Lecture 11: Parametric Estimation

In this lecture we are going to devise methods to find estimators of population parameters satisfying certain criterion. A population parameter is estimated using what is called a sample statistic. A sample statistic is a function of a random sample. Let us consider a population $f_X(\cdot;\theta)$, where θ denotes the population parameter. If θ is known then, the distribution of X is completely known. The idea of estimation comes in when θ is not known. In fact θ can be a vector or a scalar. If $X \sim$ Poisson (λ), then $\theta = \lambda$, a scalar, where as if $X \sim N(\mu, \sigma^2)$, then $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ with $\theta_1 = \mu$ & $\theta_2 = \sigma^2$.

For the moment assume $\theta \in \mathbb{R}$. Then having a random sample $X_1 \ldots X_n$ of size n from $f_X(\cdot, \theta)$, we construct a statistic symbolically given as

$$T = t(X_1, X_2 \ldots X_n)$$

We say that the random variable T is an unbiased estimator of θ if

$$E(T) = \theta.$$

Now the expectation of T is calculated based on the sampling distribution of T.

The expression $E(T) - \theta$ is called the bias in the estimation if $E(T) - \theta \neq 0$.

If $\theta \in \mathbb{R}^k$ say ie $\theta = (\theta_1 \ldots \theta_k)$, then we can have k different statistics, $T_i = t_i(X_1 \ldots X_n)$, $i = 1 \ldots k$ which can acts as estimator of θ.

Note that $\bar{X}$, the sample mean & S^2 the sample variance have been proved in the last chapter as unbiased estimators of population mean and population variance respectively.

Such an approach to get an approximation for θ is often called point ~~estimate~~ estimation. The statistic T when viewed as a random variable is called an <u>estimator</u> and the specific value of T is called an <u>estimate</u>. Sometimes $\hat{\theta}$ is denoted as an estimate of θ. There are several approaches to find an estimator. We would also sometimes refer $\hat{\theta}$ as an estimator of θ corresponding to the estimate $\hat{\theta}$.

Section 1: Method of Moments

Let $X_1 \dots X_n$ be a random sample of size n drawn from a population $f_X(x; \theta)$, where $\theta \in \mathbb{R}^k$, with $\theta = (\theta_1 \dots \theta_k)$. Let us recall the r-th moment of X, i.e

$$\mu'_r = E[X^r]$$

$$\therefore \quad \mu'_r = \int x^r f_X(x, \theta_1 \dots \theta_k)\, dx$$

$\therefore$ We can write

$$\mu'_r = \mu'_r(\theta_1 \dots \theta_k)$$

The r-th sample moment

$$M'_r = \frac{X_1^r + X_2^r + \dots + X_n^r}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

To find $\theta = (\theta_1 \dots \theta_k)$, set up the k-equations, in $\theta_1 \dots \theta_k$

$$\mu'_j(\theta_1 \dots \theta_k) = M'_j, \quad j = 1 \dots k$$

Suppose $\hat{\theta}_1, \hat{\theta}_2 \dots \hat{\theta}_k$ are the solutions to the equations. If say $\hat{\theta}_1, \hat{\theta}_2, \dots \hat{\theta}_k$, is unique, then we can consider $\hat{\theta}_1, \dots \hat{\theta}_k$ to be the estimates of $\theta_1 \dots \theta_k$.

Eg 1: $X_1 \ldots X_n$ be a random sample from a exponential population

$$f_X(x,\theta) = \theta e^{-\theta x}; \quad \theta > 0, \quad x \in (0,\infty).$$

Thus we will solve one equation

$$\therefore \quad M_1' = \mu_1' = \mu_1'(\theta) = \frac{1}{\theta}$$

$$\therefore \quad \hat{\theta} = \frac{1}{M_1'} \quad \text{So if we take a particular sample } X_1 = x_1, \ldots X_n = x_n, \text{ then } \hat{\theta} = \frac{n}{x_1 + \cdots + x_n}$$

So $\hat{\theta}$ is estimated by $\frac{1}{\bar{X}} = \frac{n}{X_1 + \cdots + X_n}$

$$\therefore \quad \hat{\Theta} = l(X_1 \ldots X_n) = \frac{n}{X_1 + \cdots + X_n}$$

$\hat{\Theta}$ is the estimator.

Eg. 2: $X_1 \ldots X_n$ be a random sample of size n, drawn from a normal population with mean μ and variance σ^2

$\therefore \theta = (\mu, \sigma^2)$.

$$\therefore \quad M_1' = \mu_1' = \mu$$

$$M_2' = \mu_2' = \sigma^2 + \mu^2$$

μ is estimated by $M_1' = \bar{X}$. ie $\bar{X}$ is the unbiased estimator of μ.

