Chapter 5

Point Estimation

Objective: In this chapter we study some statistical methods to find point estimators of population parameters and study their properties.

5.1 Introduction 226
5.2 The Method of Moments 227
5.3 The Method of Maximum Likelihood 235
5.4 Some Desirable Properties of Point Estimators 246
5.5 Other Desirable Properties of a Point Estimator 266
5.6 Chapter Summary 282
5.7 Computer Examples 283
Projects for Chapter 5 285

C. R. Rao

(*Source: http:www.science.psu.edu/alert/Rao6-2007.htm*)

Calyampudi Radhakrishna (C. R.) Rao (1920–) is a contemporary statistician whose work has influenced not just statistics, but such diverse fields as anthropology, biometry, demography, economics,

genetics, geology, and medicine. Several statistical terms and equations are named after Rao. He has worked with many other famous statisticians such as Blackwell, Fisher, and Neyman and has had dozens of theorems named after him. Rao earned an M.A. in mathematics and another M.A. in statistics, both in India, and earned his Ph.D. and Sc.D. at Cambridge University. The following was stated in the Preface to the 1991 special issue of the *Journal of Quantitative Economics* in Rao's honor: "Dr. Rao is a very distinguished scientist and a highly eminent statistician of our time. His contributions to statistical theory and applications are well known, and many of his results, which bear his name, are included in the curriculum of courses in statistics at bachelor's and master's level all over the world. He is an inspiring teacher and has guided the research work of numerous students in all areas of statistics. His early work had greatly influenced the course of statistical research during the last four decades. One of the purposes of this special issue is to recognize Dr. Rao's own contributions to econometrics and acknowledge his major role in the development of econometric research in India." The importance of statistics can be summarized in Rao's own words: "If there is a problem to be solved, seek statistical advice instead of appointing a committee of experts. Statistics can throw more light than the collective wisdom of the articulate few."

5.1 INTRODUCTION

In statistical analysis, point estimation of population parameters plays a very significant role. In studying a real-world phenomenon we begin with a random sample of size n taken from the totality of a population. The initial step in statistically analyzing these data is to be able to identify the probability distribution that characterizes this information. Because the parameters of a distribution are its defining characteristics, it becomes necessary to know the parameters. In the present chapter, we assume that the form of the population distribution is known (binomial, normal, etc.) but the parameters of the distribution (p for a binomial; μ and σ^2 for a normal, etc.) are unknown. We shall estimate these parameters using the data from our random sample. It is extremely important to have the best possible estimate of the population parameter(s). Having such estimates will lead to a better and more accurate statistical analysis.

For example, in the area of phosphate mining in Florida, we may be interested in estimating the average radioactivity from both uranium and radium in a clay settling area of a mining site. Suppose that a random sample of 10 such sites resulted in a sample average of 40 pCi/g (picocuries/gram) of radioactivity. We may use this value as an estimate of the average radioactivity for all of the settling areas of mining sites in Florida. Because many Florida crops are grown on clay settling areas, this type of estimate is important for accessing the radioactivity-associated risks that are due to eating food from the crops grown on these clay settling areas.

We will now introduce some of the more useful statistical point estimation methods, discuss their properties, and illustrate their usefulness with a number of applications. The importance of point estimates lies in the fact that many statistical formulas are based on them. For example, the point estimates of mean and standard deviation are used in the calculation of confidence intervals and in many formulas for hypothesis testing. These topics are covered in subsequent chapters. Also, in most applied problems, a certain numerical characteristic of the physical phenomenon may be of interest; however, its value may not be observable directly. Instead, suppose it is possible to observe one or more random variables, the distribution of which depends on the characteristic of interest. Our

objective will be to develop methods that use the observed values of random variables (sample data) in order to gain information about the unknown and unobservable characteristic of the population.

Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) random variables (in statistical language, a random sample) with a pdf or pf $f(x, \theta_1, \ldots \theta_l)$, where $\theta_1, \ldots, \theta_l$ are the unknown population parameters (characteristics of interest). For example, a normal pdf has parameters μ (the mean) and σ^2 (the variance). The actual values of these parameters are not known. The problem in point estimation is to determine statistics $g_i(X_1, \ldots, X_n)$, $i = 1, \ldots, l$, which can be used to estimate the value of each of the parameters—that is, to assign an appropriate value for the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_l)$ based on observed sample data from the population. These statistics are called *estimators* for the parameters, and the values calculated from these statistics using particular sample data values are called *estimates* of the parameters. Estimators of θ_i are denoted by $\hat{\theta}_i$, where $\hat{\theta}_i = g_i(X_1, \ldots, X_n)$, $i = 1, \ldots, l$. Observe that the estimators are random variables. As a result, an estimator has a distribution (which we called the sampling distribution in Chapter 4). When we actually run the experiment and observe the data, let the observed values of the random variables be $X_1, \ldots, X_n$ be $x_1, \ldots, x_n$; then, $\hat{\theta}(X_1, \ldots, X_n)$ is an estimator, and its value $\hat{\theta}(x_1, \ldots, x_n)$ is an estimate. For example, in case of the normal distribution, the parameters of interest are $\theta_1 = \mu$, and $\theta_2 = \sigma^2$, that is, $\boldsymbol{\theta} = (\mu, \sigma^2)$. If the estimators of μ and σ^2 are $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$ and $S^2 = (1/n-1)\sum_{i=1}^{n}(X_i - \overline{X})^2$ respectively, then, the corresponding estimates are $\overline{x} = (1/n)\sum_{i=1}^{n} x_i$ and $s^2 = (1/n-1)\sum_{i=1}^{n}(x_i - \overline{x})^2$, the mean and variance corresponding to the particular observed sample values. In this book, we use capital letters such as $\overline{X}$ and S^2 to represent the estimators, and lowercase letters such as $\overline{x}$ and s^2 to represent the estimates.

There are many methods available for estimating the true value(s) of the parameter(s) of interest. Three of the more popular methods of estimation are the method of moments, the method of maximum likelihood, and Bayes' method. A very popular procedure among econometricians to find a point estimator is the generalized method of moments. In this chapter we study only the method of moments and the method of maximum likelihood for obtaining point estimators and some of their desirable properties. In Chapter 11, we shall discuss Bayes' method of estimation.

There are many criteria for choosing a desired point estimator. Heuristically, some of them can be explained as follows (detailed coverage is given in Sections 5.2 through 5.5). An estimator, $\hat{\theta}$, is unbiased if the mean of its sampling distribution is the parameter θ. The bias of $\hat{\theta}$ is given by $B = E(\hat{\theta}) - \theta$. The estimator satisfies the consistency property if the sample estimator has a high probability of being close to the population value θ for a large sample size. The concept of efficiency is based on comparing variances of the different unbiased estimators. If there are two unbiased estimators, it is desirable to have the one with the smaller variance. The estimator has the sufficiency property if it fully uses all the sample information. Minimal sufficient statistics are those that are sufficient for the parameter and are functions of every other set of sufficient statistics for those same parameters. A method due to Lehmann and Scheffé can be used to find a minimal sufficient statistic.

5.2 THE METHOD OF MOMENTS

How do we find a good estimator with desirable properties? One of the oldest methods for finding point estimators is the method of moments. This is a very simple procedure for finding an estimator for one or more population parameters. Let $\mu'_k = E[X^k]$ be the kth moment about the origin of a

random variable X, whenever it exists. Let $m'_k = (1/n)\sum_{i=1}^{n} X_i^k$ be the corresponding kth sample moment. Then, the estimator of μ'_k by the method of moments is m'_k. The method of moments is based on matching the sample moments with the corresponding population (distribution) moments and is founded on the assumption that sample moments should provide good estimates of the corresponding population moments. Because the population moments $\mu'_k = h_k(\theta_1, \theta_2, \ldots, \theta_l)$ are often functions of the population parameters, we can equate corresponding population and sample moments and solve for these parameters in terms of the moments.

METHOD OF MOMENTS

Choose as estimates those values of the population parameters that are solutions of the equations $\mu'_k = m'_k, k = 1, 2, \ldots, l$. Here μ'_k is a function of the population parameters.

For example, the first population moment is $\mu'_1 = E(X)$, and the first sample moment is $\overline{X} = \sum_{i=1}^{n} X_i/n$. Hence, the moment estimator of μ'_1 is $\overline{X}$. If $k = 2$, then the second population and sample moments are $\mu'_2 = E(X^2)$ and $m'_2 = (1/n)\sum_{i=1}^{n} X_i^2$, respectively. Basically, we can use the following procedure in finding point estimators of the population parameters using the method of moments.

THE METHOD OF MOMENTS PROCEDURE

Suppose there are l parameters to be estimated, say $\theta = (\theta_1, \ldots, \theta_l)$.

1. Find l population moments, $\mu'_k, k = 1, 2, \ldots, l$. μ'_k will contain one or more parameters $\theta_1, \ldots, \theta_l$.
2. Find the corresponding l sample moments, $m'_k, k = 1, 2, \ldots, l$. The number of sample moments should equal the number of parameters to be estimated.
3. From the system of equations, $\mu'_k = m'_k, k = 1, 2, \ldots, l$, solve for the parameter $\theta = (\theta_1, \ldots, \theta_l)$; this will be a moment estimator of $\hat{\theta}$.

The following examples illustrate the method of moments for population parameter estimation.

Example 5.2.1

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli population with parameter p.

(a) Find the moment estimator for p.

(b) Tossing a coin 10 times and equating heads to value 1 and tails to value 0, we obtained the following values:

0 1 1 0 1 0 1 1 1 0

Obtain a moment estimate for p, the probability of success (head).

Solution

(a) *For the Bernoulli random variable,* $\mu_k' = E[X] = p$*, so we can use* m_1' *to estimate* p*. Thus,*

$$m_1' = \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Let

$$Y = \sum_{i=1}^{n} X_i.$$

Then, the method of moments estimator for p *is* $\hat{p} = Y/n$*. That is, the ratio of the total number of heads to the total number of tosses will be an estimate of the probability of success.*

(b) *Note that this experiment results in Bernoulli random variables. Thus, using part (a) with* $Y = 6$*, we get the moment estimate of* p *is* $\hat{p} = \frac{6}{10} = 0.6$*.*
We would use this value $\hat{p} = 0.6$*, to answer any probabilistic questions for the given problem. For example, what is the probability of exactly obtaining 8 heads out of 10 tosses of this coin? This can be obtained by using the binomial formula, with* $\hat{p} = 0.6$*, that is,* $P(X = 8) = \binom{10}{8}(0.6)^8(0.4)^{10-8}$*.*

In Example 5.2.1, we used the method of moments to find a single parameter. We demonstrate in Example 5.2.2 how this method is used for estimating more than one parameter.

Example 5.2.2

Let $X_1, \ldots, X_n$ be a random sample from a gamma probability distribution with parameters α and β. Find moment estimators for the unknown parameters α and β.

Solution

For the gamma distribution (see Section 3.2.5),

$$E[X] = \alpha\beta \quad \text{and} \quad E\left[X^2\right] = \alpha\beta^2 + \alpha^2\beta^2.$$

Because there are two parameters, we need to find the first two moment estimators. Equating sample moments to distribution (theoretical) moments, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} = \alpha\beta, \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \alpha\beta^2 + \alpha^2\beta^2.$$

Solving for α *and* β *we obtain the estimates as* $\alpha = (\overline{x}/\beta)$ *and* $\beta = \left[\{(1/n)\sum_{i=1}^{n} x_i^2 - \overline{x}^2\}/\overline{x}\right]$.

