Ch. 3 Estimation

(October 26, 2017)

1 The Nature of Statistical Inference

It is argued that it is important to develop a mathematical model purporting to provide a generalized description of the data generating process. A (univariate) probability model in the form of the parametric family of the density functions $\Phi = \{f(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and its various ramifications formulated in last chapter provides such a mathematical model. By postulating Φ as a probability model for the distribution of the observation of interested, we could go on to consider questions about the unknown parameters $\boldsymbol{\theta}$ (via *estimation* and *hypothesis tests*) as well as further observations from the probability model (*prediction*).

In the next section the important concept of a *sampling model* is introduced as a way to link the probability model postulated, say $\Phi = \{f(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, to the observed data $\boldsymbol{x} \equiv (x_1, ..., x_n)'$ available. The sampling model provided the second important ingredient needed to define a statistical model; the starting point of any "parametric" statistical inference. In short, a statistical model is defined as comprising

(a). A probability model $\Phi = \{f(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$;

(b). A sampling model $\mathbf{x} \equiv (X_1, ..., X_n)'$.

The concept of a statistical model provides the starting point of all forms of statistical inference to be considered in the sequel. To be more precise, the concept of a statistical model forms the basis of what is known as *parametric inference*. There is also a branch of statistical inference known as *non − parametric inference* where no Φ is assumed a priori.

1.1 The Sampling Model

A *sampling model* is introduced as a way to link the probability model postulated, say $\Phi = \{f(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and the observed data $\boldsymbol{x} \equiv (x_1, ..., x_n)'$ available. It is designed to model the relationship between them and refers to the way the observed data can be viewed in relation to Φ.

Definition. (Sample)
A sample (of size n) is defined to be a set of random variables $(X_1, X_2, ..., X_n)$ whose density functions coincides with the "true" density function $f(x;\boldsymbol{\theta}_0)$ as postulated by the probability model. ■

The samples are generally drawn in one of three settings:

(a). A *cross sectional sample* is a sample of a number of observational units all drawn at the same point in time.

(b). A *time series sample* is a set of observations drawn on the same observational unit at a number (usually evenly spaced) points in time.

(c). Many recently have been based on time-series cross sections, which generally consistent of the same cross section observed at several points in time. The term *panel data set* is usually fitting for this sort of study. □

Given that a sample is a set of r.v.'s related to Φ it must have a distribution which we call the distribution of the sample.

Definition. (Distribution of a Sample)
The *distribution of the sample* $\mathbf{x} \equiv (X_1, X_2, ..., X_n)'$, is defined to be the joint distribution of the r.v.'s $X_1, X_2, ..., X_n$ denoted by

$$f_{\mathbf{x}}(x_1, ..., x_n; \boldsymbol{\theta}^*) \equiv f(\boldsymbol{x}; \boldsymbol{\theta}^*). \quad ■$$

The distribution of the sample incorporates both forms of relevant information, the probability model as well as sample information. It must come as no surprise to learn that $f(\boldsymbol{x};\boldsymbol{\theta})$ plays a very important role in statistical inference. The form of $f(\boldsymbol{x};\boldsymbol{\theta})$ depends crucially on the nature of the sampling model and as well as Φ. The simplest but most widely used form of a sampling model is the one based on the idea of a random experiment $\mathcal{E}$ and is called a random sample.

Definition. (Random Sample)
A set of random variables $(X_1, X_2, ..., X_n)$ is called a random sample from $f(x;\boldsymbol{\theta})$ if the r.v.'s $X_1, X_2, ..., X_n$ are independently and identically distributed ($i.i.d$). In this case the distribution of the sample takes the form

$$f(x_1, ..., x_n; \boldsymbol{\theta}^*) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = (f(x; \boldsymbol{\theta}))^n$$

the first equality due to independence and the second due to the fact that the r.v. are identically distributed. ■

A less restrictive form of a sample model in which we call an independent sample, where the identically distributed condition in the random sample is relaxed.

Definition. (Independent Sample)
A set of random variables $(X_1, X_2, ..., X_n)$ is said to be an *independent sample* from $f(x_i;\boldsymbol{\theta}_i),\ i = 1, 2, ...n$, respectively, if the r.v.'s $X_1, X_2, ..., X_n$ are independent. In this case the distribution of the sample takes the form

$$f(x_1, ..., x_n; \boldsymbol{\theta}^*) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}_i).$$

Usually the density function $f(x_i;\boldsymbol{\theta}_i),\ i = 1, 2, ..., n$ belong to the same family but their numerical characteristics (moments, etc.) may differ. ■

If we relax the independence assumption as well we have what we can call a non-random sample.

Definition. (Non-Random Sample)
A set of random variables $(X_1, X_2, ..., X_n)$ is said to be a *non-random sample* from[1] $f(x_1, x_2, ..., x_n; \boldsymbol{\theta}^*)$ if the r.v.'s $X_1, X_2, ...X_n$ are non-*i.i.d.*. In this case the only decomposition of the distribution of the sample possible is

$$f(x_1, ..., x_n; \boldsymbol{\theta}^*) = \prod_{i=1}^{n} f(x_i|x_1, ..., x_{i-1}; \boldsymbol{\theta}_i)$$

given x_0, where $f(x_i|x_1, ..., x_{i-1}; \boldsymbol{\theta}_i),\ i = 1, 2, ...n$ represent the conditional distribution of X_i given $X_1, X_2, ..., X_{i-1}$. ■

In the context of statistical inferences need to postulate both probability as well as a sampling model and thus we define a statistical model as comprising both.

Definition. (Statistical Model):
A statistical model is defined as comprising
(a). A probability model $\Phi = \{f(x; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\}$; and
(b). A sampling model $\mathbf{x} \equiv (X_1, X_2, ..., X_n)'$. ■

It must be emphasized that the two important components of a statistical model, the probability and sampling models, are clearly inter-related. For example we cannot postulate the probability model $\Phi = \{f(x; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ if the sample $\mathbf{x}$ is non-random. This is because if the r.v.'s $X_1, X_2, ..., X_n$ are not independent the probability model must be defined in terms of their joint distribution,.i.e. $\Phi = \{f(x_1, x_2, ..., x_n; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ (for example, stock price). Moreover, in the case of an independent but not identically distributed sample we need to specify the individual density functions for each r.v. in the sample, i.e. $\Phi = \{f_k(x_k; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta, k = 1, 2, ..., n\}$. The most important implication of this relationship is that when the sampling model postulated is found to be inappropriate it means that the probability model has to be re-specified as well.

[1]Here, we may regard this set of random variables as a sample of size "one" from a multivariate point of view.

1.2 The Classical and Bayesian Approaches to Statistical Inference

The various approaches to statistical inference based on alternative interpretations of the notation of probability differ mainly in relation to *what consistitutes relevant information* for statistical inference and how it *should be processed*. In the case of the classical approach the relevant information comes in the form of a probability model $\Phi = \{f(\boldsymbol{x};\theta), \theta \in \boldsymbol{\Theta}\}$ and a sampling model $\mathbf{x} = (X_1, \cdots, X_n)'$, providing the link between Φ and the observed data $\boldsymbol{x} = (x_1, \cdots, x_n)'$. The observed data are in fact interpreted as a realization of the sampling model. This information is then processed via the distribution of the sample $f(x_1, \cdots, x_n; \boldsymbol{\theta}^*)$.

The "subjective" interpretation of probability, on the other hand, leads to a different approach to statistical inference. This is common known as the *Bayesian approach* because the discussion is based on revising prior beliefs about the unknown parameters θ in the light of the observed data using probability distribution $f(\theta)$; that is, θ is assumed to be a random variable. The revision to the prior $f(\theta)$ comes in the form of the *posterior distribution* $f(\theta|\mathbf{x})$ via Bayes's rule:

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)f(\theta)}{f(\mathbf{x})} \propto f(\mathbf{x}|\theta)f(\theta),$$

$f(\mathbf{x}|\theta)$ being the distribution of the sample and $f(\mathbf{x})$ being constraint for $\mathbf{X} = \mathbf{x}$.

1.3 An Overview of Statistical Inference

The statistical model in conjunction with the observed data enable us to consider the following questions:

(A). Are the observed data consistent with the postulated statistical model ? (*model misspecification*)

(B). Assuming that the postulated statistical model is consistent with the observed data, what can we infer about the unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$?

- **(a).** Can we decrease the uncertainty about $\boldsymbol{\theta}$ by reducing the parameters space from Θ to $\boldsymbol{\Theta}_0$ where $\boldsymbol{\Theta}_0$ is a subset of $\boldsymbol{\Theta}$. (*confidence interval estimation*)
- **(b).** Can we decrease the uncertainty about $\boldsymbol{\theta}$ by choosing a particular value in $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$, as providing the most representative value of $\boldsymbol{\theta}$? (*point estimation*)

(c). Can we consider the question that $\boldsymbol{\theta}$ belongs to some subset $\boldsymbol{\Theta}_0$ of $\boldsymbol{\Theta}$? (*hypothesis testing*)

(C). Assuming that a particular representative value $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ has been chosen what can we infer about further observations from the data generating process (DGP) as described by the postulated statistical model ? (*prediction*) □

2 Point Estimation

A (point) estimation refers to our attempt to give a numerical value to θ. Let $(\mathcal{S}, \mathcal{F}, \mathcal{P}(\cdot))$ be the probability space of reference with X a r.v. defined on this space. The following statistical model is postulated:

(a). $\Phi = \{f(x;\theta),\ \theta \in \Theta\}, \quad \Theta \subseteq \mathbb{R}$;[2]

(b). $\mathbf{x} \equiv (X_1, X_2, ..., X_n)'$ is a random sample from $f(x;\theta)$.