σ^2 is estimated by the statistic

$$\hat{\Theta}_2 = M_2' - \mu^2$$

$$\therefore \hat{\Theta}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

Now $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2\right]$

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

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

$$\therefore \quad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$\therefore$ $\hat{\theta}$ is the estimator of σ^2 and note that through the method of moments we do not have S^2 as the estimator of σ^2. Thus $\hat{\theta}$ is not an un-biased estimator of σ^2. Thus method of moments need not give us unbiased estimators always.

Section 2: Maximum Likelihood Estimator

The technique of maximum likelihood function is based on using the fact that $X_1 \ldots X_n$ are iid random variables.

Let us first look at the simple case where $\theta \in \mathbb{R}$. Let us draw a sample of size n, from a population $f_X(x;\theta)$. The likelihood function in this case is defined as

$$L(x,\theta) = L(\theta) = f_{X_1 \ldots X_n}(x_1 \ldots x_n, \theta)$$

$$L(\theta) = f_{X_1}(x_1,\theta), f_{X_2}(x_2,\theta), \ldots f_{X_n}(x_n,\theta) \rightarrow (\text{By independence})$$

$$\therefore L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i,\theta) \ *$$

We compute $\max_{\theta \in A} L(\theta)$, where A is the set in which θ is restricted to belong. We find that value of θ, which maximizes $L(\theta) = L(\theta; x_1 \ldots x_n)$, for a chosen sample. Then if $\hat{\theta}$ is the maximizer for a chosen sample i.e. $\hat{\theta} = \hat{\theta}(x_1 \ldots x_n)$, then $\hat{\theta}(x_1 \ldots x_n)$ is are called an maximum likelihood estimator estimate of θ for the chosen sample values, $X_1 = x_1, \ldots, X_n = x_n$. So the estimator of θ is

$$\hat{\theta} = \hat{\theta}(X_1 \ldots X_n).$$

Thus $\hat{\theta}$ is called the maximum likelihood estimator of θ

Noting that $\ln$ (i.e. $\log_e$) is an increasing function. We know that if $L(\theta) > 0$, $\forall \theta \in A$, then if $\hat{\theta}$ is the minimizer, we have

$$L(\theta) \leq L(\hat{\theta}), \quad \forall \theta \in A$$

$$\Rightarrow \ln L(\theta) \leq \ln L(\hat{\theta}) \quad \forall \theta \in A$$

Thus $\hat{\theta}$ is also the maximizer of $\ln L(\theta)$, over A. In most cases the set A is an open interval. Thus to compute $\hat{\theta}$ we can use the equation

$$\boxed{\frac{\partial}{\partial \theta} \ln \theta = 0} \longrightarrow (*)$$

Then also check the ex second order condition $\left.\frac{\partial^2}{\partial \theta^2} \ln \theta\right|_{\theta=\hat{\theta}} < 0$, where $\left.\frac{\partial}{\partial \theta} \ln \theta\right|_{\theta=\hat{\theta}} = 0$, $\hat{\theta}$ solves $(*)$. Such a $\hat{\theta}$ will provide an estimator or maximum likelihood estimator using the expression

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

Now suppose $\theta \in \mathbb{R}^k$, and we draw a random sample of size n $x_1 \ldots x_n$, from a population described by $f_X(x; \theta) = f_X(x, \theta_1 \ldots \theta_k)$.

In this case we have the likelihood function as

$$L(x, \theta_1 \ldots \theta_n) = f_{x_1 \ldots x_n}(x_1 \ldots x_n, \theta_1 \ldots \theta_k)$$

$$= f_{X_1}(x_1, \theta_1 \ldots \theta_k) \cdot f_2(x_2, \theta_1 \ldots \theta_k), \ldots, f_n(x_n, \theta_1 \ldots \theta_k)$$

$$= \prod_{i=1}^{n} f_{x_i}(x_i, \theta_1 \ldots \theta_k).$$

$\therefore$ Similarly, the maximum likelihood function ~~estimat~~ estimation problem as before is given as.

$$\max_{\theta \in K} L \text{ or } L(x, \theta_1 \ldots \theta_k), \text{ where } K \subset \mathbb{R}^k$$

or

$$\max_{\theta \in K} \ln L(x, \theta_1 \ldots \theta_k), \text{ where } K \subset \mathbb{R}^k$$

E.g Now if $\hat{\theta}$ is the maximizer, then as before we have $\hat{\theta} = (\hat{\theta}_1 \dots \hat{\theta}_k)$

$$\hat{\Theta}_i = \hat{\theta}_i(X_1 \dots X_n), \quad i = 1 \dots k$$

So there are k, maximum likelihood estimators. We shall now provide two examples. In the first case where $\theta \in \mathbb{R}$, and in the second case $\theta \in \mathbb{R}^2$.