Therefore, the method of moments estimators for α and β are

$$\hat{\alpha} = \frac{\overline{X}}{\hat{\beta}}$$

and

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}{\overline{X}} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n\overline{X}},$$

which implies that

$$\hat{\alpha} = \frac{\overline{X}}{\hat{\beta}} = \frac{\overline{X}^2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2} = \frac{\overline{X}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}.$$

Thus, we can use these values in the gamma pdf to answer questions concerning the probabilistic behavior of the r.v. X.

■

Example 5.2.3

Let the distribution of X be $N(\mu, \sigma^2)$.

(a) For a given sample of size n, use the method of moments to estimate μ and σ^2.

(b) The following data (rounded to the third decimal digit) were generated using Minitab from a normal distribution with mean 2 and a standard deviation of 1.5.

3.163	1.883	3.252	3.716	−0.049	−0.653	0.057	2.987
4.098	1.670	1.396	2.332	1.838	3.024	2.706	0.231
3.830	3.349	−0.230	1.496				

Obtain the method of moments estimates of the true mean and the true variance.

Solution

(a) *For the normal distribution, $E(X) = \mu$, and because $Var(X) = EX^2 - \mu^2$, we have the second moment as $E(X^2) = \sigma^2 + \mu^2$.*

Equating sample moments to distribution moments we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu_1' = \mu$$

and

$$\mu_2' = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2.$$

Solving for μ and σ^2, we obtain the moment estimators as

$$\hat{\mu} = \overline{X}$$

and

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

(b) *Because we know that the estimator of the mean is $\hat{\mu} = \overline{X}$ and the estimator of the variance is $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n} X_i^2 - \overline{X}^2$, from the data the estimates are $\hat{\mu} = 2.005$, and $\hat{\sigma}^2 = 6.12 - (2.005)^2 = 2.1$. Notice that the true mean is 2 and the true variance is 2.25, which we used to simulate the data.*

In general, using the population pdf we evaluate the lower order moments, finding expressions for the moments in terms of the corresponding parameters. Once we have population (theoretical) moments, we equate them to the corresponding sample moments to obtain the moment estimators.

Example 5.2.4

Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $[a, b]$. Obtain method of moment estimators for a and b.

Solution

Here, a and b are treated as parameters. That is, we only know that the sample comes from a uniform distribution on some interval, but we do not know from which interval. Our interest is to estimate this interval.

The pdf of a uniform distribution is

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise.} \end{cases}$$

Hence, the first two population moments are

$$\mu_1 = E(X) = \int_a^b \frac{x}{b-a}dx = \frac{a+b}{2} \quad \text{and} \quad \mu_2 = E(X^2) = \int_a^b \frac{x^2}{b-a}dx = \frac{a^2+ab+b^2}{3}.$$

The corresponding sample moments are

$$\hat{\mu}_1 = \overline{X} \quad \text{and} \quad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Equating the first two sample moments to the corresponding population moments, we have

$$\hat{\mu}_1 = \frac{a+b}{2} \quad \text{and} \quad \hat{\mu}_2 = \frac{a^2+ab+b^2}{3}$$

which, solving for a and b, results in the moment estimators of a and b,

$$\hat{a} = \hat{\mu}_1 - \sqrt{3\left(\hat{\mu}_2 - \hat{\mu}_1^2\right)} \quad \text{and} \quad \hat{b} = \hat{\mu}_1 + \sqrt{3\left(\hat{\mu}_2 - \hat{\mu}_1^2\right)}.$$

In Example 5.2.4, if $a = -b$, that is, $X_1, \ldots, X_n$ is a random sample from a uniform distribution on the interval $(-b, b)$, the problem reduces to a one-parameter estimation problem. However, in this case $E(X_i) = 0$, so the first moment cannot be used to estimate b. It becomes necessary to use the second moment. For the derivation, see Exercise 5.2.3.

It is important to observe that the method of moments estimators need not be unique. The following is an example of the nonuniqueness of moment estimators.

Example 5.2.5

Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter $\lambda > 0$. Show that both $(1/n)\sum_{i=1}^n X_i$ and $(1/n)\sum_{i=1}^n X_i^2 - \left((1/n)\sum_{i=1}^n X_i\right)^2$ are moment estimators of λ.

Solution

We know that $E(X) = \lambda$, from which we have a moment estimator of λ as $(1/n)\sum_{i=1}^n X_i$. Also, because we have $Var(X) = \lambda$, equating the second moments, we can see that

$$\lambda = E(X^2) - (EX)^2,$$

so that

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2.$$

Thus,

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

and

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2.$$

Both are moment estimators of λ. Thus, the moment estimators may not be unique. We generally choose $\overline{X}$ as an estimator of λ, for its simplicity.

It is important to note that, in general, we have as many moment conditions as the parameters. In Example 5.2.5, we have more moment conditions than parameters, because both the mean and variance of Poisson random variables are the same. Given a sample, this results in two different estimates of a single parameter. One of the questions could be, can these two estimators be combined in some optimal way? This is done by the so-called generalized method of moments (GMM). We will not deal with this topic.

As we have seen, the method of moments finds estimators of unknown parameters by equating the corresponding sample and population moments. This method often provides estimators when other methods fail to do so or when estimators are harder to obtain, as in the case of a gamma distribution. Compared to other methods, method of moments estimators are easy to compute and have some desirable properties that we will discuss in ensuing sections. The drawback is that they are usually not the "best estimators" (to be defined later) available and sometimes may even be meaningless.

EXERCISES 5.2

5.2.1. Let $X_1, \ldots, X_n$ be a random sample of size n from the geometric distribution for which p is the probability of success.

(a) Use the method of moments to find a point estimator for p.
(b) Use the following data (simulated from geometric distribution) to find the moment estimator for p:

2	5	7	43	18	19	16	11	22
4	34	19	21	23	6	21	7	12

How will you use this information? [The pdf of a geometric distribution is $f(x) = p(1-p)^{x-1}$, for $x = 1, 2, \ldots$. Also $\mu = 1/p$.]

5.2.2. Let $X_1, \ldots, X_n$ be a random sample of size n from the exponential distribution whose pdf (by taking $\theta = 1/\beta$ in Definition 2.3.7) is

$$f(x, \theta) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

(a) Use the method of moments to find a point estimator for θ.
(b) The following data represent the time intervals between the emissions of beta particles.

0.9	0.1	0.1	0.8	0.9	0.1	0.1	0.7	1.0	0.2
0.1	0.1	0.1	2.3	0.8	0.3	0.2	0.1	1.0	0.9
0.1	0.5	0.4	0.6	0.2	0.4	0.2	0.1	0.8	0.2
0.5	3.0	1.0	0.5	0.2	2.0	1.7	0.1	0.3	0.1
0.4	0.5	0.8	0.1	0.1	1.7	0.1	0.2	0.3	0.1

Assuming the data follow an exponential distribution, obtain a moment estimate for the parameter θ. Interpret.

5.2.3. Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $(\theta - 1, \theta + 1)$.

(a) Find a moment estimator for θ.

(b) Use the following data to obtain a moment estimate for θ:

$$11.72 \quad 12.81 \quad 12.09 \quad 13.47 \quad 12.37$$

5.2.4. The probability density of a one-parameter Weibull distribution is given by

$$f(x) = \begin{cases} 2\alpha x e^{-\alpha x^2}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Using a random sample of size n, obtain a moment estimator for α.

(b) Assuming that the following data are from a one-parameter Weibull population,

$$\begin{array}{lllll} 1.87 & 1.60 & 2.36 & 1.12 & 0.15 \\ 1.83 & 0.64 & 1.53 & 0.73 & 2.26 \end{array}$$

obtain a moment estimate of α.

5.2.5. Let $X_1, \ldots, X_n$ be a random sample from the truncated exponential distribution with pdf

$$f(x) = \begin{cases} e^{-(x-\theta)}, & x \geq \theta \\ 0, & \text{otherwise.} \end{cases}$$

Find the method of moments estimate of θ.

5.2.6. Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf

$$f(x, \alpha) = \frac{1 + \alpha x}{2}, \quad -1 \leq x \leq 1, \text{ and } -1 \leq \alpha \leq 1.$$

Find the moment estimators for α.

5.2.7. Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f(x) = \begin{cases} \frac{2\alpha^2}{x^3}, & x \geq \alpha \\ 0, & \text{otherwise.} \end{cases}$$

Find a method of moments estimator for α.

5.2.8. Let $X_1, \ldots, X_n$ be a random sample from a negative binomial distribution with pmf

$$p(x, r, p) = \binom{x + r - 1}{r - 1} p^x (1 - p)^x, \; 0 \leq p \leq 1, \; x = 0, 1, 2, \ldots.$$

Find method of moments estimators for r and p. [Here $E[X] = r(1-p)/p$ and $E\left[X^2\right] = r(1-p)(r-rp+1)/p^2$.]

5.2.9. Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf

$$f(x) = \begin{cases} (\theta+1)\, x^{\theta}, & 0 \le x \le 1; \theta > -1 \\ 0, & \text{otherwise.} \end{cases}$$

Use the method of moments to obtain an estimator of θ.

5.2.10. Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf

$$f(x) = \begin{cases} \frac{2\beta - 2x}{\beta^2}, & 0 < x < \beta \\ 0, & \text{otherwise.} \end{cases}$$

Use the method of moments to obtain an estimator of β.

5.2.11. Let $X_1, \ldots, X_n$ be a random sample with common mean μ and variance σ^2. Obtain a method of moments estimator for σ.

5.2.12. Let $X_1, \ldots, X_n$ be a random sample from the beta distribution with parameters α and β. Find the method of moments estimator for α and β.

5.2.13. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with unknown mean μ and variance σ^2. Show that the method of moments estimators for μ and σ^2 are, respectively, the sample mean $\overline{X}$ and $S'^2 = (1/n)\sum_{i=1}^{n}(X - \overline{X})^2$. Note that $S'^2 = [(n-1)/n]\, S^2$ where S^2 is the sample variance.

5.3 THE METHOD OF MAXIMUM LIKELIHOOD

It is highly desirable to have a method that is generally applicable to the construction of statistical estimators that have "good" properties. In this section we present an important method for finding estimators of parameters proposed by geneticist/statistician Sir Ronald A. Fisher around 1922 called the method of maximum likelihood. Even though the method of moments is intuitive and easy to apply, it usually does not yield "good" estimators. The method of maximum likelihood is intuitively appealing, because we attempt to find the values of the true parameters that would have most likely produced the data that we in fact observed. For most cases of practical interest, the performance of maximum likelihood estimators is optimal for large enough data. This is one of the most versatile methods for fitting parametric statistical models to data. First, we define the concept of a likelihood function.

Definition 5.3.1 *Let $f(x_1, \ldots, x_n; \theta), \theta \in \Theta \subseteq \mathbb{R}^k$, be the joint probability (or density) function of n random variables $X_1, \ldots, X_n$ with sample values $x_1, \ldots, x_n$. The* **likelihood function** *of the sample is given by*

$$L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta), [= L(\theta), \textit{ in a briefer notation}].$$

We emphasize that L is a function of θ for fixed sample values.

If $X_1, \ldots, X_n$ are discrete iid random variables with probability function $p(x, \theta)$, then, the likelihood function is given by

$$\begin{aligned} L(\theta) &= P(X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} P(X_i = x_i), \quad \text{(by multiplication rule for independent random variables)} \\ &= \prod_{i=1}^{n} p(x_i, \theta) \end{aligned}$$

and in the continuous case, if the density is $f(x, \theta)$, then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

It is important to note that the likelihood function, although it depends on the observed sample values $x = (x_1, \ldots, x_n)$, is to be regarded as a function of the parameter θ. In the discrete case, $L(\theta; x_1, \ldots, x_n)$ gives the probability of observing $x = (x_1, \ldots, x_n)$, for a given θ. Thus, the likelihood function is a statistic, depending on the observed sample $x = (x_1, \ldots, x_n)$.