Definition. (Estimation):
Point estimation in the context of this statistical model takes the form of constructing a mapping

$$h(\cdot):\ \mathcal{X} \to \Theta,$$

where $\mathcal{X}$ is the observation space and $h(\cdot)$ is a Borel function. The composite function (a statistic) $\hat{\theta} \equiv h(\mathbf{x}):\ \mathcal{S} \to \Theta$ is called an estimator and its value $h(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}$ an estimate of θ. It is important to distinguish between the two because the former is a *random variable* and the latter is a real number. ■

Example.
Let $f(x;\theta) = [1/\sqrt{2\pi}]exp\{-\frac{1}{2}(x-\theta)^2\},\ \theta \in \mathbb{R}$, and $\mathbf{x} = (X_1, X_2, ..., X_n)'$ be a random sample from $f(x;\theta)$. Here, $\mathcal{X} = \mathbb{R}^n$ and the following functions define estimators of θ which mapping $\mathbb{R}^n \to \mathbb{R}$:

(a). $\hat{\theta}_1 = h_1(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} X_i$,

(b). $\hat{\theta}_2 = h_2(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} X_i,\ k = 1, 2, ..., n-1$;

(c). $\hat{\theta}_3 = h_3(\mathbf{x}) = \frac{1}{2}(X_1 + X_n)$. ■

It is obvious that we can construct infinitely many such estimators. However, constructing "good" estimators is not so obvious. It is clear we need some criteria to choose between these estimators. In other words, we need to formalize what we mean

[2] so here, we are concerned with single parameter's model.

by a "good" estimator.

2.1 Finite Sample Properties of a Good Estimator

2.1.1 Unbiasedness

An estimator is constructed with the sole aim of providing us the "most representative values" of θ in the parameter space Θ, based on the available information in the form of statistical model. Given that the estimator $\hat{\theta} = h(\mathbf{x})$ is a r.v. (being a Borel function a random vector $\mathbf{x}$) any information of what we mean by a "most representative values" must be in terms of the distribution of $\hat{\theta}$, say $f(\hat{\theta})$. The obvious property to require a "good" estimator $\hat{\theta}$ of θ to satisfy is that $f(\hat{\theta})$ *is centered around* θ.

Definition. (Unbiased Estimator):
An estimator $\hat{\theta}$ of θ is said to be an *unbiased estimator* of θ if

$$E(\hat{\theta}) = \int_{-\infty}^{\infty} \hat{\theta} f(\hat{\theta}) d\hat{\theta} = \theta.$$

That is, the distribution of $\hat{\theta}$ has mean equal to the unknown parameter to be estimated.[3] ■

It must be remembered that unbiasedness is a property based on the distribution of $\hat{\boldsymbol{\theta}}$. This distribution is often called sampling distribution of $\hat{\boldsymbol{\theta}}$ in order to distinguish it from any other distribution of function of r.v.'s.

[3]Note that an alternative, but equivalent, way to define $E(\hat{\theta})$ is

$$E(\hat{\theta}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{x}) f(\boldsymbol{x}; \theta) d\boldsymbol{x}$$

where $f(\boldsymbol{x}; \theta) = f(x_1, x_2, ..., x_n; \theta)$ is the distribution of the sample, $\mathbf{x}$.

2.1.2 Efficiency

Although unbiasedness seems at first sight to be a highly desirable property it turns out in most situations there are too many unbiased estimators for this property to be used as the sole criterion for judging estimators. The question which naturally arises is "How can we choose among unbiased estimators ?" Given that the variance is a measure of dispersion, intuition suggests that the estimator with the smallest variance is in a sense better because its distribution is more "concentrated" around θ.

Definition. (Efficiency)
An unbiased estimator $\hat{\theta}$ of θ is said to be *relatively more efficient* than some other unbiased estimator $\tilde{\theta}$ if

$$Var(\hat{\theta}) < Var(\tilde{\theta}).$$

In the case of biased estimators relative efficiency can be defined in terms of the *mean square error* (MSE) which takes the form

$$\begin{aligned} MSE(\hat{\theta}, \theta_0) = E(\hat{\theta} - \theta_0)^2 &= E(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta_0)^2 \\ &= Var(\hat{\theta}) + [Bias(\hat{\theta}, \theta_0)]^2, \end{aligned}$$

the cross-product term being zero. $Bias(\hat{\theta}, \theta_0) = E(\hat{\theta}) - \theta_0$ is the bias of $\hat{\theta}$ relative to the value θ_0. ■

Definition. (Inadmissible)
For any two estimator $\hat{\theta}$ and $\tilde{\theta}$ of θ if $MSE(\hat{\theta}, \theta) \leq MSE(\tilde{\theta}, \theta),\ \theta \in \Theta$ with strict inequality holding for some $\theta \in \Theta$, $\tilde{\theta}$ is said to be inadmissible. ■

Using the concept of relatively efficiency we can compare different estimators we happen to consider. This is, however, not very satisfactory since there might be much better estimators in terms of MSE which we know nothing about. In order to avoid choosing the better of two inefficient estimators we need some *absolute measure of efficiency*. Such a measure is provided by the *Cramer-Rao* lower bound.

Definition. (Cramer-Rao Bound)
The equation

$$CR(\theta) = \frac{\left[1 + \frac{dB(\theta)}{d\theta}\right]^2}{E\left[\left(\frac{\partial \log f(\boldsymbol{x};\theta)}{\partial\theta}\right)^2\right]}$$

is the *Cramer-Rao lowest variance bound*, where $f(\boldsymbol{x},\theta)$ is the distribution of the sample and $B(\theta)$ the bias. It can be shown that for any estimator $\hat{\theta}$ of θ

$$MSE(\hat{\theta},\theta) \geq CR(\theta)$$

under the following regularity conditions on Φ:
(a). The set $A = \{x : \ f(\boldsymbol{x};\theta) > 0\}$ does not depend on θ;
(b). For each $\theta \in \Theta$ the distribution $[\partial^i \log f(\boldsymbol{x};\theta)]/(\partial\theta^i),\ i = 1,2,3$ exist for all $\boldsymbol{x} \in \mathcal{X}$;
(c). $0 < E[(\partial/\partial\theta)\log f(\boldsymbol{x};\theta)]^2 < \infty$ for all $\theta \in \Theta$.

In the case of unbiased estimators the inequality takes the form

$$Var(\hat{\theta}) \geq \left[E\left(\frac{\partial \log f(\boldsymbol{x};\theta)}{\partial\theta}\right)^2\right]^{-1},$$

the inverse of the lower bound is called *Fisher's information number* and is denoted by $I_n(\theta)$.[4]

Proof. (for the case that θ is 1×1)
Given that $f(x_1, x_2, ..., x_n;\theta)$ is the joint density function of the sample, it possesses the property that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, ..., x_n;\theta) dx_1 ... dx_n = 1,$$

or, more compactly,

$$\int_{-\infty}^{\infty} f(\boldsymbol{x};\theta) d\boldsymbol{x} = 1.$$

If we assume that the domain of $\mathbf{x}$ is independent of θ (from assumption (a), this permits straight-forward differentiation inside the integral sign) and that the derivative $\partial f(\cdot)/\partial\theta$ exists. Then differentiating the above equation with respect to $\boldsymbol{\theta}$ results in

$$\int_{-\infty}^{\infty} \frac{\partial f(\boldsymbol{x};\theta)}{\partial\theta} d\boldsymbol{x} = 0. \tag{3-1}$$

[4]It must bear in mind that the information matrix is a function of sample size n.

This equation can be reexpressed as

$$\int_{-\infty}^{\infty} \frac{\partial \ln f(\boldsymbol{x};\theta)}{\partial \theta} f(\boldsymbol{x};\theta) d\boldsymbol{x} = 0 \qquad \left(\because \frac{d}{dt} \ln f(t) = \frac{f'(t)}{f(t)} \right).$$

Therefore, it simply states that

$$E\left[\frac{\partial \ln f(\boldsymbol{x};\theta)}{\partial \theta} \right] = 0,$$

i.e., the expectation of the derivative of the natural logarithm of the likelihood function of a random sample from a regular density is zero.

Likewise, differentiating (3-1) w.r.t. θ again provides

$$\int_{-\infty}^{\infty} \frac{\partial^2 \ln f(\boldsymbol{x};\theta)}{\partial \theta^2} f(\boldsymbol{x};\theta) d\boldsymbol{x} + \int_{-\infty}^{\infty} \frac{\partial \ln f(\boldsymbol{x};\theta)}{\partial \theta} \frac{\partial f(\boldsymbol{x};\theta)}{\partial \theta} d\boldsymbol{x}$$
$$= \int_{-\infty}^{\infty} \frac{\partial^2 \ln f(\boldsymbol{x};\theta)}{\partial \theta^2} f(\boldsymbol{x};\theta) d\boldsymbol{x} + \int_{-\infty}^{\infty} \left[\frac{\partial \ln f(\boldsymbol{x};\theta)}{\partial \theta} \right]^2 f(\boldsymbol{x};\theta) d\boldsymbol{x}.$$

That is

$$Var\left[\frac{\partial \ln f(\boldsymbol{x};\theta)}{\partial \theta} \right] = -E\left[\frac{\partial^2 \ln f(\boldsymbol{x};\theta)}{\partial \theta^2} \right]. \tag{3-2}$$

Now consider the estimator $\hat{\theta} = h(\mathbf{x})$ of θ whose expectation is

$$E(h(\mathbf{x})) = \int h(\mathbf{x}) f(\boldsymbol{x};\theta) d\boldsymbol{x}. \tag{3-3}$$

Differentiating (3-3) w.r.t. θ we obtain

$$\begin{aligned} \frac{\partial E(h(\mathbf{x}))}{\partial \theta} &= \int h(\mathbf{x}) \frac{\partial f(\boldsymbol{x};\theta)}{\partial \theta} d\boldsymbol{x} \\ &= \int h(\mathbf{x}) \frac{\partial \ln f(\boldsymbol{x};\theta)}{\partial \theta} f(\boldsymbol{x};\theta) d\boldsymbol{x} \\ &= cov\left[h(\mathbf{x}), \frac{\partial \ln f(\boldsymbol{x};\theta)}{\partial \theta} \right]. \quad \left(\because E\left[\frac{\partial \ln f(\boldsymbol{x};\theta)}{\partial \theta} \right] = 0 \right). \end{aligned}$$

Since the square of the covariance is less than or equal to the product,[5] we have

$$cov\left[h(\mathbf{x}), \frac{\partial \ln f(\boldsymbol{x};\theta)}{\partial \theta} \right]^2 = \left[\frac{\partial E(h(\mathbf{x}))}{\partial \theta} \right]^2 \leq Var(h(\mathbf{x})) \cdot Var\left[\frac{\partial \ln f(\boldsymbol{x};\theta)}{\partial \theta} \right]. \tag{3-4}$$

In light of (3-2), we have

$$Var(h(\mathbf{x})) \geq \frac{\left[\frac{\partial E(h(\mathbf{x}))}{\partial \theta} \right]^2}{-E\left[\frac{\partial^2 \ln f(\boldsymbol{x};\theta)}{\partial \theta^2} \right]} = \frac{\left[1 + \frac{dB(\theta)}{d\theta} \right]^2}{E\left[\left(\frac{\partial \log f(\boldsymbol{x};\theta)}{\partial \theta} \right)^2 \right]}.$$

[5] Note that $[Cov(X,Y)]^2 = \rho^2 \sigma_X^2 \sigma_Y^2$ and $\rho^2 \leq 1$.