E.g 3: Let $X_1 \dots X_n$ be a random sample drawn from an exponential population $f_X(\cdot, \theta)$. and is given as

$$f_X(x, \theta) = \theta e^{-\theta x}, \quad x > 0, \quad \theta > 0.$$

$$\therefore L(\theta, \underline{x}) = L(\theta) = \prod_{i=1}^{n} f_{X_i}(x, \theta)$$

$$= \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

$$= \prod_{i=1}^{n} (\theta)^n e^{-\theta \sum_{i=1}^{n} x_i}$$

[The maximum likelihood estimation problem (MLE)
$\max_{\theta > 0} \ln L(\theta)$ (P1)]

$$\therefore \quad \ln L_1(\theta) = n \ln \theta + \left(-\theta \cdot \sum_{i=1}^{n} x_i\right).$$

Thus $$\frac{\partial}{\partial \theta} \ln L_1(\theta) = n \cdot \frac{1}{\theta} - \sum_{i=1}^{n} x_i = 0$$

$$\therefore \quad \frac{n}{\theta} = \sum_{i=1}^{n} x_i$$

$$\therefore \quad \frac{1}{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X}$$

or $$\theta = \frac{1}{\bar{X}}.$$

Now $$\frac{\partial^2}{\partial \theta^2} L_1(\theta) = -\frac{n}{\theta^2} = -n\bar{X}^2 < 0$$

Thus $\theta = \frac{1}{\bar{X}}$ maximizes (strictly) problem P1).

$\therefore \hat{\Theta} = \frac{1}{\bar{X}}$ is the m.l.e estimator of θ

(6) m.l.e: Short form for "maximum likelihood estimator"

Now for the case when there are more than one parameter, i.e. $\theta \in \mathbb{R}^k$, we first have to find $\hat{\theta} = (\hat{\theta}_1 \dots \hat{\theta}_k)$ satisfying

$$\frac{\partial \ln L_x(\theta_1 \dots \theta_k)}{\partial \theta_j} = 0 \quad \forall \; j = 1 \dots k.$$

Once we find a $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2 \dots \hat{\theta}_k)$, we compute the Hessian matrix of $\nabla^2 \ln L_x(\theta_1 \dots \theta_k)$, and see if this negative definite at $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$. In the above expression we take took the liberty of writing $L_x(x, \theta_1 \dots \theta_k) = L(\theta_1 \dots \theta_k)$.

When $\theta \in \mathbb{R}^2$, i.e. $\theta = (\theta_1, \theta_2)$, then we have to chat check the following. First find $(\hat{\theta}_1, \hat{\theta}_2)$ satisfying

$$\left.\begin{aligned} \frac{\partial}{\partial \theta_1} \ln L_x(\theta_1, \theta_2) &= 0 \\ \frac{\partial}{\partial \theta_2} \ln L_x(\theta_1, \theta_2) &= 0 \end{aligned}\right\}$$

Then the Hessian matrix of $\ln L_x(\theta_1, \theta_2)$ is given as

$$\nabla^2 \ln L_x(\theta_1, \theta_2) = \begin{bmatrix} \frac{\partial^2}{\partial \theta_1^2} \ln L_x & \frac{\partial^2}{\partial \theta_2 \partial \theta_1} \ln L_x \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \ln L_x & \frac{\partial^2}{\partial \theta_2^2} \ln L_x \end{bmatrix}$$

To show that $\nabla^2 \ln L_x(\theta_1, \theta_2)$ is positive definite we have $\frac{\partial^2}{\partial \theta_1^2} \ln L_x < 0$ and $\det\left[\nabla^2 \ln L_x(\theta_1, \theta_2)\right] > 0$ evaluated at $(\hat{\theta}_1, \hat{\theta}_2)$. Or find the eigen-values which both has to be negative.

Note: If we have a diagonal matrix, which is 2x2, then to check whether it is negative definite, just see if the diagonal elements are negative as they are the eigenvalues.