Example 5.3.1

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ random variables. Let $x_1, \ldots, x_n$ be the sample values. Find the likelihood function.

Solution

The density function for the normal variable is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. *Hence, the likelihood function is*

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

A statistical procedure should be consistent with the assumption that the best explanation of a set of data is provided by an estimator $\hat{\theta}$, which will be the value of the parameter θ that maximizes the likelihood function. This value of θ will be called the maximum likelihood estimator. The goal of maximum likelihood estimation is to find the parameter value(s) that makes the observed data most likely.

Definition 5.3.2 *The* **maximum likelihood estimators (MLEs)** *are those values of the parameters that maximize the likelihood function with respect to the parameter* θ. *That is,*

$$L\left(\hat{\theta}; x_1, \ldots, x_n\right) = \max_{\theta \in \Theta} L(\theta; x_1, \ldots, x_n)$$

where Θ *is the set of possible values of the parameter* θ.

The method of maximum likelihood extends to the case of several parameters. Let $X_1, \ldots, X_n$ be a random sample with joint pmf (if discrete) or pdf (if continuous)

$$L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m)$$

where the values of the parameters $\theta_1, \ldots, \theta_m$ are unknown and $x_1, \ldots, x_n$ are the observed sample values. Then, the maximum likelihood estimates $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are those values of the $\theta_i's$ that maximize the likelihood function, so that

$$f\left(x_1, \ldots, x_n; \hat{\theta}_1, \ldots, \hat{\theta}_m\right) \geq f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m)$$

for all allowable $\theta_1, \ldots, \theta_m$.

Note that the likelihood function conveys to us how feasible the observed sample is as a function of the possible parameter values. Maximum likelihood estimates give the parameter values for which the observed sample is most likely to have been generated. In general, the maximum likelihood method results in the problem of maximizing a function of single or several variables. Hence, in most situations, the methods of calculus can be used. In deriving the MLEs, however, there are situations where the techniques developed are more problem specific. Sometimes we need to use numerical methods, such as Newton's method.

In order to find a MLE, we need only to compute the likelihood function and then maximize that function with respect to the parameter of interest. In many cases, it is easier to work with the natural logarithm (ln) of the likelihood function, called the *log-likelihood function*. Because the natural logarithm function is increasing, the maximum value of the likelihood function, if it exists, will occur at the same point as the maximum value of the log-likelihood function. We now summarize the calculus-based procedure to find MLEs.

PROCEDURE TO FIND MLE

1. Define the likelihood function, $L(\theta)$.
2. Often it is easier to take the natural logarithm (ln) of $L(\theta)$.
3. When applicable, differentiate $\ln L(\theta)$ with respect to θ, and then equate the derivative to zero.
4. Solve for the parameter θ, and we will obtain $\hat{\theta}$.
5. Check whether it is a maximizer or global maximizer.

Example 5.3.2

Suppose $X_1, \ldots, X_n$ are a random sample from a geometric distribution with parameter $p, 0 \leq p \leq 1$. Find MLE $\hat{p}$.

Solution

For the geometric distribution, the pmf is given by

$$f(x, p) = p(1-p)^{x-1}, \quad 0 \leq p \leq 1, \quad x = 1, 2, 3, \ldots.$$

Hence, the likelihood function is

$$L(p) = \prod_{i=1}^{n} \left[p\,(1-p)^{x-1} \right] = p^n\,(1-p)^{-n+\sum_{i=1}^{n} x_i}.$$

Taking the natural logarithm of $L(p)$,

$$\ln L = n \ln p + \left(-n + \sum_{i=1}^{n} x_i \right) \ln (1-p).$$

Taking the derivative with respect to p, we have

$$\frac{d \ln L}{dp} = \frac{n}{p} - \frac{\left(-n + \sum_{i=1}^{n} x_i \right)}{(1-p)}.$$

Equating $\frac{d \ln L(p)}{dp}$ to zero, we have

$$\frac{n}{p} - \frac{\left(-n + \sum_{i=1}^{n} x_i \right)}{(1-p)} = 0.$$

Solving for p,

$$p = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{x}}.$$

Thus, we obtain a maximum likelihood estimator of p as

$$\hat{p} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}}.$$

We remark that $(1/\overline{X})$ is the maximum likelihood estimate of p. It can be shown that $\hat{p}$ is a global maximum. ■

Example 5.3.3

Suppose $X_1, \ldots, X_n$ are random samples from a Poisson distribution with parameter λ. Find MLE $\hat{\lambda}$.

Solution

We have the probability mass function

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots, \quad \lambda > 0.$$

Hence, the likelihood function is

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

Then, taking the natural logarithm, we have

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

and differentiating with respect to λ *results in*

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

and

$$\frac{d \ln L(\lambda)}{d\lambda} = 0, \text{ implies } \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0.$$

That is,

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}.$$

Hence, the MLE of λ *is*

$$\hat{\lambda} = \overline{X}.$$

■

It can be verified that the second derivative is negative and, hence, we really have a maximum.

Sometimes the method of derivatives cannot be used for finding the MLEs. For example, the likelihood is not differentiable in the range space. In this case, we need to make use of the special structures available in the specific situation to solve the problem. The following is one such case.

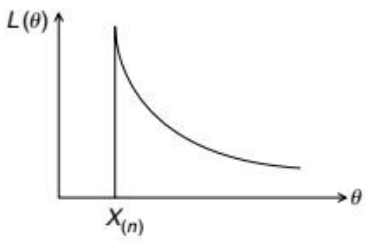
Example 5.3.4

Let $X_1, \ldots, X_n$ be a random sample from $U(0, \theta), \theta > 0$. Find the MLE of θ.

Solution

Note that the pdf of the uniform distribution is

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta \\ 0, & \text{otherwise.} \end{cases}$$



■ **FIGURE 5.1** Likelihood function for uniform probability distribution.

Hence, the likelihood function is given by

$$L(\theta, x_1, x_2, \ldots, x_n) = \begin{cases} \dfrac{1}{\theta^n}, & 0 \le x_1, x_2, \ldots, x_n \le \theta \\ 0, & \text{otherwise.} \end{cases}$$

When $\theta \ge \max(x_i)$, the likelihood is $(1/\theta^n)$, which is positive and decreasing as a function of θ (for fixed n). However, for $\theta < \max(x_i)$ the likelihood drops to 0, creating a discontinuity at the point $\max(x_i)$ (this is the minimum value of θ that can be chosen which still satisfies the condition $0 \le x_i \le \theta$), and Figure 5.1 shows that the maximum occurs at this point. Hence, we will not be able to find the derivative. Thus, the MLE is the largest order statistic,

$$\hat{\theta} = \max(X_i) = X_{(n)}.$$

■

In the previous example, because $E(X) = (\theta/2)$, we can see that $\theta = 2E(X)$. Hence, the method of moments estimator for θ is $\hat{\theta} = 2\overline{X}$. Sometimes the method of moments estimator can give meaningless results. To see this, suppose we observe values 3, 5, 6, and 18 from a $U(0, \theta)$ distribution. Clearly, the maximum likelihood estimate of θ is 18, whereas the method of moments estimate is 16, which is not quite acceptable, because we have already observed a value of 18.

As mentioned earlier, if the unknown parameter θ represents a vector of parameters, say $\theta = (\theta_1, \ldots, \theta_l)$, then the MLEs can be obtained from solutions of the system of equations

$$\frac{\partial}{\partial \theta} \ln L(\theta_1, \ldots, \theta_n) = 0, \quad \text{for} \quad i = 1, \ldots, l.$$

These are called the *maximum likelihood equations* and the solutions are denoted by $(\hat{\theta}_1, \ldots, \hat{\theta}_l)$.

Example 5.3.5

Let $X_1, \ldots, X_n$ be $N(\mu, \sigma^2)$.

(a) If μ is unknown and $\sigma^2 = \sigma_0^2$ is known, find the MLE for μ.

(b) If $\mu = \mu_0$ is known and σ^2 is unknown, find the MLE for σ^2.

(c) If μ and σ^2 are both unknown, find the MLE for $\theta = (\mu, \sigma^2)$.

Solution

In order to avoid notational confusion when taking the derivative, let $\theta = \sigma^2$. *Then, the likelihood function is*

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

or

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

(a) *When* $\theta = \theta_0 = \sigma_0^2$ *is known, the problem reduces to estimating the only one parameter,* μ. *Differentiating the log-likelihood function with respect to* μ,

$$\frac{\partial}{\partial\mu} \left(\ln L(\mu, \theta_0) \right) = \frac{2 \sum_{i=1}^{n} (x_i - \mu)}{2\theta_0}.$$

Setting the derivative equal to zero and solving for μ,

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

From this,

$$\sum_{i=1}^{n} x_i = n\mu \quad \text{or} \quad \mu = \bar{x}.$$

Thus, we get $\hat{\mu} = \overline{X}$.

(b) *When* $\mu = \mu_0$ *is known, the problem reduces to estimating the only one parameter,* $\sigma^2 = \theta$. *Differentiating the log-likelihood function with respect to* θ,

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

Setting the derivative equal to zero and solving for θ, *we get*

$$\hat{\theta} = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (X_i - \mu_0)^2}{n}.$$

(c) *When both μ and θ are unknown, we need to differentiate with respect to both μ and θ individually:*

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

and

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

Setting the derivatives equal to zero and solving simultaneously, we obtain

$$\hat{\mu} = \overline{X},$$

$$\hat{\sigma}^2 = \hat{\theta} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n} = S'^2.$$

Note that in (a) and (c), the estimates for μ are the same; however, in (b) and (c), the estimates for σ^2 are different.

■

At times, the maximum likelihood estimators may be hard to calculate. It may be necessary to use numerical methods to approximate values of the estimate. The following example gives one such case.

Example 5.3.6

Let $X_1, \ldots, X_n$ be a random sample from a population with gamma distribution and parameters α and β. Find MLEs for the unknown parameters α and β.

Solution

The pdf for the gamma distribution is given by

$$f(x) = \begin{cases} \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, & x > 0, \quad \alpha > 0, \quad \beta > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood function is given by

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

Taking the logarithms gives

$$\ln L = -n \ln \Gamma(\alpha) - n\alpha \ln \beta + (\alpha - 1)\sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} \frac{x}{\beta}.$$

Now taking the partial derivatives with respect to α and β and setting both equal to zero, we have

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

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

Solving the second one to get β in terms of α, we have

$$\beta = \frac{\overline{x}}{\alpha}.$$

Substituting this β in the first equation, we have to solve

$$-n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - n \ln \frac{\overline{x}}{\alpha} + \sum_{i=1}^{n} \ln x_i = 0$$

for $\alpha > 0$. There is no closed-form solution for α and β. In this case, one can use numerical methods such as the Newton--Raphson method to solve for α, and then use this value to find β.

There are many references available on the Web. Explaining the Newton–Raphson method, for instance, http://web.as.uky.edu/statistics/users/viele/sta601s08/nummax.pdf gives the algorithm for the gamma distribution.

In only a few cases are we able to obtain a simple form for the maximum likelihood equation that can be solved by setting the first derivative to zero. Often we cannot write an equation that can be differentiated to find the MLE parameter estimates. This is especially true in the situation where the model is complex and involves many parameters. Evaluating the likelihood exhaustively for all values of the parameters becomes almost impossible, even with modern computers. This is why so-called *optimization* algorithms have become indispensable to statisticians. The purpose of an optimization algorithm is to find as fast as possible the set of parameter values that make the observed data most likely. There are many such algorithms available. We describe the Newton–Raphson method in Project 5F, and another powerful algorithm, known as the EM algorithm, is given in Section 13.4.