If the estimator is unbiased, $E(h(\mathbf{x})) = \theta$ and

$$Var(h(\mathbf{x})) \geq \frac{1}{E\left[\left(\frac{\partial \log f(\boldsymbol{x};\theta)}{\partial \theta}\right)^2\right]}.$$ ■

The above result can be extend to a multivariate parameters case.

Definition. (Multi-Parameters's Cramér-Rao Theorem)
An unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is said to be fully efficient if

$$\begin{aligned} Var(\hat{\boldsymbol{\theta}}) &= E\left[\left(\frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right]^{-1} \\ &= E\left[-\frac{\partial^2 \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right]^{-1} \\ &= (I_n(\boldsymbol{\theta}))^{-1}, \end{aligned}$$

where

$$\begin{aligned} I_n(\boldsymbol{\theta}) &= E\left[\left(\frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right] \\ &= E\left[-\frac{\partial^2 \log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right], \end{aligned}$$

is called the *sample information matrix*. ■

Example.
For a random sample of size n from a (univariate) normal distribution, the Cramer-Rao variance lower bound for unbiased estimator $\boldsymbol{\theta} = (\mu, \sigma^2)'$ is derived as following.

$$\begin{aligned} f(x_1, x_2, ..., x_n; \boldsymbol{\theta}) = f(\boldsymbol{x}, \boldsymbol{\theta}) &= \prod_{i=1}^{n}\left\{(2\pi)^{-1/2}(\sigma^2)^{-1/2}\exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]\right\} \\ &= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]. \end{aligned}$$

The sample log-likelihood function therefore is

$$\ln f(x_1, x_2, ..., x_n; \boldsymbol{\theta}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

The first and second derivatives of log likelihood w.r.t. the parameters is

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu),\\
\frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2,\\
\frac{\partial^2 \ln L}{\partial \mu^2} &= -\frac{n}{\sigma^2},\\
\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2,\\
\frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} &= -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu).
\end{aligned}$$

The sample information matrix therefore is

$$\begin{aligned}
I_n(\boldsymbol{\theta}) &= E\left[\left(\frac{\partial \ln f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right]\\
&= E\left[-\frac{\partial^2 \ln f(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right]\\
&= -E\begin{bmatrix} \frac{\partial^2 \ln L}{\partial \mu^2} & \frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} & \frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} \end{bmatrix}\\
&= \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}. \quad (how?)
\end{aligned}$$

The Cramer-Rao low variance Bound therefore is

$$I_n^{-1}(\mu,\sigma) = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}. \qquad \blacksquare$$

2.1.3 Sufficiency

Efficiency can be seen as a property indicating that the estimator "utilizes" all the information contained in the statistical model. An important concept related to the information of a statistical model is the concept of a *sufficient statistic* introduced by Fisher (1922) as a way to reduce the sampling information by discarding only the information of no relevance to any inference about θ. In other words, a statistic $\tau(\mathbf{x})$ is said to be sufficient for θ if it makes no difference whether we use $\mathbf{x}$ or $\tau(\mathbf{x})$ in inference

concerning θ. Obviously in such a case we would prefer to work with $\tau(\mathbf{x})$ instead of $\mathbf{x}$, the former being of lower dimensionality.

Definition. (Sufficient Statistic)
A statistic $\tau(\cdot): \mathcal{X} \to \mathbb{R}^m$, $n > m$ (n is the size of a sample), is called *sufficient for* $\boldsymbol{\theta}$ if the conditional distribution $f(\boldsymbol{x}|\tau(\boldsymbol{x}) = \tau)$ is independent of $\boldsymbol{\theta}$, i.e. $\boldsymbol{\theta}$ does not appear in $f(\boldsymbol{x}|\tau(\boldsymbol{x}) = \tau)$ and the domain of $f(\cdot)$ does not involve $\boldsymbol{\theta}$. ■

That is, knowing that particular outcomes is, adds nothing to our information about $\boldsymbol{\theta}$ that we do not already know, having been told that $\tau(\boldsymbol{x}) = \tau$. Verifying this directly by deriving $f(\boldsymbol{x}|\tau(\boldsymbol{x}) = \tau)$ and showing that is independent of $\boldsymbol{\theta}$ can be a very difficult exercise. One indirect way of verifying sufficiency is provided by the following Theorem.

Theorem. (Fisher-Neyman Factorization)
The statistics $\tau(\mathbf{x})$ is sufficient for $\boldsymbol{\theta}$ if and only if there exists a factorization of the form

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = f(\tau(\boldsymbol{x}); \boldsymbol{\theta}) \cdot s(\mathbf{x}),$$

where $f(\tau(\boldsymbol{x}); \boldsymbol{\theta})$ is the density function of $\tau(\mathbf{x})$ and depends on $\boldsymbol{\theta}$ and $s(\mathbf{x})$, some functions of $\mathbf{x}$ independent of $\boldsymbol{\theta}$. ■

Even this result, however, is of no great help because we have to have the statistics $\tau(\mathbf{x})$ as well as its distribution to begin with. Lehmann and Scheffe (1950) provides us with a very convenient way to derive *minimal sufficient statistics*.

Intuition suggests that, since efficiency is related to *full utilization*[6] of the information in the statistical model, and sufficiency can be seen as a *maximal reduction of such information without losing any relevant information* as far as inference about $\boldsymbol{\theta}$ is concerned, there must be a direct relationship between the two properties. A relationship along the lines that when an efficient estimator is needed we should look no further than the sufficient statistics, is provided by the following Theorem. It states that a best unbiased estimator (if it exists) must be the function of sufficient statistic.

[6]Think of using all the n sample is relatively efficient than using less of them in a sample of size n.

Theorem. (Rao-Blackwell Theorem):
Let $\tau(\mathbf{x})$ be a sufficient statistic for $\boldsymbol{\theta}$ and $\kappa(\mathbf{x})$ be an estimator of $\boldsymbol{\theta}$, then

$$E(h(\tau)-\theta)^2 \leq E(\kappa(\mathbf{x})-\theta)^2, \quad \boldsymbol{\theta} \in \Theta,$$

where $h(\tau) = E(\kappa(\mathbf{x})|\tau(\mathbf{x}) = \tau)$, i.e. the conditional expectation of $\kappa(\mathbf{x})$ given $\tau(\mathbf{x}) = \tau$. ■

From the above discussion of the properties of unbiasedness, relative and fully efficiency and sufficiency we can see that these properties are directly related to the distribution of the estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. As argued repeats, deriving the distribution of Borel functions of r.v.'s such as $\hat{\boldsymbol{\theta}} = h(\mathbf{x})$ is a very difficult exercise and very few results are available in the literature. These existing results are mainly related to simple functions of normally distributed r.v.'s . For the cases where no such results are available (which is the rule rather than the exception) we have to resort to *asymptotic results*. This implies that we need to extend the above list of criteria for 'good' estimators to include *asymptotic properties of estimators*. These asymptotic properties will refer to the behavior of $\hat{\boldsymbol{\theta}}_n$ as $n \to \infty$. In order to emphasis the distinction between these asymptotic properties and the properties considered so far we will call the latter *finite sample properties*. The finite sample properties are related directly to the distribution of $\hat{\boldsymbol{\theta}}_n$, say $f(\hat{\boldsymbol{\theta}}_n)$. On the other hand, the asymptotic properties are related to the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$.

2.2 Asymptotic Properties

2.2.1 Consistency

A natural property to require estimators to have is that as $n \to \infty$ the probability of $\hat{\theta}$ being close to the true θ should increase as well. We formalize this idea using the concept of convergence in probability associated with the weak law of large numbers.

Definition. (Convergence in Probability)
An estimator $\hat{\theta}_n = h(\mathbf{x})$ is said to be (weak) consistent for θ if

$$lim_{n\to\infty}\, Pr(|\hat{\theta}_n - \theta| < \varepsilon) = 1,$$

we write $\hat{\theta}_n \xrightarrow{p} \theta$. Vectors and matrices are said to converge in probability provided each element converges in probability. ■

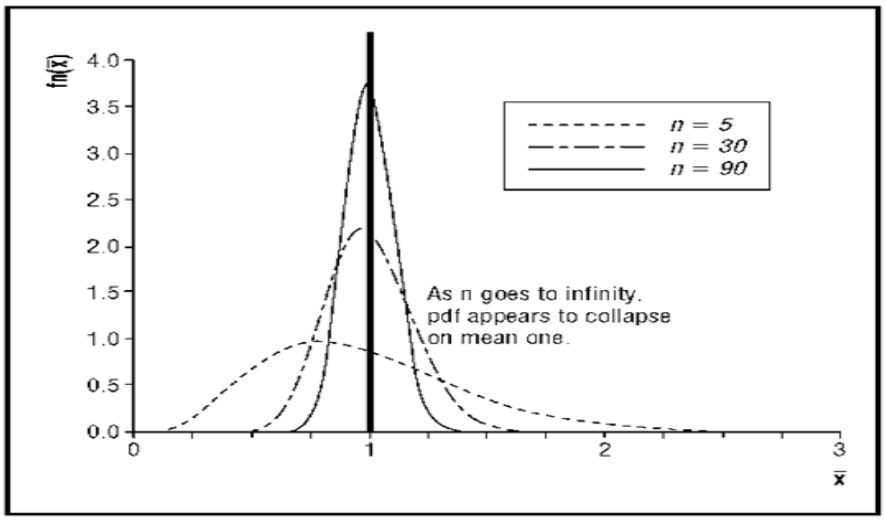

Figure (3-1). $\bar{X}$ is Consistent Estimator of μ.