E.g: Let us consider a random sample from a normal population, with mean μ and variance σ^2.

$$\therefore L(x, \mu, \sigma^2) = \prod_{i=1}^{n} f_{x_i}(x, \mu, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma(\sqrt{2\pi})} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

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

where $\sigma > 0$, $\infty < \mu < -\infty$

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

$\therefore$ The maximum likelihood estimation problem; ie MLE problem is

$$\max_{\substack{\sigma>0 \\ -\infty<\mu<+\infty}} \ln\left(L(x, \mu, \sigma^2)\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$\therefore$ ~~∂L~~ $\frac{\partial}{\partial\mu}\ln L(x, \mu, \sigma^2) = 0$ & $\frac{\partial}{\partial\sigma^2}\ln L(x, \mu, \sigma^2) = 0$

$\Downarrow$ $\Downarrow$

$\therefore \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$ & ~~$-\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum(x_i-\mu)^2 = 0$~~

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

$$\therefore \left.\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum x_i = \bar{x} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \end{aligned}\right\} \rightarrow \text{The MLE estimators?}$$

To conferm that these are MLE estimators we have to establish the fact that these are maximizers. The simplest way to realize this is that $\ln L(x, \mu, \sigma^2)$ is a ~~convex~~ concave function of (μ, σ^2) and hence the critical point is a global maximizer. But this is beyond the scope of this course. Hence we check the usual way stated in the previous page.

The Hessian matrix $\nabla^2 \ln L_1(x, \mu, \sigma^2)$ evaluated at $(\hat{\mu}, \hat{\sigma}^2)$ is given as

$$\nabla^2 \ln L_1(x, \hat{\mu}, \hat{\sigma}^2) = \begin{bmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{1}{\hat{\sigma}^4}\left[\frac{n}{2} - n^2\right] \end{bmatrix}$$

Thus $-\frac{n}{\hat{\sigma}^2} < 0$ & $\frac{1}{\hat{\sigma}^4}\left[\frac{n}{2} - n^2\right] < 0$, hence

$\nabla^2 \ln L_1(x, \hat{\mu}, \hat{\sigma}^2)$

to be negative definite. Thus $\hat{\mu}$ & $\hat{\sigma}^2$ are the maximizers of the mle problem.

Detail computation of $\nabla^2 \ln L_1(x, \hat{\mu}, \hat{\sigma}^2)$:

$$\text{We have } \frac{\partial}{\partial \mu} \ln L_1(x, \mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

$$\therefore \frac{\partial^2}{\partial \mu^2} \ln L_1(x, \hat{\mu}, \hat{\sigma}^2) = \frac{\sum n - n\mu}{\sigma^2} = -\frac{n}{\hat{\sigma}^2}$$

$$\frac{\partial}{\partial \sigma^2} \frac{\partial}{\partial \mu} \ln L_1(x, \hat{\mu}, \hat{\sigma}^2) = \frac{\partial}{\partial \sigma^2 \partial \mu} \ln L_1(x, \hat{\mu}, \hat{\sigma}^2)$$

$$= -\hat{\sigma}^{-4} \sum_{i=1}^{n} (x_i - \hat{\mu})$$

$$= -\hat{\sigma}^{-4} \left[\sum_{i=1}^{n} (x_i - \bar{x})\right]$$

$$= -\hat{\sigma}^{-4} \left(\sum_{i=1}^{n} x_i - n\bar{x}\right) = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} \frac{\partial \ln L}{\partial \sigma^2, \mu} = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\therefore \frac{\partial}{\partial \mu} \cdot \frac{\partial}{\partial \sigma^2} \ln L_1(x, \hat{\mu}, \hat{\sigma}^2) = \frac{1}{2\hat{\sigma}^4} 2 - \left(\sum_{i=1}^{n} (x_i - \lambda_i)\right) = 0$$

$$\frac{\partial}{\partial \sigma^2}\left(\frac{\partial}{\partial \sigma^2} \ln L_1(x, \hat{\mu}, \hat{\sigma}^2)\right) = \frac{n}{2}\hat{\sigma}^{-4} - n^2\hat{\sigma}^{-4} = \frac{1}{\hat{\sigma}^4}\left[\frac{n}{2} - n^2\right].$$

$$\left(\text{We use } \sum_{i=1}^{n} (x_i - \bar{x})^2 = n\hat{\sigma}^2\right)$$

We would like to note that second order condition only tells us that the ~~note~~ estimators that we have got are really only local minimizer albeit strict one. To establish that they are global we need to use ~~convexity~~ Concavity. The second order check also establishes the concavity of functions here.

