\lambda} = 0 \quad \Rightarrow \quad \lambda = \bar{x}$$

Thus the M.L.E. for λ is the sample mean $\bar{x}$.