Sometimes, it may be necessary to estimate a function of a parameter. The following invariance property of maximum likelihood estimators is very useful in those cases.

Theorem 5.3.1 *Let $h(\theta)$ be a one-to-one function of θ. If $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_l)$ is the MLE of $\theta = (\theta_1, \ldots, \theta_l)$, then the MLE of a function $h(\theta) = (h_1(\theta), \ldots, h_k(\theta))$ of these parameters is $h(\hat{\theta}) = (h_1(\hat{\theta}), \ldots, h_k(\hat{\theta}))$ for $1 \leq k \leq l$.*

As a consequence of the invariance property, in Example 5.3.5, we can obtain the estimator of the true standard deviation as $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{(1/n) \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$.

It is also known that, under very general conditions on the joint distribution of the sample and for a large sample size n, the MLE $\hat{\theta}$ is approximately the minimum variance unbiased estimator (this concept is introduced in the next section) of θ.

EXERCISES 5.3

5.3.1. Let $X_1, \ldots, X_n$ be a random sample recorded as heads or tails resulting from tossing a coin n times with unknown probability p of heads. Find the MLE $\hat{p}$ of p. Also using the invariance property, obtain an MLE for $q = 1 - p$. How would you use the results you have obtained?

5.3.2. Suppose $X_1, \ldots, X_n$ are a random sample from an exponential distribution with parameter θ. Find the MLE of $\hat{\theta}$. Also using the invariance property, obtain an MLE for the variance.

5.3.3. Let X be a random variable representing the time between successive arrivals at a checkout counter in a supermarket. The values of X in minutes (rounded to the nearest minute) are

1	2	3	7	11	4	13
12	7	3	2	11	7	2

Assume that the pdf of X is $f(x) = (1/\theta)e^{-(x/\theta)}$. Use these data to find MLE $\hat{\theta}$. How can you use this estimate you have just derived?

5.3.4. Let $X_1, \ldots, X_n$ be a random sample from the truncated exponential distribution with pdf

$$f(x) = \begin{cases} e^{-(x-\theta)}, & x \geq \theta \\ 0, & \text{otherwise.} \end{cases}$$

Show that the MLE of θ is $\min(X_i)$.

5.3.5. The pdf of a random variable X is given by

$$f(x) = \begin{cases} \frac{2x}{\alpha^2} e^{-x^2/\alpha^2}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Using a random sample of size n, obtain MLE $\hat{\alpha}$ for α.

5.3.6. The pdf of a random variable X is given by

$$P(X = n) = \frac{1}{n!} \exp\left(\alpha n - e^{\alpha}\right), \quad n = 0, 1, 2, \ldots.$$

Using a random sample of size n, obtain MLE $\hat{\alpha}$ for α.

5.3.7. Let $X_1, \ldots, X_n$ be a random sample from a two-parameter Weibull distribution with pdf

$$f(x) = \begin{cases} \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)^{\alpha}}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find the MLEs of α and β.

5.3.8. Let $X_1, \ldots, X_n$ be a random sample from a Rayleigh distribution with pdf

$$f(x) = \begin{cases} \frac{x}{\alpha} e^{-x^2/2\alpha}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find the MLEs of α.

5.3.9. Let $X_1, \ldots, X_n$ be a random sample from a two-parameter exponential population with density

$$f(x, \theta, \upsilon) = \frac{1}{\theta} e^{-\frac{(x-\upsilon)}{\theta}}, \quad \text{for} \quad x \geq \upsilon, \quad \theta > 0.$$

Find MLEs for θ and υ when both are unknown.

5.3.10. Let $X_1, \ldots, X_n$ be a random sample from the shifted exponential distribution with pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda(x-\theta)}, & x \geq \theta \\ 0, & \text{otherwise.} \end{cases}$$

Obtain the maximum likelihood estimators of θ and λ.

5.3.11. Let $X_1, \ldots, X_n$ be a random sample on $[0, 1]$ with pdf

$$f(x) = \frac{\Gamma(2\theta)}{\Gamma(\theta)^2} [x(1-x)]^{\theta-1}, \quad \theta > 0.$$

What equation does the maximum likelihood estimate of θ satisfy?

5.3.12. Let $X_1, \ldots, X_n$ be a random sample with pdf

$$f(x) = \begin{cases} (\alpha+1)x^{\alpha}, & 0 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find the MLE of α.

5.3.13. Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution with pdf

$$f(x) = \begin{cases} \frac{1}{3\theta+2}, & 0 \leq x \leq 3\theta + 2 \\ 0, & \text{otherwise.} \end{cases}$$

Obtain the MLE of θ.

5.3.14. Let $X_1, \ldots, X_n$ be a random sample from a Cauchy distribution with pdf

$$f(x) = \frac{1}{\pi \left[1 + (x-\beta)^2\right]}, \quad -\infty < x < \infty.$$

Find the MLE for β.

5.3.15. The following data represent the amount of leakage of a fluorescent dye from the bloodstream into the eye in patients with abnormal retinas:

1.6	1.4	1.2	2.2	1.8	1.7
1.8	6.3	2.4	2.3	18.9	22.8

Assuming that these data come from a normal distribution, find the maximum likelihood estimate of (μ, σ).

5.3.16. Let $X_1, \ldots, X_n$ be a random sample from a population with gamma distribution and parameters α and β. Show that the MLE of $\mu = \alpha\beta$ is the sample mean $\hat{\mu} = \overline{X}$.

5.3.17. The lifetimes X of a certain brand of component used in a machine can be modeled as a random variable with pdf $f(x) = (1/\theta)\, e^{-(x/\theta)}$. The reliability $R(x)$ of the component is defined as $R(x) = 1 - F(x)$. Suppose $X_1, X_2, \ldots, X_n$ are the lifetimes of n components randomly selected and tested. Find the MLE of $R(x)$.

5.3.18. Using the method explained in Project 4A, generate 20 observations of a random variable having an exponential distribution with mean and standard deviation both equal to 2. What is the maximum likelihood estimate of the population mean? How much is the observed error?

5.3.19. Let $X_1, \ldots, X_n$ be a random sample from a Pareto distribution (named after the economist Vilfredo Pareto) with shape parameter a. The density function is given by

$$f(x) = \begin{cases} \dfrac{a}{x^{a+1}}, & x \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

(The Pareto distribution is a skewed, heavy-tailed distribution. Sometimes it is used to model the distribution of incomes.) Show that the maximum likelihood estimator of a is

$$\hat{a} = \frac{n}{\sum_{i=1}^{n} \ln (X_i)}.$$

5.3.20. Let $X_1, \ldots, X_n$ be a random sample from $N(\theta, \theta)$, $0 < \theta < \infty$. Find the maximum likelihood estimate of θ.

5.4 SOME DESIRABLE PROPERTIES OF POINT ESTIMATORS

Two different methods of finding estimators for population parameters have been introduced in the preceding sections. We have seen that it is possible to have several estimators for the same parameter. For a practitioner of statistics, an important question is going to be which of many available sample statistics, such as mean, median, smallest observation, or largest observation, should be chosen to represent all of the sample? Should we use the method of moments estimator, the maximum

likelihood estimator, or an estimator obtained through some other method of least squares (we will see this method in Chapter 8)? Now we introduce some common ways to distinguish between them by looking at some desirable properties of these estimators.

5.4.1 Unbiased Estimators

It is desirable to have the property that the expected value of an estimator of a parameter is equal to the true value of the parameter. Such estimators are called unbiased estimators.

Definition 5.4.1 *A point estimator $\hat{\theta}$ is called an* **unbiased estimator** *of the parameter θ if $E(\hat{\theta}) = \theta$ for all possible values of θ. Otherwise $\hat{\theta}$ is said to be* **biased**. *Furthermore, the* **bias** *of $\hat{\theta}$ is given by*

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

Note that the bias is nothing but the expected value of the (random) error, $E(\hat{\theta} - \theta)$. Thus, the estimator is unbiased if the bias is 0 for all values of θ. The bias occurs when a sample does not accurately represent the population from which the sample is taken. It is important to observe that in order to check whether $\hat{\theta}$ is unbiased, it is not necessary to know the value of the true parameter. Instead, one can use the sampling distribution of $\hat{\theta}$. We demonstrate the basic procedure through the following example.

Example 5.4.1

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli population with parameter p. Show that the method of moments estimator is also an unbiased estimator.

Solution

We can verify that the moment estimator of p is

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

Because for binomial random variables, $E(Y) = np$, it follows that

$$E(\hat{p}) = E\left(\frac{Y}{n}\right) = \frac{1}{n}E(Y) = \frac{1}{n} \cdot np = p.$$

Hence, $\hat{p} = Y/n$ is an unbiased estimator for p.

In fact, we have the following result, which states that the sample mean is always an unbiased estimator of the population mean.

Theorem 5.4.1 *The mean of a random sample $\overline{X}$ is an unbiased estimator of the population mean μ.*

Proof. Let $X_1, \ldots, X_n$ be random variables with mean μ. Then, the sample mean is $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$.

$$E\overline{X} = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n} \cdot n\mu = \mu.$$

Hence, $\overline{X}$ is an unbiased estimator of μ. □

How is this interpreted in practice? Suppose that a data set is collected with n numerical observations $x_1, \ldots, x_n$. The resulting sample mean may be either less than or greater than the true population mean, μ (remember, we do not know this value). If the sampling experiment was repeated many times, then the average of the estimates calculated over these repetitions of the sampling experiment will equal the true population mean.

If we have to choose among several different estimators of a parameter θ, it is desirable to select one that is unbiased. The following result states that the sample variance $S^2 = (1/n-1)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is an unbiased estimator of the population variance σ^2. This is one of the reasons why in the definition of the sample variance, instead of dividing by n, we divide by $(n-1)$.

Theorem 5.4.2 *If S^2 is the variance of a random sample from an infinite population with finite variance σ^2, then S^2 is an unbiased estimator for σ^2.*

Proof. Let $X_1, \ldots, X_n$ be iid random variables with variance $\sigma^2 < \infty$. We have

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

Because $E\{(X_i - \mu)^2\} = \sigma^2$ and $E\{\left(\overline{X} - \mu\right)^2\} = \sigma^2/n$, it follows that

$$E\left(S^2\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n}\sigma^2 - n\frac{\sigma^2}{n}\right] = \sigma^2.$$

Hence, S^2 is an unbiased estimator of σ^2. □

It is important to observe the following:

1. S^2 is not an unbiased estimator of the variance of a finite population.
2. Unbiasedness may not be retained under functional transformations, that is; if $\hat{\theta}$ is an unbiased estimator of θ, it does not follow that $f(\hat{\theta})$ is an unbiased estimator of $f(\theta)$.
3. Maximum likelihood estimators or moment estimators are not, in general, unbiased.
4. In many cases it is possible to alter a biased estimator by multiplying by an appropriate constant to obtain an unbiased estimator.

The following example will show that unbiased estimators need not be unique.

Example 5.4.2

Let $X_1, \ldots, X_n$ be a random sample from a population with finite mean μ. Show that the sample mean $\overline{X}$ and $\frac{1}{3}\overline{X} + \frac{2}{3}X_1$ are both unbiased estimators of μ.

Solution

By Theorem 1, $\overline{X}$ is unbiased. Now

$$E\left[\tfrac{1}{3}\overline{X} + \tfrac{2}{3}X_1\right] = \tfrac{1}{3}\mu + \tfrac{2}{3}\mu = \mu.$$

Hence, $\frac{1}{3}\overline{X} + \frac{2}{3}X_1$ is also an unbiased estimator of μ.