A crash course on Concavity Let I be an interval in $\mathbb{R}$ and $f: I \to \mathbb{R}$.

The function $f: I \to \mathbb{R}$ is called concave if for any x, y in I & $\lambda \in [0,1]$

$$f(\lambda y + (1-\lambda)x) \geq \lambda f(y) + (1-\lambda) f(x). \longrightarrow (A)$$

Note: If x & y are in I; then for any $\lambda \in [0,1]$, $\lambda y + (1-\lambda)x \in I$

Thus if f is differentiable, then from (A), by Taylor's Theorem.

$$f(x) + \lambda f'(x)(y-x) + o(\lambda) \geq \lambda f(y) + (1-\lambda) f(x), \quad \text{for all } \lambda \in (0,1)$$

where $\frac{o(\lambda)}{\lambda} \to 0$ as $\lambda \to 0$.

$$\therefore \quad \lambda f'(x)(y-x) + o(\lambda) \geq \lambda (f(y) - f(x))$$

$$\therefore \text{ As } \lambda \to 0, \quad f'(x)(y-x) \geq f(y) - f(x)$$

If $f'(x) = 0 \Rightarrow f(x) \geq f(y) \Rightarrow x$ is a global maximum maximizer

Let $f: I \times I \to \mathbb{R}$, then f is concave if for any $(x_1, x_2) \in I \times I$ and $(y_1, y_2) \in I_1 \times I_2$ & $\lambda \in (0,1)$

$$f(\lambda y_1 + (1-\lambda) x_1, \ \lambda y_2 + (1-\lambda) x_2) \geq \lambda f(y_1, y_2) + (1-\lambda) f(x_1, x_2)$$

Again using Taylor's Theorem in 2-dimensions we have $\forall (x_1, x_2) \& (y_1, y_2)$

$$\frac{\partial f}{\partial x_1}(y_1 - x_1) + \frac{\partial f}{\partial x_2}(y_2 - x_2) \geq f(y_1, y_2) - f(x_1, x_2).$$

$$\therefore \frac{\partial f}{\partial x_1} = 0 \ \& \ \frac{\partial f}{\partial x_2} = 0, \quad f(x_1, x_2) \geq f(y_1, y_2), \quad \forall (y_1, y_2) \in I \times I$$

$\Rightarrow$ (x_1, x_2) is the global maximizer.

E.g: $-\alpha (x-\mu)^2 = f(x)$, $\alpha > 0$, $\mu \in \mathbb{R}$ is concave in x

$f(x) = -\ln x$, $x > 0$, i.e $x \in \mathbb{R}_{++}$ is concave in x.

~~If we fix x_0, then~~
f is concave on I if $f''(x) \leq 0$, $\forall x$ on I
f is concave on $I \times I$ if $\nabla^2 f$ is negative definite on $I \times I$

In our study in the example the log likelihood functions can be shown to be concave.

Section 3 Fisher Information & Cramer-Rao Inequality

The maximum likelihood estimation technique, is very popular, and the reason for this is follows. Given a sample observation,

$X_1 = x_1 \ldots\ldots X_n = x_n$ the likelihood function $L(x, \theta)$ provides us the "likelihood" or the frequency of the occurence of the observation $x_1, x_2, \ldots\ldots x_n$. Since the parameter θ is unknown the maximum Suppose we ask the question: What is the distribution for which the observation observations $x_1 \ldots x_n$ occur the maximum number of time? This can be done by finding the θ, which maximizes the joint pdf, since it is the parameter θ, which specifies the distribution We shall now see how the likelihood function can be used to build some important measures related to, estimators.

We shall begin with the notion of Fisher Information, but let us first write down the two basic assumptions we need. We will first study only the single parameter case.

Assumptions

A) The pdf $f_X(x, \theta)$ is twice contin differentiable as a function of the parameter θ.

B) We can differentiate ~~the~~ with respect to θ under the integral sign for the integral $\int f_X(x, \theta)\, d\theta$. In fact we will assume that we can differentiate twice under the integral sign with respect to θ.

C) The parameter θ belongs to an open interval.

The Fisher information function is built on the idea of a Fisher score function.

Let $X_1 \ldots X_n$ be a random sample from a population $f_X(x, \theta)$.