Theorem.
Given $g : \mathbb{R}^k \to \mathbb{R}^l$ and any sequence random $k \times 1$ vector $\boldsymbol{\theta}_n$ such that $\boldsymbol{\theta}_n \xrightarrow{p} \boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is $k \times 1$, if g is continuous at $\boldsymbol{\theta}$, then $g(\boldsymbol{\theta}_n) \xrightarrow{p} g(\boldsymbol{\theta})$. ■

Another convergence concept often encountered in the context of time series data is that of convergence in the mean square error.

Definition. (Convergence in Mean Square Error)
If $\hat{\theta}_n$ is an estimator of θ which satisfies the following properties:
(a) $lim_{n\to\infty}\, E(\hat{\theta}_n) = \theta$;
(b) $lim_{n\to\infty}\, Var(\hat{\theta}_n) = 0$,

then $\hat{\theta}_n$ *converges in mean square to* θ, written $\hat{\theta}_n \xrightarrow{m.s.} \theta$. ■

A sufficient condition for weakly consistent of $\hat{\theta}_n$ is formalized in the following results.

Theorem. (Convergence in Mean Square Implies Convergence in Probability)

$$\text{If } \hat{\theta}_n \xrightarrow{m.s.} \theta, \text{ then } \hat{\theta}_n \xrightarrow{p} \theta.$$

Proof.
Using Chebyshev inequality,

$$Pr\left(|\hat{\theta}_n - E(\hat{\theta}_n)| \geq \varepsilon\right) \leq \frac{Var(\hat{\theta}_n)}{\varepsilon^2}, \ i.e.$$
$$Pr\left(|\hat{\theta}_n - E(\hat{\theta}_n)| < \varepsilon\right) \geq 1 - \frac{Var(\hat{\theta}_n)}{\varepsilon^2}.$$

Hence, if as $n \to \infty$ such that $lim_{n\to\infty}\, E(\hat{\theta}_n) = \theta$ and $lim_{n\to\infty}\, Var(\hat{\theta}_n) = 0$, then

$$Pr\left(|\hat{\theta}_n - \theta| < \varepsilon\right) = 1,$$

the result is immediate obtained. ■

It is important to note that these are only sufficient conditions for consistency (not necessary); that is, consistency is not equivalent to the above conditions, since for consistency $Var(\hat{\theta}_n)$ need not even exist.

A stronger form of consistency associated with almost sure convergence is a very desirable asymptotic property for estimators.

Definition. (Almost Sure Convergence)
An estimator $\hat{\theta}_n$ is said to be a strongly consistent estimator of θ if

$$Pr\left(lim_{n\to\infty}\hat{\theta}_n = \theta\right) = 1,$$

and is denoted by $\hat{\theta}_n \xrightarrow{a.s.} \theta$. ■

Without further conditions, no necessary relationship holds between convergence in mean-square error and almost sure convergence.

Theorem. (Laws of Large Numbers)
Given restrictions on the dependence, heterogeneity, and moments of a sequence of random variable $\{Z_t\}$,

$$\bar{Z}_n - \bar{\mu}_n \xrightarrow{a.s.} 0,$$

where $\bar{Z}_n = n^{-1}\sum_{t=1}^{n} Z_t$ and $\bar{\mu}_n = E(\bar{Z}_n)$. ■

Example.
Let $\{Z_t\}$ be a sequence of *i.i.d.* random variables with mean μ and variance σ^2 and with finite fourth moment. Then

$$\begin{aligned} n^{-1}\left(\sum_{t=1}^{n} Z_t\right) - n^{-1}E\left(\sum_{t=1}^{n} Z_t\right) &\equiv \bar{Z}_n - \mu \xrightarrow{a.s.} 0, \ i.e. \\ \bar{Z}_n &\xrightarrow{a.s.} \mu. \end{aligned}$$

■

Theorem.
Given $g : \mathbb{R}^k \to \mathbb{R}^l$ and any sequence random $k \times 1$ vector $\boldsymbol{\theta}_n$ such that $\boldsymbol{\theta}_n \xrightarrow{a.s.} \boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is $k \times 1$, if g is continuous at $\boldsymbol{\theta}$, then $g(\boldsymbol{\theta}_n) \xrightarrow{a.s.} g(\boldsymbol{\theta})$. ■

2.2.2 Asymptotic Normality

A second asymptotic property of a estimator is its *convergence in distribution*.

Definition. (Convergence in Distribution)
Let X_n be a sequence of random variables indexed by the sample size, and assume that X_n has cdf $F_n(x)$. If $lim_{n\to\infty}|F_n(x) - F(x)| = 0$ at all continuity point of $F(x)$, where $F(x)$ is the distribution function of a random variable X, then X_n *converges in distribution* to the random variable X, denoted as $X_n \xrightarrow{d} X$ and $F(x)$ is called the

limiting distribution of X_n. ■

Extending the limit theorem of distribution to $\hat{\theta}_n$ leads to the property of asymptotic normality, i.e. central limit theorem.

Definition. (Asymptotic Normality)
An estimator $\hat{\theta}_n$ is said to be asymptotic normal if two sequence $\{V_n(\theta),\ n \geq 1\}$ and $\{\hat{\theta}_n,\ n \geq 1\}$ exist such that

$$(V_n(\theta))^{-\frac{1}{2}}(\hat{\theta}_n - \theta) \xrightarrow{d} Z \sim N(0,1),$$

where $V_n(\theta) = \lim_{n\to\infty} Var(\hat{\theta}_n)$. ■

In most cases of interest in practice such as the case of a random sample, $V_n(\theta)$ is of order $1/n$, denoted by $V_n(\theta) = O(1/n)$. In such a case asymptotic normality can be written in the following form:

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \sim N(0, V(\theta)),$$

where $V(\theta) = n \cdot V_n(\theta)$ represent the *asymptotic variance* of $\hat{\theta}$.

Definition. (Asymptotic Unbiasedness)
An estimator $\hat{\theta}_n$ with $V_n(\hat{\theta}) = O(1/n)$ is said to be *asymptotically unbiased* if

$$E\left[\sqrt{n}(\hat{\theta}_n - \theta)\right] = 0 \ \ as\ n \to \infty.$$ ■

It must be emphasized that asymptotic unbiasedness is a stronger condition than $lim_{n\to\infty}E(\hat{\theta}_n) = \theta$; the former specifying the rate of convergence.

Example.
Let $X_1, X_2, ..., X_n$ be a *i.i.d.* sample with mean μ and variance σ^2. Denote $\hat{\theta}_n = \bar{X}_n = \sum_{i=1}^n X_i$. Then $Var(\bar{X}_n) = Var(\sum X_i/n) = \frac{1}{n}\frac{n\cdot\sigma^2}{n} = \frac{1}{n}\cdot\sigma^2 = O(1/n)$. In this case,

$$(V_n(\theta))^{-\frac{1}{2}}(\hat{\theta}_n - \theta) \equiv \left(\frac{1}{n}\cdot\sigma^2\right)^{-\frac{1}{2}}(\bar{X}_n - \mu) \equiv \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1),$$

is the familiar central limit theorem. We may also write

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2),$$

where we read $\sigma^2 = V(\theta) = n \cdot V_n(\theta)$, which is the asymptotic variance. Here, $V_n(\theta) = \lim_{n\to\infty} Var(\bar{X}_n)$. ■

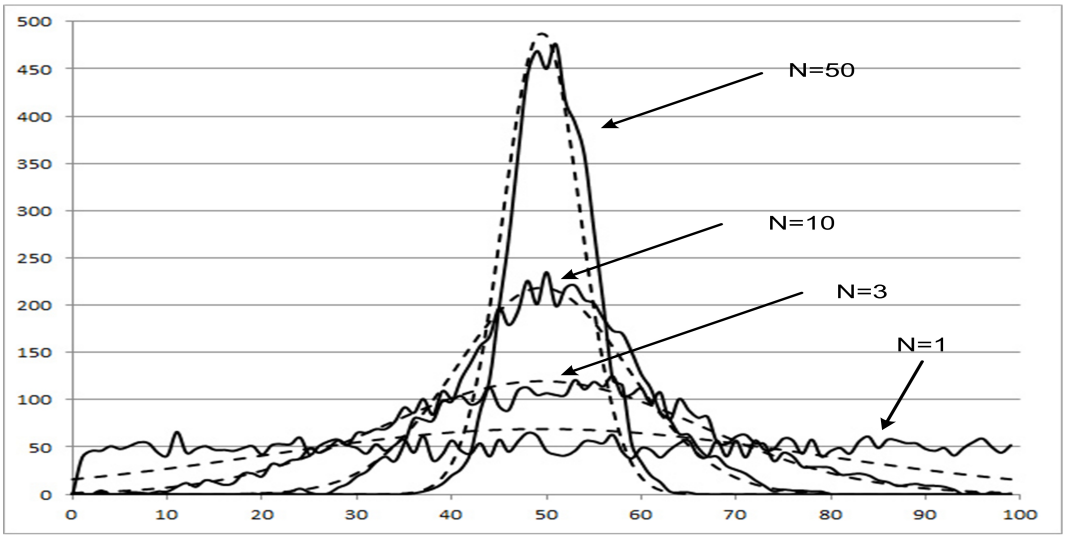

Figure (3-2). The Sampling Distribution becomes Normal Regardless of Shape of Population.

We now derive the most important result in limiting distribution theory–the central limiting theorem in a random sample case.