How many unbiased estimators can we find? In fact, the following example shows that if we have two unbiased estimators, there are infinitely many unbiased estimators.

Example 5.4.3

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators of θ. Show that

$$\hat{\theta}_3 = a\hat{\theta}_1 + (1-a)\,\hat{\theta}_2,\ 0 \le a \le 1$$

is an unbiased estimator of θ. Note that $\hat{\theta}_3$ is a convex combination of $\hat{\theta}_1$ and $\hat{\theta}_2$. In addition, assume that $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent, and $Var(\hat{\theta}_1) = \sigma_1^2$ and $Var(\hat{\theta}_2) = \sigma_2^2$. How should the constant a be chosen in order to minimize the variance of $\hat{\theta}_3$?

Solution

We are given that $E(\hat{\theta}_1) = \theta$ and $E(\hat{\theta}_2) = \theta$. Therefore,

$$\begin{aligned} E\left(\hat{\theta}_3\right) &= E\left[a\hat{\theta}_1 + (1-a)\,\hat{\theta}_2\right] = aE\hat{\theta}_1 + (1-a)\,E\hat{\theta}_2 \\ &= a\theta + (1-a)\,\theta = \theta. \end{aligned}$$

Hence $\hat{\theta}_3$ is unbiased. By independence,

$$\begin{aligned} Var\left(\hat{\theta}_3\right) &= Var\left[a\hat{\theta}_1 + (1-a)\,\hat{\theta}_2\right] \\ &= a^2 Var\left(\hat{\theta}_1\right) + (1-a)^2\, Var\left(\hat{\theta}_2\right) \\ &= a^2\sigma_1^2 + (1-a)^2\,\sigma_2^2. \end{aligned}$$

To find the minimum,

$$\frac{d}{da} Var\left(\hat{\theta}_3\right) = 2a\sigma_1^2 - 2(1-a)\sigma_2^2 = 0,$$

gives us

$$a = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Because $\frac{d^2}{da^2} V(\hat{\theta}_3) = 2\sigma_1^2 + 2\sigma_2^2 > 0$, $V(\hat{\theta}_3)$ *has a minimum at this value of* $'a'$. *Thus, if* $\sigma_1^2 = \sigma_2^2$, *then* $a = 1/2$.

■

Example 5.4.4

Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Show that the method of moments estimator for the population parameter β is unbiased.

Solution

From Section 5.2, we have seen that the method of moments estimator for β *is the sample mean* $\overline{X}$, *and the population mean is* β. *Because* $E(\overline{X}) = \mu = \beta$, *the method of moments estimator for the population parameter* β *is unbiased.*

■

As we have seen, there can be many unbiased estimators of a parameter θ. Which one of these estimators can we choose? If we have to choose an unbiased estimator, it will be desirable to choose the one with the least variance. If an estimator is biased, then we should prefer the one with low bias as well as low variance. Generally, it is better to have an estimator that has low bias as well as low variance. This leads us to the following definition.

Definition 5.4.2 *The* **mean square error** *of the estimator* $\hat{\theta}$, *denoted by* $MSE(\hat{\theta})$, *is defined as*

$$MSE\left(\hat{\theta}\right) = E\left(\hat{\theta} - \theta\right)^2.$$

Through the following calculations, we will now show that the MSE is a measure that combines both bias and variance.

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

because letting $B = E(\hat{\theta}) - \theta$, we get

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + B^2.$$

B is called the *bias* of the estimator. Also, $E(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) = 0$.

Because the bias is zero for unbiased estimators, it is clear that $MSE(\hat{\theta}) = Var(\hat{\theta})$. Mean square error measures, on average, how close an estimator comes to the true value of the parameter. Hence, this could be used as a criterion for determining when one estimator is "better" than another. However, in general, it is difficult to find $\hat{\theta}$ to minimize $MSE(\hat{\theta})$. For this reason, most of the time, we look only at unbiased estimators in order to minimize $Var(\hat{\theta})$. This leads to the following definition.

Definition 5.4.3 *The unbiased estimator $\hat{\theta}$ that minimizes the mean square error is called the* **minimum variance unbiased estimator (MVUE)** of θ.

Example 5.4.5

Let X_1, X_2, X_3 be a sample of size $n = 3$ from a distribution with unknown mean μ, $-\infty < \mu < \infty$, where the variance σ^2 is a known positive number. Show that both $\hat{\theta}_1 = \overline{X}$ and $\hat{\theta}_2 = [(2X_1 + X_2 + 5X_3)/8]$ are unbiased estimators for μ. Compare the variances of $\hat{\theta}_1$ and $\hat{\theta}_2$.

Solution

We have

$$E(\hat{\theta}_1) = E(\overline{X}) = \frac{1}{3} \cdot 3\mu = \mu,$$

and

$$E(\hat{\theta}_2) = \frac{1}{8}[2EX_1 + EX_2 + 5EX_3]$$
$$= \frac{1}{8}[2\mu + \mu + 5\mu] = \mu.$$

Hence, both $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased estimators.
However,

$$Var(\hat{\theta}_1) = \frac{\sigma^2}{3},$$

whereas

$$Var(\hat{\theta}_2) = Var\left(\frac{2X_1 + X_2 + 5X_3}{8}\right)$$
$$= \frac{4}{64}\sigma^2 + \frac{1}{64}\sigma^2 + \frac{25}{64}\sigma^2 = \frac{30}{64}\sigma^2.$$

Because $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$, we see that $\overline{X}$ is a better unbiased estimator in the sense that the variance of $\overline{X}$ is smaller.

■

It is important to observe that the maximum likelihood estimators are not always unbiased, but it can be shown that for such estimators the bias goes to zero as the sample size increases.

5.4.2 Sufficiency

In the statistical inference problems on a parameter, one of the major questions is: Can a specific statistic replace the entire data without losing pertinent information? Suppose $X_1, \ldots, X_n$ is a random sample from a probability distribution with unknown parameter θ. In general, statisticians look for ways of reducing a set of data so that these data can be more easily understood without losing the meaning associated with the entire collection of observations. Intuitively, a statistic U is a sufficient statistic for a parameter θ if U contains all the information available in the data about the value of θ. For example, the sample mean may contain all the relevant information about the parameter μ, and in that case $U = \overline{X}$ is called a sufficient statistic for μ. An estimator that is a function of a sufficient statistic can be deemed to be a "good" estimator, because it depends on fewer data values. When we have a sufficient statistic U for θ, we need to concentrate only on U because it exhausts all the information that the sample has about θ. That is, knowledge of the actual n observations does not contribute anything more to the inference about θ.

Definition 5.4.4 *Let $X_1, \ldots, X_n$ be a random sample from a probability distribution with unknown parameter θ. Then, the statistic $U = g(X_1, \ldots, X_n)$ is said to be* **sufficient** for θ if the conditional pdf or pf of $X_1, \ldots, X_n$ given $U = u$ does not depend on θ for any value of u. An estimator of θ that is a function of a sufficient statistic for θ is said to be a **sufficient estimator** of θ.

Example 5.4.6

Let $X_1, \ldots, X_n$ be iid Bernoulli random variables with parameter θ. Show that $U = \sum_{i=1}^{n} X_i$ is sufficient for θ.

Solution

The joint probability mass function of $X_1, \ldots, X_n$ is

$$f(X_1, \ldots, X_n; \theta) = \theta^{\sum_{i=1}^{n} X_i} (1-\theta)^{n-\sum_{i=1}^{n} X_i}, \quad 0 \le \theta \le 1.$$

Because $U = \sum_{i=1}^{n} X_i$ we have

$$f(X_1, \ldots, X_n; \theta) = \theta^{U} (1-\theta)^{n-U}, \quad 0 \le U \le n.$$

Also, because $U \sim B(n, \theta)$, we have

$$f(u; \theta) = \binom{n}{u} \theta^{U} (1-\theta)^{n-U}.$$

Also,

$$f(x_1, \ldots, x_n \,|U = u) = \frac{f(x_1, \ldots, x_n, u)}{f_U(u)} = \begin{cases} \frac{f(x_1, \ldots, x_n)}{f_U(u)}, & u = \sum x_i \\ 0, & \text{otherwise.} \end{cases}$$

Therefore,

$$f(x_1, \ldots, x_n | U = u) = \begin{cases} \dfrac{\theta^u (1-\theta)^{n-u}}{\binom{n}{u} \theta^u (1-\theta)^{n-u}} = \dfrac{1}{\binom{n}{u}} & \text{if } u = \sum x_i \\ 0, & \text{otherwise.} \end{cases}$$

which is independent of θ. Therefore U is sufficient for θ.

Example 5.4.7

Let $X_1, \ldots, X_n$ be a random sample from $U(0, \theta)$. That is,

$$f(x) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 < x < \theta \\ 0, & \text{otherwise.} \end{cases}$$

Show that $U = \max_{1 \le i \le n} X$ is sufficient for θ.

Solution

The joint density or the likelihood function is given by

$$f(x_1, \ldots, x_n; \theta) = \begin{cases} \frac{1}{\theta^n}, & \text{if } 0 < x_1, \ldots, x_n < \theta \\ 0, & \text{otherwise.} \end{cases}$$

The joint pdf $f(x_1, \ldots, x_n; \theta)$ can be equivalently written as

$$f(x_1, \ldots, x_n; \theta) = \begin{cases} \frac{1}{\theta^n}, & \text{if } x_{\min} > 0,\ x_{\max} < \theta \\ 0, & \text{otherwise.} \end{cases}$$

Now, we can compute the pdf of U.

$$\begin{aligned} F(u) &= P(U \le u) = P(X_1, \ldots, X_n \le u) \\ &= \prod_{i=1}^{n} P(X_i \le u) \text{ (because of independence)} \\ &= \prod_{i=1}^{n} \left(\int_0^u \frac{1}{\theta} dx \right) = \frac{u^n}{\theta^n}, \ 0 < u < \theta. \end{aligned}$$

The pdf of U may now be obtained as

$$f(u) = \frac{d}{du} F(u) = \frac{n u^{n-1}}{\theta^n}, \quad 0 < u < \theta$$

Moreover,

$$f(x_1, \ldots, x_n \,|u) = \begin{cases} \frac{f(x_1,\ldots,x_n,u)}{f_U(u)} = \frac{f(x_1,\ldots,x_n)}{f_U(u)}, & \text{if } u = x_{\max} \text{ and } x_{\min} > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Using the expressions for $f(x_1, \ldots, x_n)$ *and* $f_U(u)$ *we obtain*

$$f(x_1, \ldots, x_n \,|u = u) = \begin{cases} \frac{1/\theta^n}{nu^{n-1}/\theta^n} = \frac{1}{nu^{n-1}}, & \text{if } u = x_{\max} \text{ and } x_{\min} > 0 \\ 0, & \text{otherwise} \end{cases}$$

$f(X_1, \ldots, X_n \,|U)$ *is a function of* u *and* $x_{\min}$ *which is independent of* θ. *Hence,* $U = \max_{1 \le i \le n} X_i$ *is sufficient for* θ.

■

The outcome $X_1, \ldots, X_n$ is always sufficient, but we will exclude this trivial statistic from consideration. In the previous two examples, we were given a statistic and asked to check whether it was sufficient. It can often be tedious to check whether a statistic is sufficient for a given parameter based directly on the foregoing definition. If the form of the statistic is not given, how do we guess what is the sufficient statistic? Now think of working out the conditional probability by hand for each of our guesses! In general, this will be a tedious way to go about finding sufficient statistics. Fortunately, the Neyman–Fisher factorization theorem makes it easier to spot a sufficient statistic. The following result will give us a convenient way of verifying sufficiency of a statistic through the likelihood function.