The Fischer score function for the ith observation is given as the r.v. $F(X_i, \theta)$ whose value for the observation $X = x_i$ is given as

$$F(x_i, \theta) = \frac{\partial}{\partial \theta} \ln f_X(x_i, \theta)$$

$$\therefore F(x_i, \theta) = \frac{1}{f_X(x_i, \theta)} \cdot \frac{\partial}{\partial \theta} f_X(x_i, \theta)$$

$$\therefore E\left[F(X_i, \theta)\right] = \int_{-\infty}^{\infty} \frac{1}{f_X(x_i, \theta)} \frac{\partial}{\partial \theta} f_X(x_i, \theta)\, f_X(x_i, \theta)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f_X(x_i, \theta)\, dx$$

$$= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f_X(x_i, \theta)\, dx = \frac{\partial}{\partial \theta}(1) = 0.$$

$$\therefore E\left[F(X_i, \theta)\right] = 0$$

The Total Fisher score function for a sample of size n is given as the r.v.

$$F_n(\theta) = \sum_{i=1}^{n} F(X_i, \theta)$$

[Note $E(F_n(\theta)) = 0$]

The Fisher information about the i-th observation is denoted as $I(\theta)$ and is given as

$$I(\theta) = \mathrm{Var}\left(F(X_i, \theta)\right)$$

$$\therefore \quad I_i(\theta) = \text{Var}\left(F(x_i,\theta)\right)$$

$$= E\left[\left(F(x_i,\theta)\right)^2\right] \qquad \left(\because E\left(F(x_i,\theta)\right) = 0\right)$$

$$= E\left[\left(\frac{\partial}{\partial\theta} \ln f(x_i,\theta)\right)^2\right]$$

$$= \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial\theta} \ln f(x_i,\theta)\right]^2 f_X(x,\theta)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{f(x,\theta)} \cdot \frac{1}{f_X^2(x_i,\theta)} \left[\frac{\partial}{\partial\theta} f_X(x_i,\theta)\right]^2 f_X(x,\theta)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{f_X(x_i,\theta)} \left[\frac{\partial}{\partial\theta} f_X(x_i,\theta)\right]^2 dx$$

Usually in text to one writes

$$\boxed{I_i(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \ln f_X(x_i,\theta)\right)^2\right]}$$

For a random sample of size n, the Fisher. information is given as.

E.g1: Consider a random sample ~~drawn from~~ of size n drawn from an exponential population; $f_X(x,\theta)$, given as

$$\Rightarrow f_X(x,\theta) = \theta e^{-\theta x}; \quad x>0, \ \theta>0$$

Thus

$$F(x_i,\theta) = \frac{\partial}{\partial\theta} \ln\left(\theta e^{-\theta x_i}\right)$$

$$= \frac{1}{\theta e^{-\theta x_i}} \left[\frac{\partial}{\partial\theta}\left(\theta e^{-\theta x_i}\right)\right]$$

$$= \frac{1}{\theta e^{-\theta x_i}} \left[e^{-\theta x_i} + \theta e^{-\theta x_i}(-x_i)\right]$$

$$\therefore \quad \bar{F}(x_i, \theta) = \frac{1}{\theta}\left[1 - \theta x_i\right]$$

$$= \frac{(1-\theta x_i)}{\theta}$$

$$\therefore \quad \bar{F}(X_i, \theta) = \left[\frac{1-\theta X_i}{\theta}\right]$$

$$\therefore \quad I(\theta) = E\left[\left(\bar{F}(x_i, \theta)\right)^2\right]$$

$$= \int_0^\infty \frac{(1-\theta x_i)^2}{\theta} \cdot \theta e^{-\theta x_i} dx$$

$$= \int_0^\infty \frac{(1-\theta x_i)^2}{\theta^2} \cdot \theta e^{-\theta x_i} dx$$

$$= \frac{1}{\theta}\int_0^\infty (1-\theta x_i)^2 e^{-\theta x_i} dx$$

$$= \frac{1}{\theta}\int_0^\infty (1 - 2\theta x_i + \theta^2 x_i^2) \cdot e^{-\theta x_i} dx$$

$$= \frac{1}{\theta}\int_0^\infty e^{-\theta x_i} dx - \frac{2}{\theta}\int_0^\infty x_i \cdot \theta e^{-\theta x_i} dx + \frac{1}{\theta}\int_0^\infty x_i^2 \theta e^{-\theta x_i} dx.$$

$$= \frac{1}{\theta^2}\int_0^\infty \theta e^{-\theta x_i} dx - \frac{2}{\theta} E(x_i) + \frac{1}{\theta} E(x_i^2)$$