Theorem. (Lindeberg-Lévy)
Let X_t, $t = 1, 2, .., n$ be a sequence of *i.i.d.* random scalars, with $E(X_t) = \mu$ and $Var(X_t) = \sigma^2 < \infty$. Then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0, 1).$$

Proof.
Let $f(\lambda)$ be the characteristic function of $X_t - \mu$ and let $f_n(\lambda)$ be the characteristic

function of $n^{-1/2}\sum_{t=1}^{n}(\mathcal{Z}_t-\mu)/\sigma = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$. From equation (2-4) of Ch.2 we have

$$f_n(\lambda) = [f(\lambda/(\sigma\sqrt{n}))]^n$$

or

$$\log f_n(\lambda) = n\log f(\lambda/(\sigma\sqrt{n})).$$

Taking a Taylor expansion of $f(\lambda)$ around $\lambda = 0$ gives[7]

$$\begin{aligned} f(\lambda) &= f(0) + f'(0)\lambda + \frac{f''(0)\lambda^2}{2!} + o(\lambda^2) \\ &= 1 - \sigma^2\lambda^2/2 + o(\lambda^2). \end{aligned}$$

Hence[8]

$$\begin{aligned} \lim_{n\to\infty}\log f_n(\lambda) &= \lim_{n\to\infty}\{n\cdot\log[1-\lambda^2/(2n)+o(\lambda^2/n)]\} \\ &= \lim_{n\to\infty} n\{-\lambda^2/(2n)+o(\lambda^2/n)+o[-\lambda^2/(2n)+o(\lambda^2/n)]\} \\ &= -\lambda^2/2 + \lim_{n\to\infty}\{n\cdot o(\lambda^2/n) + n\cdot o[-\lambda^2/(2n)+o(\lambda^2/n)]\} \\ &= -\lambda^2/2. \end{aligned}$$

Hence $f_n(\lambda)\to\exp(-\lambda^2/2)$, which is the characteristic function of standard normal distribution. It follows from the uniqueness theorem that $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\xrightarrow{d} Z\sim N(0,1)$. ■

Finally, in relation to the variance of the asymptotic normal distribution we can define the concept of asymptotic efficiency in a multivariate case.

Definition (Asymptotic Efficiency)
An asymptotically normal estimator $\hat{\boldsymbol{\theta}}_n$ is said to be *asymptotically efficient* if $V_n(\boldsymbol{\theta}) = (I_n(\boldsymbol{\theta}))^{-1}$, where

$$\begin{aligned} I_n(\boldsymbol{\theta}) &= E\left[\left(\frac{\partial\log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial\log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)'\right] \\ &= E\left[-\frac{\partial^2\log f(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right], \end{aligned}$$

i.e. the asymptotic variance achieve the limit of the Cramer-Rao lower bound. See Chapter 4 for a detailed coverage of asymptotic theorems.

[7] To see this result, recall from Taylor's theorem that $f(\lambda) = f(0) + f'(0)\lambda + \frac{f''(0+\xi)\lambda^2}{2!}$, $0<\xi<\lambda$. If we assume that $f''(\lambda)$ is continuous at 0, then $f''(0+\xi) = f''(0) + o(1)$, where $o(1)\to 0$ as $\xi\to 0$. We can therefore write $f(\lambda) = f(0) + f'(0)\lambda + \frac{f''(0)\lambda^2}{2!} + o(\lambda^2)$.

[8] Similarly, by Taylor expansion $g(x)$ around $x = 0$ we have $g(x) = g(0) + xg'(0) + o(x)$. So, $\ln(1+x) = \ln(1) + x + o(x) = x + o(x)$.

3 Methods of Estimation

The purpose of this section is to consider various methods for constructing "good estimators" for the unknown parameters $\boldsymbol{\theta}$. These methods to be discussed are the *least-square method*, *the method of moment* and the *maximum likelihood estimation method.*

3.1 The Method of Least-Squares

The method of least-square was first introduced by Legendre in 1805 and Gauss in 1809 in the context of astronomical measurement. The problem as posed at the time was one of approximating a set of noisy observations $y_i,\ i = 1, 2, ..., n$, with some known functions $g_i(\theta_1, \theta_2, ..., \theta_m),\ i = 1, 2, ..., n$, which depended on the unknown parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)',\ m < n$. Legendre suggest minimizing the squared errors

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - g_i(\boldsymbol{\theta}))^2 \quad the\ least - squares.$$

Assuming differentiability of $g_i(\boldsymbol{\theta}),\ i = 1, 2, ..., n,\ [\partial l(\boldsymbol{\theta})]/\partial\boldsymbol{\theta} = 0$ gives rise to the so called *normal equation* of the form

$$(-2)\sum_{i=1}^{n}[y_i - g_i(\boldsymbol{\theta})]\frac{\partial}{\partial\theta_k}g_i(\boldsymbol{\theta}) = 0, \quad k = 1, 2, ..., m.$$

In this form the least-square method has nothing to do with the statistical model developed above, it is merely an interpolation method in approximation theory.

Gauss extended the least square method to statistical method with the probabilistic structure attached to the error term. He posed the problem in the form

$$\begin{aligned} y_i &= g_i(\boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, 2, .., n, \\ \varepsilon_i &\overset{i.i.d.}{\sim} (0, \sigma^2), \quad i = 1, 2, ..., n. \end{aligned}$$

In this form the problem can be viewed as one of estimation in the context of statistical model:

$$\Phi = \left\{ f(y_i; \boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{(2\pi)}}\exp\left(-\frac{1}{2\sigma^2}(y_i - g_i(\boldsymbol{\theta}))^2\right), \boldsymbol{\theta} \in \Theta \right\}$$

by transferring the probabilistic assumption from ε_i to y_i, the observable r.v., and consider $\mathbf{y} = (Y_1, Y_2, ..., Y_n)'$ be an independent sample from $f(y_i; \boldsymbol{\theta}), i = 1, 2, ..., n$. Gauss went on to derive what we have called the distribution of the sample

$$f(\mathbf{y}; \boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - g_i(\boldsymbol{\theta}))^2\right),$$

and suggested that maximization of $f(\mathbf{y}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ gives rise to the same estimator of $\boldsymbol{\theta}$ as minimization the square errors

$$\sum_{i=1}^{n}(y_i - g_i(\boldsymbol{\theta}))^2.$$

As we will see below, the above maximization can be seen as a forerunner of the maximum likelihood method.

3.2 The Method of Moments

In the context of a probability model, Φ, however unknown parameters of interest are not only associated with the mean but also with the higher moments. This prompted Pearson in 1894 to suggest the *method of moment* as a general estimation method.

Let us assume that $\mathbf{x} = (X_1, X_2, ..., X_n)'$ is a random sample from $f(x; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^k$. The raw moments of $f(x; \boldsymbol{\theta})$, $\mu'_r \equiv E(x^r)$, $r \geq 1$, are by definition functions of the unknown parameters, since

$$\mu'_r(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} x^r f(x; \boldsymbol{\theta}) dx, \quad r \geq 1.$$

In order to apply the method we need to express the unknown parameter $\boldsymbol{\theta}$ in the form

$$\theta_i = g_i(\mu'_1, \mu'_2, ..., \mu'_k), \quad i = 1, 2, ..., k,$$

where g_is are continuous functions. The method of moments based on the substitutional idea, proposes estimating θ_i using

$$\hat{\theta}_i = g_i(m_1, m_2, ..., m_k), \quad i = 1, 2, ..., k,$$

as the estimator of θ_i, $i = 1, 2, ..., k$, where $m_r = (1/n)\sum_{i=1}^{n} X_i^r$, $r \geq 1$ represent the sample raw moments. The justification of the method is based on the fact that if $\mu'_1, \mu'_2, ..., \mu'_k$ are one-to one function of $\boldsymbol{\theta}$ then since (by laws of large numbers)

$$m_r \xrightarrow{a.s.} \mu'_r, \quad r \geq 1,$$

it follows that (by Slutzky Theorem)

$$\hat{\theta}_i \xrightarrow{a.s.} \theta_i, \quad i = 1, 2, ..., k.$$

Example.
Let $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, ..., n$, then $\mu_1' = \mu$, $\mu_2' = \sigma^2 + \mu^2$ and $m_1 = (1/n)\sum_{i=1}^n X_i$, $m_2 = (1/n)\sum_{i=1}^n X_i^2$. The method suggests

$$\begin{aligned} \hat{\mu} &= m_1 = \bar{X}_n, \\ \hat{\sigma}^2 &= m_2 - (m_1)^2 = \frac{1}{n}\sum_i X_i^2 - (\bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n (X_i^2 - \bar{X}_n^2). \end{aligned}$$

■

Although the method of moments usually yield consistent estimator they are in general inefficient. This is taken by Fisher in several papers in the 1920s and 30s arguing in favor of the MLE. The controversy between Pearson and Fisher about the relative merits of their respective methods of estimations ends in the mid-1930s with Fisher the winner and the absolute dominance then of the maximum likelihood method.

The basic reason for the inefficiency of the estimator based on the method of moment is not hard to find. It is due to the fact that the method does not use any information relating to the probability model Φ apart from the assumption that raw moments of the order k exist.

3.2.1 Generalized Method of Moments, GMM

The general statement of GMM was developed by Hansen (1982).[9] We begin GMM with an example.