NEYMAN–FISHER FACTORIZATION CRITERIA

Theorem 5.4.3 *Let* U *be a statistic based on the random sample* $X_1, \ldots, X_n$. *Then,* U *is a sufficient statistic for* θ *if and only if the joint pdf (or pf)* $f(x_1, \ldots, x_n; \theta)$ *(which depends on the parameter* θ*) can be factored into two nonnegative functions.*

$$f(x_1, \ldots, x_n; \theta) = g(u, \theta)\, h(x_1, \ldots, x_n), \quad \text{for all } x_1, \ldots, x_n,$$

where $g(u, \theta)$ *is a function only of* u *and* θ *and* $h(x_1, \ldots, x_n)$ *is a function of only* $x_1, \ldots, x_n$ *and not of* θ.

Proof. (Discrete case.) We will only give the proof in the discrete case, even though the result is also true for the continuous case. First suppose that $U(X_1, \ldots, X_n)$ is sufficient for θ. Then, $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ if and only if $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ and $U(X_1, \ldots, X_n) = U(x_1, \ldots, x_n) = u(say)$. Therefore

$$\begin{aligned} f(x_1, \ldots, x_n; \theta) &= P_\theta\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \text{ and } U = u\right) \\ &= P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|U = u)\, P_\theta(U = u). \end{aligned}$$

□

Because U is assumed to be sufficient for θ, the conditional probability $P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|U = u)$ does not depend on θ. Let us denote this conditional probability by $h(x_1, \ldots, x_n)$. Clearly $P_\theta(U = u)$ is a function of u and θ. Let us denote this by $g(u, \theta)$.

It now follows from the equation above that

$$f(x_1, \ldots, x_n; \theta) = g(u, \theta)\, h(x_1, \ldots, x_n)$$

as was to be shown.

To prove the converse, assume that

$$f(x_1, \ldots, x_n; \theta) = g(u, \theta)\, h(x_1, \ldots, x_n).$$

Define the set A_u by

$$A_u = \{(x_1, \ldots, x_n) : U(x_1, \ldots, x_n) = u\}.$$

That is, A_u is the set of all $(x_1, \ldots, x_n)$ such that U maps it into u. We note that A_u does not depend on θ. Now

$$\begin{aligned} &P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|U = u) \\ &\quad = \frac{P_\theta\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \text{ and } U = u\right)}{P_\theta(U = u)} \\ &\quad = \begin{cases} \frac{P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \text{ and } U = u)}{P_\theta(U = u)}, & \text{if } (x_1, \ldots, x_n) \in A_u \\ 0, & \text{if } (x_1, \ldots, x_n) \notin A_u. \end{cases} \end{aligned}$$

If $(x_1, \ldots, x_n) \notin A_u$, then, clearly,

$$f(x_1, \ldots, x_n; \theta) = P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|U = u)$$

which is independent of θ.

If $(x_1, \ldots, x_n) \in A_u$, then, using the factorization criterion, we obtain

$$\begin{aligned} &P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|U = u) \\ &\quad = \frac{P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{P_\theta(U = u)} \\ &\quad = \frac{f(x_1, \ldots, x_n; \theta)}{P_\theta(U = u)} = \frac{g(u, \theta)\, h(x_1, \ldots, x_n)}{\sum\limits_{(x_1, \ldots, x_n) \in A_u} g(u, \theta)\, h(x_1, \ldots, x_n)} \\ &\quad = \frac{g(u, \theta)\, h(x_1, \ldots, x_n)}{g(u, \theta) \sum\limits_{(x_1, \ldots, x_n) \in A_u} h(x_1, \ldots, x_n)} = \frac{h(x_1, \ldots, x_n)}{\sum\limits_{(x_1, \ldots, x_n) \in A_u} h(x_1, \ldots, x_n)} \end{aligned}$$

Therefore, the conditional distribution of $X_1, \ldots, X_n$ given U does not depend on θ, proving that U is sufficient.

One can use the following procedure to verify that a given statistic is sufficient. This procedure is based on factorization criteria rather than using the definition of sufficiency directly.

PROCEDURE TO VERIFY SUFFICIENCY

1. Obtain the joint pdf or pf $f_\theta(x_1, \ldots, x_n)$.
2. If necessary, rewrite the joint pdf or pf in terms of the given statistic and parameter so that one can use the factorization theorem.
3. Define the functions g and h, in such a way that g is a function of the statistic and parameter only and h is a function of the observations only.
4. If step 3 is possible, then the statistic is sufficient. Otherwise, it is not sufficient.

In general, it is not easy to use the factorization criterion to show that a statistic U is *not* sufficient. We now give some examples using the factorization theorem.

Example 5.4.8

Let $X_1, \ldots, X_n$ denote a random sample from a geometric population with parameter p. Show that $\overline{X}$ is sufficient for p.

Solution

For the geometric distribution, the pf is given by

$$f(x, p) = \begin{cases} p(1-p)^{x-1}, & x \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Hence, the joint pf is

$$f(x_1, \ldots, x_n; p) = p^n (1-p)^{-n+\sum_{i=1}^{n} x_i}$$

$$= \begin{cases} p^n (1-p)^{n\overline{x}-n}, & \text{if } x_1, \ldots, x_n \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Take,

$$g(\overline{x}, p) = p^n (1-p)^{n\overline{x}-n} \quad \text{and} \quad h(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } x_i \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, $\overline{X}$ is sufficient for p.

■

Example 5.4.9

Let $X_1, \ldots, X_n$ denote a random sample from a $U(0, \theta)$ with pdf

$$f_\theta(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta, \quad \theta > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Show that $X_{(n)} = \max_{1 \le i \le n} X_i$ is sufficient for θ, using the factorization theorem.

Solution

The likelihood function of the sample is

$$f_\theta(x_1, \ldots, x_n) = \begin{cases} \frac{1}{\theta^n}, & \text{if } 0 < x_1, \ldots, x_n < \theta, \\ 0, & \text{otherwise.} \end{cases}$$

We can now write $f_\theta(x_1, \ldots, x_n)$ *as*

$$f_\theta(x_1, \ldots, x_n) = h(x_1, \ldots, x_n)\, g\left(\theta, x_{(n)}\right), \text{ for all } x_1, \ldots, x_n$$

where

$$h(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } x_1, \ldots, x_n > 0 \\ 0, & \text{otherwise} \end{cases}$$

and

$$g\left(\theta; x_{(n)}\right) = \begin{cases} \frac{1}{\theta^n}, & \text{if } 0 < x_{(n)} < \theta, \\ 0, & \text{otherwise.} \end{cases}$$

From the factorization theorem, we now conclude that $X_{(n)}$ *is sufficient for* θ. *In the next definition, we introduce the concept of joint sufficiency.*

■

Definition 5.4.5 *Two statistics* U_1 *and* U_2 *are said to be* **jointly sufficient** *for the parameters* θ_1 *and* θ_2 *if the conditional distribution of* $X_1, \ldots, X_n$ *given* U_1 *and* U_2 *does not depend on* θ_1 *or* θ_2. *In general, the statistic* $U = (U_1, \ldots, U_n)$ *is jointly sufficient for* $\theta = (\theta_1, \ldots, \theta_n)$ *if the conditional distribution of* $X_1, \ldots, X_n$ *given* U *is free of* θ.

Now we state the factorization criteria for joint sufficiency analogous to the single population parameter case.

THE FACTORIZATION CRITERIA FOR JOINT SUFFICIENCY

Theorem 5.4.4 *The two statistics U_1 and U_2 are jointly sufficient for θ_1 and θ_2 if and only if the likelihood function can be factored into two non-negative functions,*

$$f(x_1, \ldots, x_n; \theta_1, \theta_2) = g(u_1, u_2; \theta_1, \theta_2)\, h(x_1, \ldots, x_n)$$

where $g(u_1, u_2; \theta_1, \theta_2)$ is only a function of $u_1, u_2; \theta_1$ and θ_2, and $h(x_1, x_n)$ is free of θ_1 or θ_2.

Example 5.4.10

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$.

(a) If μ is unknown and $\sigma^2 = \sigma_0^2$ is known, show that $\overline{X}$ is a sufficient statistic for μ.

(b) If $\mu = \mu_0$ is known and σ^2 is unknown, show that $\sum_{i=1}^{n} (X_i - \mu_0)^2$ is sufficient for σ^2.

(c) If μ and σ^2 are both unknown, show that $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$ are jointly sufficient for μ and σ^2.

Solution

The likelihood function of the sample is

$$L = \frac{1}{(2\pi)^{n/2}\,\sigma^n} \exp\left[-\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{2\sigma^2}\right]$$

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

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

(a) *When $\sigma^2 = \sigma_0^2$ is known, use the factorization criteria, with*

$$g(\overline{x}, \mu) = \exp\left(\frac{2n\mu\overline{x} - n\mu^2}{2\sigma_0^2}\right)$$

and

$$h(x_1, \ldots, x_n) = (2\pi)^{-n/2}\,\sigma^{-n} \exp\left(-\frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2}\right).$$

Therefore, $\overline{X}$ is sufficient for μ.

(b) *When* $\mu = \mu_0$ *is known, let*

$$g\left(\sum_{i=1}^{n}(X_i-\mu)^2, \sigma^2\right) = \sigma^{-n}\exp\left|-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right|$$

and

$$h(x_1,\ldots,x_n) = \frac{1}{(2\pi)^{n/2}}.$$

Thus, $\sum_{i=1}^{n}(X_i-\mu)^2$ *is sufficient for* σ^2.

(c) *When both* μ *and* σ^2 *are unknown, use*

$$g\left(\sum_{i=1}^{n}x_i, \sum_{i=1}^{n}x_i^2, \mu, \sigma^2\right) = \sigma^{-n}\exp\left|-\frac{\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2}{2\sigma^2}\right|$$

and

$$h(x_1,\ldots,x_n) = \frac{1}{(2\pi)^{n/2}}.$$

Hence, $\sum_{i=1}^{n}X_i$ *and* $\sum_{i=1}^{n}X_i^2$ *are jointly sufficient for* μ *and* σ^2.

Example 5.4.11

Suppose that we have a random sample $X_1,\ldots,X_n$ from a discrete distribution given by

$$f_\theta(x) = C(\theta)\,2^{-x/\theta}, \quad x = \theta, \theta+1, \theta+2, \ldots; \quad \theta > 0$$

where $C(\theta) > 0$ is a normalizing constant. Using the factorization theorem, find a sufficient statistic for θ.

Solution

The joint density function $f(x_1,\ldots,x_n;\theta)$ *of the sample* $X_1,\ldots,X_n$ *is*

$$f(x_1,\ldots,x_n;\theta) = \begin{cases} C(\theta)\,2^{-\sum_{i=1}^{n}(x_i/\theta)}, & x_1, x_2, \ldots, x_n \text{ are integers } \geq \theta \\ 0, & \text{otherwise}. \end{cases}$$

The function $f(x_1,\ldots,x_n;\theta)$ *can be written as*

$$f(x_1,\ldots,x_n;\theta) = h(x_1,\ldots,x_n)\,C(\theta)\,2^{-\sum_{i=1}^{n}(x_i/\theta)}\,g_1\left(\theta, x_{(1)}\right)$$

where $x_{(1)} = \min_i (x_1, \ldots, x_n)$, *and*

$$h(x_1, x_2, \ldots, x_n) = \begin{cases} 1, & \text{if } x_j - x_{(1)} \geq 0 \text{ is an integer for } j = 1, 2, \ldots, n \\ 0, & \text{otherwise} \end{cases}$$

and

$$g_1(\theta, x_{(1)}) = \begin{cases} 1, & \text{if } x_{(1)} \geq \theta \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$f(x_1, \ldots, x_n; \theta) = h(x_1, \ldots, x_n)\, g\left(\theta, \sum x_i, x_{(1)}\right)$$

■

where $g\left(\theta, \sum x_i, x_{(1)}\right) = C(\theta) 2^{-\sum_{i=1}^{n} (x_i/\theta)} g_1\left(\theta, x_{(1)}\right)$. Using the factorization theorem, we conclude that $\left(\sum x_i,\ x_{(1)}\right)$ is jointly sufficient for θ. This result shows that even for a single parameter, we may need more than one statistic for sufficiency.