$$= \frac{1}{\theta^2} - \frac{2}{\theta}\cdot\frac{1}{\theta} + \frac{1}{\theta}\left[\mathrm{Var}(x_i) + (E(x_i))^2\right]$$

$$= \frac{1}{\theta^2} - \frac{2}{\theta^2} + \frac{1}{\theta}\left[\frac{1}{\theta^2} + \frac{1}{\theta^2}\right]$$

$$= \frac{1}{\theta^2} - \frac{2}{\theta^2} + \frac{1}{\theta^2}\left[\frac{2}{\theta^2}\right] = -\frac{1}{\theta^2} + \frac{2}{\theta^4}$$

$$= \frac{2}{\theta^4} - \frac{1}{\theta^2} = \frac{1}{\theta^2}\left[\frac{2}{\theta^2} - 1\right]$$

$$= \frac{1}{\theta^2} > 0.$$

So the Fisher information for the whole sample is given as

$$I_n(\theta) = \text{Var}\left(F_n(\theta)\right)$$

$$= \text{Var}\left(F(x_1,\theta) + \cdots + F(x_n,\theta)\right)$$

Since $x_1 \ldots x_n$ are independent, $F(x_1,\theta), \ldots, F(x_n,\theta)$ are independent. Thus

$$I_n(\theta) = \text{Var}\left(F_{x_1}(x_1,\theta)\right) + \cdots + \text{Var}\left(F(x_n,\theta)\right)$$

$$= \sum I_{x_i}(\theta) \quad \text{(~~$I(x_i$~~)}$$

$$\therefore \quad I_n(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

But each x_i's are having the same distribution.

$$\therefore \quad \boxed{I_n(\theta) = n\, I_i(\theta)}$$

The Cramer-Rao Inequality

A random sampling is called regular if its Fischer information is continuous, strictly positive ~~and bounded~~ for all θ in the given range.

For example if we consider a random sample ~~from~~ $x_1 \ldots x_n$ from an exponential distribution, then

$$I_n(\theta) = n\, I_i(\theta)$$

$$= \frac{n}{\theta^2} > 0$$

$I_n(\theta)$ is of course continuous on $\theta > 0$. However, $I_n(\theta)$ is not bounded above, as a function of θ. But as $I_n(\theta)$ is continuous and positive we can consider the sampling from an exponential distribution is regular.

Theorem 11.1: Cramer-Rao Bound (Actually a lower bound)

Let $T = \hat{\theta}(X_1 \dots X_n)$ be a statistic which an estimator of θ of the population $f(x, \theta)$. Let the bias $b_n(\theta) = E[\hat{\theta}] - \theta$ be continuously differentiable. Then

$$\text{Var}(\hat{\theta}) \geq \frac{(1 + b_n'(\theta))^2}{I_n(\theta)}$$

Proof:

$$\theta + b_n(\theta) = E[\hat{\theta}] = \int_{\mathbb{R}^n} \hat{\theta}_n(x_1 \dots x_n) f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)\, dx_1 \dots dx_n$$

Now differentiation under the integral sign we have.

$$1 + b_n'(\theta) = \int_{\mathbb{R}^n} \hat{\theta}_n(x_1 \dots x_n) \frac{\partial f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)}{\partial \theta} dx_1 \dots dx_n$$

$$= \int_{\mathbb{R}^n} \hat{\theta}(x_1 \dots x_n) \frac{f_X(x_1 \dots x_n, \theta)}{f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)} \cdot \frac{\partial}{\partial \theta} f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)\, dx_1 \dots dx_n$$

$$= \int_{\mathbb{R}^n} \hat{\theta}(x_1 \dots x_n) F_n(\theta) f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)\, dx_1 \dots dx_n$$

$$F_n(\theta) = \sum_{i=1}^{n} F(x_i, \theta) = \sum \frac{1}{f_X(x_i, \theta)} \frac{\partial}{\partial \theta} f_X(x_i, \theta)$$

$$= \sum_{i=1}^{n} \frac{1}{f_{X_i}(x_i \theta)} \frac{\partial}{\partial \theta} f_{X_1 \dots X_n}(x_1 \dots x_n, \theta)$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f_X(x_i, \theta)$$

$$= \frac{\partial}{\partial \theta} \left[\sum_{i=1}^{n} \ln f_{X_i}(x_i, \theta) \right]$$

$$= \frac{\partial}{\partial \theta} \left[\ln \prod_{i=1}^{n} f_{X_i}(x_i, \theta) \right] = \frac{\partial}{\partial \theta} \ln f_{X_1 \dots X_n}(x_1 \dots x_n \theta)$$