Example.
Consider a random sample Y_t, $t = 1, ..., n$ from a standard t distribution with ν degree of freedom. Provided that $\nu > 2$, a standard t variables has population mean 0 and

[9]Hansen was awarded the Nobel Prize in Economics in 2013 for his contribution at GMM.

variance given by $\mu_2' = E(Y_t^2) = \frac{\nu}{(\nu-2)}$. Since $m_2 = n^{-1}\sum_{t=1}^{n} y_t^2$ would be a consistent estimator for μ_2', i.e. $m_2 \xrightarrow{a.s.} \mu_2'$, classical method of moments would suggest a consistent estimator of ν can be obtained by finding a solution to

$$\frac{\nu}{(\nu-2)} = m_2 \tag{3-5}$$

or

$$\hat{\nu}_n = \frac{2m_2}{m_2 - 1}.$$

That is to estimate a single population moment ν, we have used a single moment m_2. However, we might also have made used of other moments. If $\nu > 4$, the population fourth moment of a standard t variable is $\mu_4' = E(Y_t^4) = \frac{3\nu^2}{(\nu-2)(\nu-4)}$, and we might expect to find a solution to

$$\frac{3\nu^2}{(\nu-2)(\nu-4)} = m_4, \tag{3-6}$$

where $m_4 = n^{-1}\sum_{t=1}^{n} y_t^4$. We cannot choose the single parameter ν so that to match the sample second moment (3-5) and the sample fourth moment (3-6). However, we might try to choose ν so as to as close as possible to both, by minimizing a criterion function such as

$$Q(\nu;\ y_n, y_{n-1}, ..., y_1) = \mathbf{g}'\mathbf{W}\mathbf{g},$$

where

$$\mathbf{g} = \begin{bmatrix} \left\{ m_2 - \frac{\nu}{(\nu-2)} \right\} \\ \left\{ m_4 - \frac{3\nu^2}{(\nu-2)(\nu-4)} \right\} \end{bmatrix}.$$

Here, $\mathbf{W}$ is a (2×2) positive definite symmetric weighting matrix reflecting the importance given to matching each of the moments. ■

The derivation of GMM begins by defining the population moment condition.

Definition. (Population Moment Condition)
Let $\mathbf{w}_t$ be a vector of random variables, $\boldsymbol{\theta}_0$ be a $p \times 1$ vector of parameters, and $\mathbf{g}(\cdot)$ be a $q \times 1$ vector valued function. The population moment condition is defined

$$E[\mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0)] = 0. \tag{3-7}$$

The q rows of the population moment condition in (3-7) are sometimes described as *orthogonality conditions.* ■

Definition. (Sample Moment Condition)
The sample moment condition is derived from the average population moment condition,

$$\mathbf{g}_n(\mathbf{w}, \boldsymbol{\theta}_0) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{g}(\mathbf{w}_t, \boldsymbol{\theta}_0).$$ ■

Definition. (GMM Estimators)
The GMM estimator is defined as the value of $\boldsymbol{\theta}$ that minimizes

$$Q_n(\boldsymbol{\theta}) = \mathbf{g}_n(\mathbf{w}, \boldsymbol{\theta}_0)'\mathbf{W}_n\mathbf{g}_n(\mathbf{w}, \boldsymbol{\theta}_0).$$

Thus the GMM estimator is defined as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}}\ Q_n(\boldsymbol{\theta})$$

where $\mathbf{W}_n$ is a $q \times q$ positive semi-definite matrix. ■

In order to operationalize the GMM estimator, q, the number of moments, will be required to greater than or equal to p, the number of unknown parameters.

3.3 The Maximum Likelihood Estimation Method

The maximum likelihood method of estimation was formulated by Fisher and extended by various authors such as Cramer, Rao and Wald. In the current statistical literature the method of likelihood is by far the most widely used method of estimation and plays a very important role in hypothesis testing.

3.3.1 The Likelihood Function

Consider the statistical model (univariate):

(a). $\Phi = \{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\}$;

(b). $\mathbf{x} = (X_1, X_2, ..., X_n)'$ a sample from $f(x;\boldsymbol{\theta})$, where $\mathbf{x}$ takes values in $\mathcal{X} = \mathbb{R}^n$, the observation space.

The distribution of the sample $f(x_1, x_2, ..., x_n;\boldsymbol{\theta})$ describe how the density changes as $\mathbf{x}$ takes different value in $\mathcal{X}$ for a given $\boldsymbol{\theta} \in \Theta$. In deriving the likelihood function we reason as follows:

> "Since $f(\mathbf{x};\boldsymbol{\theta})$ incorporates all the information in the statistical model it makes a lot of intuitive sense to reverse the argument in deriving $f(\boldsymbol{x};\boldsymbol{\theta})$ and consider the question which value of $\boldsymbol{\theta} \in \Theta$ is mostly supported by a given sample realisation $\mathbf{x} = \boldsymbol{x}$? That is, the value $\boldsymbol{\theta}$ under which $\boldsymbol{x}$ would have had the highest "likelihood" of arising must be intuitively our best choice of $\boldsymbol{\theta}$."

Definition. (Likelihood Function)
Using this intuitive argument the likelihood function is defined by

$$L(\boldsymbol{\theta};\mathbf{x}) = k(\mathbf{x})D(\mathbf{x};\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta,$$

where $k(\mathbf{x}) > 0$ is a function of $\mathbf{x}$ only (not $\boldsymbol{\theta}$). In particular

$$L(\cdot;\mathbf{x}) : \Theta \to [0, 1].$$ ■

The presence of $k(\mathbf{x})$ (which we always chose arbitrarily to be equal to one) in the definition of $L(\boldsymbol{\theta};\mathbf{x})$ implies that the likelihood function is non-unique; any monotonic transformation of it represents the same information. In particular:

(a) $\log L(\boldsymbol{\theta};\mathbf{x})$, *the log likelihood functions; and*
(b) $\frac{\partial \log L(\boldsymbol{\theta};\mathbf{x})}{\partial \boldsymbol{\theta}} \equiv s(\boldsymbol{\theta};\mathbf{x})$, *the score function,*

incorporate the same information as $L(\boldsymbol{\theta};\mathbf{x})$.

3.3.2 The Maximum Likelihood Estimator (MLE)

Definition. (MLE)
Given that the likelihood function represents the support given to the various $\boldsymbol{\theta} \in \Theta$ given $\mathbf{x} = \boldsymbol{x}$, it is natural to define the *maximum likelihood estimator* of $\boldsymbol{\theta}$ to be a Borel function $\hat{\boldsymbol{\theta}} : \mathcal{X} \to \boldsymbol{\Theta}$ such that

$$L(\hat{\boldsymbol{\theta}}; \mathbf{x}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}; \mathbf{x}),$$

and there may be one, none or many such MLE's. Note that

$$\log L(\hat{\boldsymbol{\theta}}; \mathbf{x}) \geq \log L(\boldsymbol{\theta}^*; \mathbf{x}), \quad for\ all\ \boldsymbol{\theta}^* \in \boldsymbol{\Theta}. \qquad \blacksquare$$

In the case where $L(\boldsymbol{\theta}; \mathbf{x})$ is differentiable, the MLE can be derived as a solution of the equations

$$\frac{\partial \log L(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}} \equiv s(\boldsymbol{\theta}; \mathbf{x}) = 0,$$

refereed as the *likelihood equation*.

Example. (Univariate Normal Distribution, *i.i.d.* sample)
Let $X_1, X_2, ..., X_n$ be a random sample from the $N(\mu, \sigma^2)$. Find the MLEs of $\boldsymbol{\theta} = (\mu, \sigma^2)'$.

For random sampling from a normal distribution, the log-likelihood and its derivatives are

$$\begin{aligned}
\ln L(\mu, \sigma^2) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2, \\
\frac{\partial \ln L}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu), \\
\frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2.
\end{aligned}$$

Therefore, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(X_i) = \bar{X}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is indeed MLE since

$$\begin{aligned}
\left.\frac{\partial^2 \ln L}{\partial \mu^2}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= -\frac{n}{\hat{\sigma}^2} < 0, \\
\left.\frac{\partial^2 \ln L}{\partial (\sigma^2)^2}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{-n}{2\hat{\sigma}^4} < 0, \\
\left.\frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= -\frac{1}{\hat{\sigma}^4}\sum_{i=1}^{n}(X_i - \bar{X}) = 0.
\end{aligned}$$

We can deduce that

$$\mathbf{H}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{-n}{2\hat{\sigma}^4} \end{bmatrix}$$

is a negative definite matrix. (How ?) ■

The MLE estimators may be biased in finite sample as is illustrated in the following example.

Example.

The expected value the MLE estimators, $\hat{\mu}$, from a random sample of $N(\mu, \sigma^2)$ is

$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\left(\sum_{i=1}^{n} E(X_i)\right) = \frac{1}{n}n\mu = \mu.$$

The expected value the MLE estimators, $\hat{\sigma}^2$ is derived as follows. Denote $\mathbf{x} = (X_1, X_2, ..., X_n)'$ and $\boldsymbol{\mu} = (\mu, \mu, ..., \mu)'$ be the vector of this random sample and its mean vector respectively. Then the joint distribution of this sample would be that

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I}).$$

We know that

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \mathbf{x}'\mathbf{M}_0\mathbf{x} = (\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}_0(\mathbf{x} - \boldsymbol{\mu}),$$

since $\mathbf{M}_0(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{M}_0\mathbf{x} - \mathbf{M}_0\boldsymbol{\mu} = \mathbf{M}_0\mathbf{x}$, where $\mathbf{M}_0 = \mathbf{I} - 1/n(\mathbf{i}\mathbf{i}')$.

From the results of "quadratic forms related to the normal distribution" at section 8.2.3 of Ch. 2, we have

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{\mathbf{x}'\mathbf{M}_0\mathbf{x}}{\sigma^2} = \frac{(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}_0(\mathbf{x} - \boldsymbol{\mu})}{\sigma^2} \sim \chi^2_{trace\ of\ \mathbf{M}_0} \equiv \chi^2_{n-1}.$$

Therefore,

$$E\left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}\right)=(n-1),$$

or

$$E\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=(n-1)\sigma^2.$$

The expected value of the MLE estimator of $\sigma^2, \hat{\sigma}^2$ is

$$E(\hat{\sigma}^2)=E\left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}\right)=\frac{n-1}{n}\sigma^2\neq\sigma^2,$$

which is biased. An unbiased estimator of σ^2 would be

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

Example.

Let $X_1, X_2, ..., X_n$ be a random sample from the density $f(x;\theta)=\theta(1-\theta)^x$, $x=0,1,...,$ $0<\theta<1$. Find the MLEs of θ.