When using the factorization criterion, one has to be careful in cases where the range space depends on the parameter.

Using the factorization criterion, we can prove the following result, which says that if we have a unique maximum likelihood estimator, then that estimator will be a function of the sufficient statistic.

Theorem 5.4.5 *If U is a sufficient statistic for θ, the maximum likelihood estimator of θ, if unique, is a function of U.*

Proof. Because U is sufficient, by Theorem 5.4.1, the joint pdf can be factored as

$$f(x_1, \ldots, x_n; \theta) = g(u, \theta)\, h(x_1, \ldots, x_n).$$

This depends on θ only through the statistic U. To maximize L we need to maximize $g(U, \theta)$. □

Many common distributions such as Poisson, normal, gamma, and Bernoulli are members of the *exponential family* of probability distributions. The exponential family of distributions has density functions of the form

$$f(x; \theta) = \begin{cases} \exp\left[k(x)c(\theta) + S(x) + d(\theta)\right], & \text{if } x \in B \\ 0, & x \notin B \end{cases}$$

where B does not depend on the parameter θ.

Example 5.4.12
Write the following in exponential form.

(a) $\dfrac{e^{-\lambda}\lambda^x}{x!}$

(b) $p^x(1-p)^{1-x}$

(c) $\dfrac{1}{\sqrt{2\pi}}e^{-(x-\mu)^2/2}$

Solution

(a) *We have*

$$\frac{e^{-\lambda}\lambda^x}{x!} = \exp\left[x\ln\lambda - \ln x! - \lambda\right].$$

Here $k(x) = x$, $c(\lambda) = \ln\lambda$, $S(x) = -\ln(x!)$, *and* $d(\lambda) = -\lambda$.

(b) *Similarly,*

$$p^x(1-p)^{1-x} = \exp\left[x\ln\left(\frac{p}{1-p}\right) + \ln(1-p)\right], \quad x = 0 \text{ or } 1.$$

(c) *This is the standard normal density.*

$$\frac{1}{\sqrt{2\pi}}e^{-(x-\mu)^2/2} = \exp\left[x\mu - \frac{x^2}{2} - \frac{\mu^2}{2} - \frac{1}{2}\ln(2\pi)\right], \quad -\infty < x < \infty.$$

Note that in the previous example, for each of the cases, $\sum_{i=1}^n X_i$ is a sufficient statistic for the parameter. In the next result, we give a generalization of this fact.

Theorem 5.4.6 *Let* $X_1, \ldots, X_n$ *be a random sample from a population with pdf or pmf of the exponential form*

$$f(x;\theta) = \begin{cases} \exp\left[k(x)c(\theta) + S(x) + d(\theta)\right], & \text{if } x \in B \\ 0, & x \notin B \end{cases}$$

where B *does not depend on the parameter* θ. *The statistic* $\sum_{i=1}^n k(X_i)$ *is sufficient for* θ.

Proof. The joint density

$$f(x_1, \ldots, x_n;\theta) = \exp\left[c(\theta)\sum_{i=1}^n k(x_i) + \sum_{i=1}^n S(x_i) + nd(\theta)\right]$$

$$= \left\{\exp\left[c(\theta)\sum_{i=1}^n k(x_i) + nd(\theta)\right]\right\}\left\{\exp\left[\sum_{i=1}^n S(x_i)\right]\right\}.$$

Using the factorization theorem, the statistic $\sum_{i=1}^n k(X_i)$ is sufficient. □

It does not follow that every function of a sufficient statistic is sufficient. However, any one-to-one function of a sufficient statistic is also sufficient. Every statistic need not be sufficient. When they do exist, sufficient estimators are very important, because if one can find a sufficient estimator it is ordinarily possible to find an unbiased estimator based on the sufficient statistic. Actually, the following theorem shows that if one is searching for an unbiased estimator with minimal variance, it has to be restricted to functions of a sufficient statistics.

RAO–BLACKWELL THEOREM

Theorem 5.4.7 *Let $X_1, \ldots, X_n$ be a random sample with joint pf or pdf $f(x_1, \ldots, x_n; \theta)$ and let $U = (U_1, \ldots, U_n)$ be jointly sufficient for $\theta = (\theta_1, \ldots, \theta_n)$. If T is any unbiased estimator of $k(\theta)$, and if $T^* = E(T|U)$, then:*

- **(a)** *T^* is an unbiased estimator of $k(\theta)$.*
- **(b)** *T^* is a function of U, and does not depend on θ.*
- **(c)** *$Var(T^*) \leq Var(T)$ for every θ, and $Var(T^*) < Var(T)$ for some θ unless $T^* = T$ with probability 1.*

Proof.

(a) By the property of conditional expectation and by the fact that T is an unbiased estimator of $k(\theta)$,

$$E(T^*) = E(E(T|U)) = E(T) = k(\theta).$$

Hence, T^* is an unbiased estimator of $k(\theta)$.

(b) Because U is sufficient for θ, the conditional distribution of any statistic (hence, for T), given U, does not depend on θ. Thus, $T^* = E(T|U)$ is a function of U.

(c) From the property of conditional probability, we have the following:

$$\begin{aligned} Var(T) &= E(Var(T|U)) + Var(E(T|U)) \\ &= E(Var(T|U)) + Var(T^*). \end{aligned}$$

□

Because $Var(T|U) \geq 0$ for all u, it follows that $E(Var(T|U)) \geq 0$. Hence, $Var(T^*) \leq Var(T)$. We note that $Var(T^*) = Var(T)$ if and only if $Var(T|U) = 0$ or T is a function of U, in which case $T^* = T$ (from the definition of $T^* = E(T|U) = T$).

In particular, if $k(\theta) = \theta$, and T is an unbiased estimator of θ, then $T^* = E(T|U)$ will typically give the MVUE of θ. If T is the sufficient statistic that best summarizes the data from a given distribution with parameter θ, and we can find some function g of T such that $E(g(T)) = \theta$, it follows from the Rao–Blackwell theorem that $g(T)$ is the UMVUE for θ.

EXERCISES 5.4

5.4.1. Let $X_1, \ldots, X_n$ be a random sample from a population with density

$$f(x) = \begin{cases} e^{-(x-\theta)}, & \text{for } x > \theta \\ 0, & \text{otherwise.} \end{cases}$$

(a) Show that $\overline{X}$ is a biased estimator of θ.

(b) Show that $\overline{X}$ is an unbiased estimator of $\mu = 1 + \theta$.

5.4.2. The mean and variance of a finite population $\{a_1, \ldots, a_N\}$ are defined by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} a_i \text{ and } \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(a_i - \mu)^2.$$

For a finite population, show that the sample variance S^2 is a biased estimator of σ^2.

5.4.3. For an infinite population with finite variance σ^2, show that the sample standard deviation S is a biased estimator for σ. Find an unbiased estimator of σ. [We have seen that S^2 is an unbiased estimator of σ^2. From this exercise, we see that a function of an unbiased estimator need not be an unbiased estimator.]

5.4.4. Let $X_1, \ldots, X_n$ be a random sample from an infinite population with finite variance σ^2. Define

$$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Show that S'^2 is a biased estimator for σ^2, and that the bias of S'^2 is $-\frac{\sigma^2}{n}$. Thus, S'^2 is negatively biased, and so on average underestimates the variance. Note that S'^2 is the MLE of σ^2.

5.4.5. Let $X_1, \ldots, X_n$ be a random sample from a population with the mean μ. What condition must be imposed on the constants $c_1, c_2, \ldots, c_n$ so that

$$c_1X_1 + c_2X_2 + \cdots + c_nX_n$$

is an unbiased estimator of μ?

5.4.6. Let $X_1, \ldots, X_n$ be a random sample from a geometric distribution with parameter θ. Find an unbiased estimate of θ.

5.4.7. Let $X_1, \ldots, X_n$ be a random sample from $U(0, \theta)$ distribution. Let $Y_n = \max\{X_1, \ldots, X_n\}$. We know (from Example 5.3.4) that $\hat{\theta}_1 = Y_n$ is a maximum likelihood estimator of θ.

(a) Show that $\hat{\theta}_2 = 2\overline{X}$ is a method of moments estimator.

(b) Show that $\hat{\theta}_1$ is a biased estimator, and $\hat{\theta}_2$ is an unbiased estimator of θ.

(c) Show that $\hat{\theta}_3 = \frac{n+1}{n}\hat{\theta}_1$ is an unbiased estimator of θ.

5.4.8. Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance 1. Show that $\hat{\mu}^2 = \overline{X}^2$ is a biased estimator of μ^2, and compute the bias.

5.4.9. Let $X_1, \ldots, X_n$ be a random sample from an $N(\mu, \sigma^2)$ distribution. Show that the estimator $\hat{\mu} = \overline{X}$ is the MVUE for μ.

5.4.10. Let $X_1, \ldots, X_{n_1}$ be a random sample from an $N(\mu_1, \sigma^2)$ distribution and let $Y_1, \ldots, Y_{n_2}$ be a random sample from a $N(\mu_2, \sigma^2)$ distribution. Show that the pooled estimator

$$\hat{\sigma}^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

is unbiased for σ^2, where S_1^2 and S_2^2 are the respective sample variances.

5.4.11. Let $X_1, \ldots, X_n$ be a random sample from an $N(\mu, \sigma^2)$ distribution. Show that the sample median, M, is an unbiased estimator of the population mean μ. Compare the variances of $\overline{X}$ and M. [Note: For the normal distribution, the mean, median, and mode all occur at the same location. Even though both $\overline{X}$ and M are unbiased, the reason we usually use the mean instead of the median as the estimator of μ is that $\overline{X}$ has a smaller variance than M.]

5.4.12. Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ. Show that the sample mean $\overline{X}$ is sufficient for λ.

5.4.13. Let $X_1, \ldots, X_n$ be a random sample from a population with density function

$$f_\sigma(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x|}{\sigma}\right), \quad -\infty < X < \infty, \quad \sigma > 0.$$

Find a sufficient statistic for the parameter σ.

5.4.14. Show that if $\hat{\theta}$ is a sufficient statistic for the parameter θ and if the maximum likelihood estimator of θ is unique, then the maximum likelihood estimator is a function of this sufficient statistic $\hat{\theta}$.

5.4.15. Let $X_1, \ldots, X_n$ be a random sample from an exponential population with parameter θ.

(a) Show that $\sum_{i=1}^n X_i$ is sufficient for θ. Also show that $\overline{X}$ is sufficient for θ.

(b) The following is a random sample from exponential distribution.

1.5	3.0	2.6	6.8	0.7	2.2	1.3	1.6	1.1	6.5
0.3	2.0	1.8	1.0	0.7	0.7	1.6	3.0	2.0	2.5
5.7	0.1	0.2	0.5	0.4					

(i) What is an unbiased estimate of the mean?

(ii) Using part (a) and these data, find two sufficient statistics for the parameter θ.

5.4.16. Let $X_1, \ldots, X_n$ be a random sample from a one-parameter Weibull distribution with pdf

$$f(x) = \begin{cases} 2\alpha x e^{-\alpha x^2}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find a sufficient statistic for α.