(By independence).

$$\therefore \quad F_n(\theta) = \frac{1}{f_{X_1 X_2 \cdots X_n}(x_1 \ldots x_n, \theta)} \frac{\partial}{\partial \theta} f_{X_1, \ldots X_n}(x_1, \ldots x_n, \theta)$$

$$\therefore \quad 1 + b_n'(\theta) = E\left[\hat{\theta} F_n(\theta)\right] = \operatorname{cov}\left(\hat{\theta}, F_n(\theta)\right)$$

Note that $\operatorname{Cov}(\hat{\theta}, F_n(\theta)) = E\left[\hat{\theta} F_n(\theta)\right] - E[\hat{\theta}]\, E[F_n(\theta)]$

$= E\left[\hat{\theta} F_n(\theta)\right]$ $\quad (\because E[F_n(\theta)] = 0$ as derived earlier)

$\therefore$ Let ρ_n^2 = correlation coefficient of $\hat{\theta}$, ~~$F_n(\theta)$~~ and $F_n(\theta)$

$$\therefore \quad \rho_n^2 = \frac{\operatorname{Cov}\left(\hat{\theta}, F_n(\theta)\right)}{\operatorname{Var}(\hat{\theta}) \operatorname{Var}(F_n(\theta))}$$

But $\rho_n^2 \leq 1$ as $\rho_n \in [-1, +1]$

$$\Rightarrow \quad \frac{\operatorname{Cov}\left(\hat{\theta}, F_n(\theta)\right)}{\operatorname{Var}(\hat{\theta}) \operatorname{Var}(F_n(\theta))} \leq 1.$$

Hence

$$\operatorname{Var}(\hat{\theta}) \geq \frac{\operatorname{Cov}\left(\hat{\theta}, F_n(\theta)\right)}{\operatorname{Var}(F_n(\theta))}$$

$$\Rightarrow \quad \boxed{\operatorname{Var}(\hat{\theta}) = \frac{\left(1 + b_n'(\theta)\right)^2}{I_n(\theta)}} \quad \square$$

This is called the Cramer-Rao Lower bound. Now if $\hat{\theta}$ is an unbiased estimator, then $b_n(\theta) = 0$ for any θ in the given range. Thus we have

$$\boxed{\operatorname{Var}(\hat{\theta}) \geq \frac{1}{I_n(\theta)}}$$

This is called the Cramer-Rao Inequality. An unbiased estimator $\hat{\theta}$ is called efficient if

$$\boxed{\operatorname{Var}(\hat{\theta}) = \frac{1}{I_n(\theta)}}$$

Let us finish our discussion with an example by drawing a sample of size n, from a normal population $N(\mu, \sigma^2)$, where μ is unknown but σ^2 is ~~not~~ known.

We shall show that $\bar{X}_n = \bar{X} = \frac{X_1 + \cdots + X_n}{n}$ is indeed an efficient estimator of μ. $\bar{X}_n$ is known to be unbiased as we have already shown. Now $Var(\bar{X}_n) = \frac{\sigma^2}{n}$. We have

$$f_X(x, \mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

$$\therefore \quad \ln f_{X_i}(x_i, \mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2}$$

In the random-variable form we have

$$\ln f_{X_i}(X_i, \mu) = \frac{1}{2}\ln(2\pi\sigma^2) - \frac{(X_i-\mu)^2}{2\sigma^2}$$

$$F(X_i, \theta) = \frac{X - \theta\mu}{\sigma^2}$$

$$\therefore \quad I_n(\theta) = n I_i(\theta) = n\,Var(F(x_i, \mu)) = n\,E\left[(F(x_i, \mu))^2\right]$$

$$= n\,E\left[\left(F(x_i, \mu)\right)^2\right]$$

$$= n\,E\left[\frac{(x_i-\mu)^2}{\sigma^4}\right]$$

$$= \frac{n}{\sigma^4} E\left[(x_i-\mu)^2\right]$$

$$= \frac{n}{\sigma^4} \cdot Var(x_i)$$

$$= \frac{n}{\sigma^4} \cdot \sigma^2 = \frac{n}{\sigma^2}$$

Now $Var(\bar{X}_n) = \frac{\sigma^2}{n} = \frac{1}{I_n(\theta)}$, proving that $\bar{X}_n$ is indeed an efficient estimator.

The part on Fisher Information is partially based on Chapter 1, of the book titled: Mathematical Statistics: Asymptotic minimax theory. by A. Korostelev and O. Korostelova, AMS-2011.