Since this is a random sample, the likelihood function takes the form

$$\begin{aligned}L(\theta;X_1,X_2,...,X_n) &= f(X_1;\theta)f(X_2;\theta)...f(X_n;\theta)\\ &= \theta(1-\theta)^{X_1}\cdot\theta(1-\theta)^{X_2}\cdots\theta(1-\theta)^{X_n}\\ &= \theta^n(1-\theta)^{\sum_{i=1}^{n}X_i},\end{aligned}$$

and the log-likelihood is therefore

$$\ln L(\theta;X_1,X_2,...,X_n)=n\ln\theta+\sum_{i=1}^{n}X_i\ln(1-\theta).$$

The FOC condition is

$$\frac{\partial\ln L(\theta;X_1,X_2,...,X_n)}{\partial\theta}=\frac{n}{\theta}-\frac{\sum_{i=1}^{n}X_i}{1-\theta}=0$$

which has a solution at

$$\hat{\theta}=\frac{1}{1+1/n\sum_{i=1}^{n}X_i}=\frac{1}{1+\bar{X}}.$$ ■

Exercise 1.
Check second order condition to make sure that the solution from FOC is indeed a MLE in the last example. ■

We can extend our probability model to a multivariate one and estimate the parameters by MLE.

Example. (Multivariate Normal Distribution, *i.i.d.* Sample)
Let $\mathbf{x} \equiv (\mathbf{x}_1', \mathbf{x}_2', ..., \mathbf{x}_T')'$ be a random sample of size T from $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i.e. $\mathbf{x}_t \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}), t = 1, 2, ..., T$. The likelihood function takes the form

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x}) &= \prod_{t=1}^{T} \left[(2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{x}_t - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}) \right\} \right] \\ &= (2\pi)^{-nT/2} |\boldsymbol{\Sigma}|^{-T/2} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} (\mathbf{x}_t - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}) \right\}, \end{aligned}$$

and the log-likelihood function therefore is

$$\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x}) = -\frac{nT}{2} \ln(2\pi) - \frac{T}{2} \ln(|\boldsymbol{\Sigma}|) - \frac{1}{2} \sum_{t=1}^{T} (\mathbf{x}_t - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}).$$

The FOC is

$$\begin{aligned} \frac{\partial \ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{x})}{\partial \boldsymbol{\mu}} &= -\boldsymbol{\Sigma}^{-1} \sum_{t=1}^{T} (\mathbf{x}_t - \boldsymbol{\mu}) = 0 \\ \Rightarrow \quad & \sum_{t=1}^{T} \mathbf{x}_t = T\boldsymbol{\mu} \\ \Rightarrow \quad & \hat{\boldsymbol{\mu}} = \frac{\sum_{t=1}^{T} \mathbf{x}_t}{T} = \bar{\mathbf{x}}_T, \end{aligned}$$

and

$$\begin{aligned}\frac{\partial \ln L(\boldsymbol{\mu},\boldsymbol{\Sigma};\mathbf{x})}{\partial \boldsymbol{\Sigma}^{-1}} &= \frac{T}{2}\frac{\partial \ln |\boldsymbol{\Sigma}^{-1}|}{\partial \boldsymbol{\Sigma}^{-1}} - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{x}_t-\boldsymbol{\mu})(\mathbf{x}_t-\boldsymbol{\mu})' \\ &= \frac{T}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{x}_t-\boldsymbol{\mu})(\mathbf{x}_t-\boldsymbol{\mu})' = 0 \\ \Rightarrow\ T\hat{\boldsymbol{\Sigma}} &= \sum_{t=1}^{T}(\mathbf{x}_t-\bar{\mathbf{x}}_T)(\mathbf{x}_t-\bar{\mathbf{x}}_T)' \\ \Rightarrow\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_t-\bar{\mathbf{x}}_T)(\mathbf{x}_t-\bar{\mathbf{x}}_T)'. \qquad \blacksquare\end{aligned}$$

It would be an *erroneous conclusion* that deriving the MLE is a matter of a simple differentiation and solution of linear equations. Let us consider an example where the derivation is not as straightforward.

Example.

Let $\mathbf{z} = (\mathbf{z}_1', \mathbf{z}_2', ..., \mathbf{z}_n')'$ where $\mathbf{z}_i = (X_i, Y_i)$, be a random sample from

$$f(x,y;\rho) = \frac{(1-\rho^2)^{-1/2}}{2\pi}\exp\left\{-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)\right\},$$

then the log-likelihood function is

$$\log L(\rho; X, Y) = -\log 2\pi - \frac{n}{2}\log(1-\rho^2) - \frac{1}{2(1-\rho^2)}\sum_{i=1}^{n}(X_i^2 - 2\rho X_iY_i + Y_i^2),$$

and the FOC for a maximum is

$$\frac{d\log L}{d\rho} = \frac{n(-2)}{2(1-\rho^2)}\rho - \rho\frac{\sum_{i=1}^{n}(X_i^2+Y_i^2)}{(1-\rho^2)^2} + \frac{1+\rho^2}{(1-\rho^2)^2}\sum_{i=1}^{n}(X_iY_i) = 0.$$

However this is a nonlinear equation that *cannot* be solved explicitly for ρ and additional search techniques are needed to locate the maximum value such as using numerical methods.

3.3.3 A Short Tour to Numerical Method

We consider the general problem of maximizing a function of several variables:

$$\text{maximize}_{\boldsymbol{\theta}}\ F(\boldsymbol{\theta}),$$

where $F(\boldsymbol{\theta})$ may be log-likelihood or some other function. An efficient means of solving most nonlinear maximization problem is by an *iterative algorithm*:

> "Beginning from initial value $\boldsymbol{\theta}_0$, at entry to iteration t, if $\boldsymbol{\theta}_t$, is not the optimal value for $\boldsymbol{\theta}$, compute direction vector $\boldsymbol{\Delta}_t$, and step size λ_t, then
>
> $$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t."$$

The most commonly used algorithm are gradient method and template for most gradient method in common use is the Newton's method. The basis for Newton's method is a linear Taylor series approximation. Expanding the first-order conditions,

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (\textit{which may have nonlinear solution})$$

in a linear Taylor series around an arbitrary $\boldsymbol{\theta}^0$ yield

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \simeq \mathbf{g}^0 + \mathbf{H}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0),$$

where the superscript indicates that the term is evaluated at $\boldsymbol{\theta}^0$ and $\mathbf{g}$ and $\mathbf{H}$ are the gradient vector and Hessian matrix, respectively. If $F(\boldsymbol{\theta})$ attains a local maximum at $\boldsymbol{\theta}_1$, then we must necessarily have $\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}_1} = \mathbf{0}$. Solving for $\boldsymbol{\theta}_1$, we obtain

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - \mathbf{H}_0^{-1}\mathbf{g}_0.$$

If now we approximate $\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ with another linear function, by again applying Taylor's expansion in a neighborhood of $\boldsymbol{\theta}_1$,[10] and then repeat the same process as before with $\boldsymbol{\theta}_1$ used instead of $\boldsymbol{\theta}_0$, we obtain

$$\boldsymbol{\theta}_2 = \boldsymbol{\theta}_1 - \mathbf{H}_1^{-1}\mathbf{g}_1.$$

Further repetitions of this process lead to the iteration,

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \mathbf{H}_i^{-1}\mathbf{g}_i, \quad i = 0, 1, 2, ...$$

[10] It is noted that here, for a given initial $\boldsymbol{\theta}_0$, we can obtain the value of $\boldsymbol{\theta}_1$.

The iterations stop at the value of $\boldsymbol{\theta}^*$ such that $\mathbf{g}^* \equiv \frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}^*} = \mathbf{0}$. Under this circumstance, $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$ and therefore $\mathbf{g}_{i+1} = \mathbf{g}_i = \mathbf{0}$. In this case, $\lambda_t = 1$ and $\boldsymbol{\Delta}_t = -\mathbf{H}_t^{-1}\mathbf{g}_t$.

The Newton-Raphson method requires finding the inverse of the Hessian matrix $\mathbf{H}$ at each iteration. This can be computationally involved, especially if the number of the variables in $\boldsymbol{\theta}$ is large. Furthermore, the method may fail to converge if $\mathbf{H}_i$ is not positive definite. This can occur, for example, when $\boldsymbol{\theta}_i$ is far from the location $\boldsymbol{\theta}^*$ of the true maximum. If, however, the initial point $\boldsymbol{\theta}_0$, is close to $\boldsymbol{\theta}^*$, then convergence occurs at a rapid rate.

3.3.4 Finite Sample Properties of MLE

Let us discuss the finite sample properties of MLE in this subsection. One of the most important properties of MLE is the *invariance principle*.

Result. (Invariance Principle)
Let $\hat{\boldsymbol{\theta}}$ be a MLE of $\boldsymbol{\theta}$. If $g(\cdot) : \Theta \to \mathbb{R}$ is a Borel function of $\boldsymbol{\theta}$ then a MLE of $g(\boldsymbol{\theta})$ exists and is given by $g(\hat{\boldsymbol{\theta}})$. ■

This means that if the MLE of $\boldsymbol{\theta}$ is available then for functions such as $\boldsymbol{\theta}^k$, $e^{\boldsymbol{\theta}}$, $\log \boldsymbol{\theta}$, its MLE is derived by substitution $\hat{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$, i.e. $\hat{\boldsymbol{\theta}}^k$, $e^{\hat{\boldsymbol{\theta}}}$ and $\log \hat{\boldsymbol{\theta}}$ are MLE's of these functions.

As seen in the case of $\hat{\sigma}^2$, MLE are not in general unbiased estimator. In one particular case, when unbiasedness is accompanied by full efficiency, however, the two coincide.

Result. (Unbiasedness is Accomplished with Full-Efficiency)
In the case where Φ satisfies the regularity conditions and $\hat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$ whose variance achieves the Cramer-Rao lower bound, then the likelihood estimation has a unique solution equal to $\hat{\boldsymbol{\theta}}$. This suggests that any unbiased fully efficient estimator $\hat{\boldsymbol{\theta}}$ can be derived as a solution of the likelihood equation, as in the case of $\hat{\mu}$.■

The properties mostly emphasised by Fisher in support of the MLE was the properties of sufficiency.

Result. (Sufficiency)
If $\tau(\mathbf{x})$ is a sufficient statistic for θ and a unique MLE $\hat{\theta}$ of θ exists then $\hat{\theta}$ is a function of $\tau(\mathbf{x})$. ■

3.3.5 Asymptotic Properties ($i.i.d.$ case and θ is a scalar) of MLE

Although MLE's enjoy several optimum finite sample properties, as seen above, their asymptotic properties provide the main justification for the almost universal appeal to the method of maximum likelihood. As argued below, under certain regularity conditions, MLE can be shown to be consistent, asymptotically normal and asymptotically efficient.

Let us begin the discussion of asymptotic properties enjoyed by MLE by considering the simplest possible case where the statistical model is as follows:
(a). probability model, $\Phi = \{f(\boldsymbol{x};\theta), \theta \in \mathbf{\Theta}\}$, θ being a scalar;
(b). sampling model, $\mathbf{x} \equiv (X_1, X_2, ..., X_n)'$ is a random sample from $f(\boldsymbol{x};\theta)$.

For our purpose it suffices to repeat the regularity conditions on Φ:

(RC 1.) The set $A = \{x : f(\boldsymbol{x};\theta) > 0\}$ does not depend on θ.

(RC 2.) For each $\theta \in \mathbf{\Theta}$ the derivatives $[\partial^i \log f(\boldsymbol{x};\theta)]/(\partial\theta^i)$, $i = 1, 2, 3$, exist for all $\boldsymbol{x} \in \mathcal{X}$.

(RC 3.) $0 < E[(\partial/\partial\theta)\log f(\boldsymbol{x};\theta)]^2 < \infty$ for all $\theta \in \mathbf{\Theta}$.

(RC 4.) For every $\theta \in \mathbf{\Theta}$,

$$\left|\frac{\partial^i \log f(x;\theta)}{\partial\theta^i}\right| \leq h_i(x), \quad i = 1, 2, 3,$$

where the function $h_1(x)$ and $h_2(x)$ are integrable over $(-\infty, \infty)$, i.e.

$$\int_{-\infty}^{\infty} h_i(x)dx < \infty, \quad i = 1, 2,$$

and

$$\int_{-\infty}^{\infty} h_3(x) f(x;\theta) dx < k,$$

where k does not depend on θ.

Theorem.

Under above regular conditions, the likelihood equation $[\partial \log L(\theta;\mathbf{x})]/\partial\theta = 0$ admits a sequence of solution $\{\hat{\theta}_n,\ n \geq 1\}$ such that:

(a). Consistency:[11]

$$\hat{\theta}_n \xrightarrow{a.s.} \theta, \quad \text{i.e. strong consistency.}$$

(b). Asymptotic normality:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{L} N(0, I(\theta)^{-1}).$$

(c). Asymptotic efficiency:

$$I(\theta) = \lim_{n\to\infty} \left[\frac{1}{n} I_n(\theta)\right],$$

i.e. the asymptotic variance of $\hat{\theta}_n$ equal the limit of the Cramer-Rao lower bound for consistent estimator (that is, $Var(\hat{\theta}_n) = \frac{1}{n} \cdot (I(\theta))^{-1} = (I_n(\theta))^{-1}$).

Proof.

Take a Taylor expansion of $[\partial \log L(\hat{\theta}_n;\mathbf{x})]/\partial\theta$ at $\hat{\theta} = \theta_0$ we would get

$$\frac{\partial \log L(\hat{\theta})}{\partial\theta} = \frac{\partial \log L(\theta_0)}{\partial\theta} + \frac{\partial^2 \log L(\theta_0)}{\partial\theta^2}(\hat{\theta}_n - \theta_0) + o_p(n) = 0,$$

where $o_p(n)$ refers to all terms of order n.[12] The above expansion is based on RC 2 and RC 4.

[11] This implies that $\hat{\theta}_n \xrightarrow{P} \theta_0$, i.e. weak consistency.

[12] Here, $o_p(n)$ is from $\sum_1^n o_p(1)$, and $o_p(1)$ is from the fact that the remainder for the nth degree Taylor polynomial $\lim_{\hat{\theta}_n \to \theta_0} R_n(x) = 0$. See Arnold, p.232; Fulk, p.122; and Hamilton, p.713.

In view of the fact the $\log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta)$ and the $f(x_i;\theta), i = 1, 2, ..., n$, is *i.i.d.*, we can express the above Taylor expansion in the form

$$\underbrace{-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(x_i;\theta_0)}{\partial\theta^2}}_{B_n}\cdot(\hat{\theta}_n - \theta_0) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(x_i;\theta_0)}{\partial\theta}}_{A_n} + o_p(1). \tag{3-8}$$

Using the strong law of large numbers for the *i.i.d.* r.v.'s,

$$A_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(x_i;\theta_0)}{\partial\theta} \xrightarrow{a.s.} E\left(\frac{\partial \log f(x;\theta_0)}{\partial\theta}\right) = 0$$

and

$$B_n = \left(-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(x_i;\theta_0)}{\partial\theta^2}\right) \xrightarrow{a.s.} E\left(-\frac{\partial^2 \log f(x;\theta_0)}{\partial\theta^2}\right) \equiv I_i(\theta),$$

where $\sum_{i=1}^{n} I_i(\theta) = I_n(\theta)$ or $I_n(\theta) = nI(\theta)$.[13] This in turn implies that

$$(\hat{\theta}_n - \theta_0) \xrightarrow{a.s.} 0, \quad or \quad \hat{\theta}_n \xrightarrow{a.s.} \theta_0. \ \Diamond$$

To show the normality we multiple all the terms of the Taylor expansion by $\sqrt{n}$ to get

$$B_n\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(x_i;\theta_0)}{\partial\theta} + o_p(\sqrt{n}).$$

Using the central limit theorem for *i.i.d.* r.v.'s we can show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(x_i;\theta_0)}{\partial\theta} \xrightarrow{L} N(0, I_i(\theta)).$$

Given that $B_n \xrightarrow{a.s.} I_i(\theta)$ we can deduce that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{L} N(0, I_i(\theta)^{-1}) + o_p(\sqrt{n}) \equiv N(0, nI_n^{-1}(\theta)) + o_p(\sqrt{n}).$$

Hence,

$$(\hat{\theta}_n - \theta) \xrightarrow{L} N(0, I_n(\theta)^{-1}), \text{ or}$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N(0, I(\theta)^{-1}). \quad \blacksquare \tag{3-9}$$

[13] see Arnold, p.271.

Example.

Let $\mathbf{x} \equiv (X_1, X_2, ..., X_n)'$ be a random sample from $f(x;\theta) = 1/\theta$, where $0 \leq x \leq \theta$. Since the range of the random variables $(X_1, X_2, ..., X_n)$ depends on the parameters θ. This function violates regularity condition 1 (RC 1) above and is excluded from the properties below. ■

Example.

Let $\mathbf{x} \equiv (X_1, X_2, ..., X_n)'$ be a random sample from $N(0, \sigma^2)$. Then

$$\frac{\partial^3 \log f(\mathbf{x})}{\partial \sigma^6} = -\frac{n}{\sigma^6} + \frac{3\sum X^2}{\sigma^8} \to \infty, \quad as\ \sigma^2 \to 0,$$

that is, the third derivative is not bounded in the open interval $0 < \sigma^2 < \infty$, the regularity condition 4 (RC 4) is not satisfied. ■

3.3.6 Estimating the Asymptotic Variance of the MLE, $\boldsymbol{\theta}$ is $k \times 1$

The asymptotic covariance matrix of the MLE is a matrix of parameters that must be estimated. The followings are three methods to estimate this variance.

(a). If the form of the expected value of the second derivative of the log-likelihood is known, we can evaluate the information matrix at $\hat{\boldsymbol{\theta}}$ to estimate the covariance matrix for the MLE,

$$\widehat{[\mathbf{I}_n(\boldsymbol{\theta})]^{-1}} = \left\{ -E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\}^{-1}.$$

If the expected value of the second derivative of the log-likelihood is complicated, two alternative estimators is

(b).

$$\widetilde{[\mathbf{I}_n(\boldsymbol{\theta})]^{-1}} = \left\{ -\left[\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}\ \partial \hat{\boldsymbol{\theta}}'} \right] \right\}^{-1}, \text{ and}$$

(c). the BHHH estimator

$$[\mathbf{I}_n(\check{\boldsymbol{\theta}})]^{-1} = \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i'\right]^{-1}, \ where\ \hat{\mathbf{g}}_i = \left.\frac{\partial \ln f(x_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \ is\ k \times 1.$$

The reason for using the BHHH is that

$$\ln L(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i;\boldsymbol{\theta}).$$

Therefore,

$$\frac{\partial \ln L}{\partial \boldsymbol{\theta}} = \mathbf{g} = \sum_{i=1}^{n} \mathbf{g}_i, \quad (\mathbf{g}_i = \frac{\partial \ln f(x_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, is\ k \times 1),$$

and

$$\begin{aligned} -E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] &= E\left[\left(\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\right)'\right] \\ &= E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{g}_i \mathbf{g}_j'\right], \end{aligned}$$

we drop the unequal subscript since $\ln f(x_i;\boldsymbol{\theta}),\ i = 1, 2, ..., n$ is a random sample to obtain

$$-E\left[\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] = E\left[\sum_{i=1}^{n} \mathbf{g}_i \mathbf{g}_i'\right].$$

Exercise 2.

Suppose a random sample $(X_i, Y_i)',\ i = 1, ..., 20$ are generated by a model of the form

$$f(x, y; \beta) = \frac{1}{x+\beta} \exp(-y/(x+\beta)).$$

From the following data, find the maximum likelihood estimate of β and compute the asymptotic variance of this MLE by the above three methods.

$X = 12, 16, 18, 16, 12, 12, 16, 12, 10, 12, 16, 20, 12, 16, 10, 18, 16, 20, 12, 16,$ and

$Y = 20.5, 31.5, 47.7, 26.2, 44, 8.28, 30.8, 17.2, 19.9, 9.96, 55.8, 25.2, 29, 85.5, 15.1,$
$28.5, 21.4, 17.7, 6.42, 84.9.$ ■

Beaumont Tower, MSU.

End of this Chapter