(b) Using part (a), find an UMVUE for α.

5.4.17. Let $X_1, \ldots, X_n$ be a random sample from a population with density function

$$f(x) = \begin{cases} \dfrac{1}{\theta}, & -\dfrac{\theta}{2} \le x \le \dfrac{\theta}{2}, \ \theta > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Show that $\left(\min_{1 \le i \le n} X_i, \max_{1 \le i \le n} X_i \right)$ is sufficient for θ.

5.4.18. Let $X_1, \ldots, X_n$ be a random sample from a $G(1, \beta)$ distribution.

(a) Show that $U = \sum_{i=1}^{n} X_i$ is a sufficient statistic for β.

(b) The following is a random sample from a $G(1, \beta)$ distribution.

0.3	3.4	0.4	1.8	0.7	1.0	0.1	2.3	3.7	2.0
0.3	3.7	0.1	1.3	1.2	3.3	0.2	1.3	0.6	0.4

Find a sufficient statistic for β.

5.4.19. Show that X_1 is not sufficient for μ, if $X_1, \ldots, X_n$ is a sample from $N(\mu, 1)$.

5.4.20. Let $X_1, \ldots, X_n$ be a random sample from the truncated exponential distribution with pdf

$$f(x) = \begin{cases} e^{\theta - x}, & x > \theta \\ 0, & \text{otherwise.} \end{cases}$$

Show that $X_{(1)} = \min(X_i)$ is sufficient for θ.

5.4.21. Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf

$$f(x) = \begin{cases} \theta x^{\theta - 1}, & 0 < x < 1, \quad \theta > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Show that $U = X_1, \ldots, X_n$ is a sufficient statistic for θ.

5.4.22. Let $X_1, \ldots, X_n$ be a random sample of size n from a Bernoulli population with parameter p. Show that $\hat{p} = \overline{X}$ is the UMVUE for p.

5.4.23. Let $X_1, \ldots, X_n$ be a random sample from a Rayleigh distribution with pdf

$$f(x) = \begin{cases} \frac{2x}{\alpha} e^{-x^2/\alpha}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Show that $\sum_{i=1}^{n} X_i^2$ is sufficient for the parameter α.

5.5 OTHER DESIRABLE PROPERTIES OF A POINT ESTIMATOR

In this section, we discuss a few more properties of point estimators that can be used in choosing a particular estimator.

5.5.1 Consistency

It is a desirable property that the values of an estimator be closer to the value of the true parameter being estimated as the sample size becomes larger. To this end, we now introduce the notion of consistent estimators. Consistency is a large-sample, or asymptotic, property. That is, it describes the behavior of estimators as the sample size n becomes infinitely large. In this section, we use the notation $\hat{\theta}_n$ for $\hat{\theta}$ to show the dependence of the estimator on the sample size n.

Definition 5.5.1 *The estimator $\hat{\theta}_n$ is said to be a* **consistent estimator** *of θ if, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\left[\left|\hat{\theta}_n - \theta\right| \le \varepsilon\right] = 1$$

or equivalently,

$$\lim_{n\to\infty} P\left[\left|\hat{\theta}_n - \theta\right| > \varepsilon\right] = 0.$$

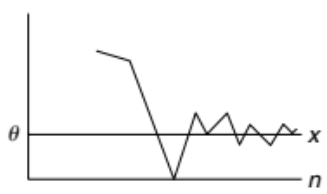
The statement "$\hat{\theta}_n$ is a consistent estimator of θ" is equivalent to "$\hat{\theta}_n$ converges in probability to θ." That is, the sample estimator should have a high probability of being close to the population value θ for large sample size n. The idea of consistency can be observed in Figure 5.2, where $\hat{\theta}_n$ converges to θ. If it did not, $\hat{\theta}_n$ would not be a consistent estimator of θ.

If the estimator is unbiased, we have the following result, which gives a sufficient condition for the consistency of an estimator. However, it is important to note that a consistent estimator need not be unbiased, and hence this result is not a necessary condition.

A SUFFICIENT CONDITION FOR CONSISTENCY OF AN UNBIASED ESTIMATOR

Theorem 5.5.1 *An unbiased estimator $\hat{\theta}_n$ of θ is a consistent estimator for θ if*

$$\lim_{n\to\infty} Var\left(\hat{\theta}_n\right) = 0.$$



■ **FIGURE 5.2** Consistency of an estimator.

The proof of this theorem follows directly from Chebyshev's inequality. A general version of this result is proved in Theorem 5.5.3.

Example 5.5.1

Let $X_1, \ldots, X_n$ be a random sample with true mean μ and finite variance. Then, the sample mean $\overline{X}$ is a consistent estimator of the population mean μ.

Solution

We show this result in two ways.

(i) *Using Chebyshev's inequality,* $P\{|X - \mu| \geq \varepsilon\} \leq \frac{Var(x)}{\varepsilon^2}$, *we obtain*

$$P\left[\left|\overline{X} - \mu\right| \leq k\right] \geq 1 - \frac{\sigma_{\overline{X}}^2}{k^2}$$

$$= 1 - \frac{\sigma^2}{k^2 n} \to 1 \text{as } n \to \infty.$$

Hence, $\overline{X}$ *is a consistent estimator of* μ.

(ii) *First note that* $\overline{X}$ *is an unbiased estimator of* μ. *Because* $Var(\overline{X}) = (\sigma^2/n)$, *we have*

$$\lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$$

Thus, from the previous theorem, $\overline{X}$ *is a consistent estimator of* μ.

We can generalize Theorem 5.5.1 even when the estimator is biased. The following result states that the mean square error of $\hat{\theta}_n$ decreases to zero as more and more observations are incorporated into its computation.

TEST FOR CONSISTENCY

Theorem 5.5.2 *Let* $\hat{\theta}_n$ *be an estimator of* θ *and let* $Var(\hat{\theta}_n)$ *be finite. If*

$$\lim_{n\to\infty} E\left[(\hat{\theta}_n - \theta)^2\right] = 0$$

then $\hat{\theta}_n$ *is a consistent estimator of* θ.

Proof. Using Chebyshev's inequality, we obtain

$$P\left[|\hat{\theta}_n - \theta| \geq \varepsilon\right] \leq \frac{E\left[(\hat{\theta}_n - \theta)^2\right]}{s^2}.$$

Because

$$\lim_{n\to\infty} E\left[(\hat{\theta}_n - \theta)^2\right] = 0, \text{ [by hypothesis]}$$

the right-hand side converges to zero. Thus,

$$\lim_{n\to\infty} P\left[\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right] = 0.$$

Consequently $\hat{\theta}_n$ is a consistent estimator of θ. □

Furthermore, we know that

$$E\left[\left(\hat{\theta}_n - \theta\right)^2\right] = Var\left(\hat{\theta}_n\right) + \left[B\left(\hat{\theta}_n\right)\right]^2,$$

and for unbiased estimators, the bias $B\left(\hat{\theta}_n\right)$ is zero. As a result, Theorem 5.5.1 is a particular case of Theorem 5.5.3. We now summarize the procedure for testing for consistency of an estimator as follows:

PROCEDURE TO TEST FOR CONSISTENCY

1. Check whether the estimator $\hat{\theta}_n$ is unbiased or not.
2. Calculate $Var\left(\hat{\theta}_n\right)$ and $B\left(\hat{\theta}_n\right)$, the bias of $\hat{\theta}_n$.
3. An unbiased estimator is consistent if $Var\left(\hat{\theta}_n\right) \to 0$ as $n \to \infty$.
4. A biased estimator is consistent if both

$$Var\left(\hat{\theta}_n\right) \to 0 \text{ and } B\left(\hat{\theta}_n\right) \to 0 \text{ as } n \to \infty.$$

Example 5.5.2

Let $X_1, \ldots, X_n$ be a random sample from $N\left(\mu, \sigma^2\right)$ population.

(a) Show that the sample variance S^2 is a consistent estimator for σ^2.

(b) Show that the maximum likelihood estimators for μ and σ^2 are consistent estimators for μ and σ^2.

Solution

(a) *We have already seen that* $ES^2 = \sigma^2$, *and hence,* S^2 *is an unbiased estimator of* σ^2. *Because the sample is drawn from a normal distribution, we know that* $\left[(n-1)S^2/\sigma^2\right]$ *has a chi-square distribution with* $(n-1)$ *d.f. and*

$$Var\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1).$$

Thus,

$$2(n-1) = Var\left(\frac{(n-1)S^2}{\sigma^2}\right) = \frac{(n-1)^2}{\sigma^4} Var\left(S^2\right).$$

This implies that

$$Var(S^2) = \frac{2\sigma^4}{n-1} \to 0 \text{ as } n \to \infty.$$

Hence, S^2 is a consistent estimator of the variance of a normal population.

(b) *We have seen that the MLE of μ is $\hat{\mu} = \overline{X}$, and that of σ^2 is $\hat{\sigma}_n^2 = (1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2$. Now $\hat{\mu}$ is an unbiased estimator of μ, and $Var(\overline{X}) = (\sigma^2/n) \to 0$ as $n \to \infty$. Therefore, from Theorem 5.5.1, $\overline{X}$ is a consistent estimator for μ.*

Now we will use the identity

$$E\left[(\hat{\theta}_n - \theta)^2\right] = Var(\hat{\theta}_n) + \left[B(\hat{\theta}_n)\right]^2$$

to show that the MLE for σ^2 is biased with

$$E(\hat{\sigma}_n^2) = \frac{n-1}{n}\sigma^2$$

and

$$B(\hat{\sigma}_n^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2.$$

Thus, $\hat{\sigma}_n^2 = (1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2 = ((n-1)/n)S^2$. Using part (a), we get

$$Var(\hat{\sigma}_n^2) = \frac{(n-1)^2}{n^2}Var(S^2)$$

$$= \frac{(n-1)^2\,2\sigma^4}{n^2(n-1)} = \frac{2(n-1)(\sigma^2)^2}{n^2}.$$

Therefore,

$$\lim_{n\to\infty} B(\hat{\sigma}_n^2) = \lim_{n\to\infty}\frac{-\sigma^2}{n} = 0, \text{ and } \lim_{n\to\infty} Var(\hat{\sigma}_n^2)$$

$$= \lim_{n\to\infty}\frac{2(n-1)(\sigma^2)^2}{n^2} = 0.$$

By Theorem 5.5.3,

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is a consistent estimator of σ^2.

■

From the foregoing example we can see that consistent estimators need not be unique. It turns out that most of the MLEs and method of moments estimators derived for important probability distributions are consistent.

5.5.2 Efficiency

We have seen that there can be more than one unbiased estimator for a parameter θ. We have also mentioned that the one with the least variance is desirable. Here, we introduce the concept of efficiency, which is based on comparing variances of the different unbiased estimators. If there are two unbiased estimators, it is desirable to have the one with a smaller variance.

Definition 5.5.2 *If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators for θ, the* **efficiency** *of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is the ratio*

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}.$$

If $Var(\hat{\theta}_2) > Var(\hat{\theta}_1)$, or equivalently, $e(\hat{\theta}_1, \hat{\theta}_2) > 1$, then, $\hat{\theta}_1$ is **relatively more efficient** *than $\hat{\theta}_2$. That is $\hat{\theta}_1$ has a smaller variance as compared to the variance of $\hat{\theta}_2$.*

We summarize the following procedure to compare the efficiencies of the different unbiased estimators.

PROCEDURE TO TEST RELATIVE EFFICIENCY

1. Check for unbiasedness of $\hat{\theta}_1$ and $\hat{\theta}_2$.
2. Calculate the variances of $\hat{\theta}_1$ and $\hat{\theta}_2$.
3. Calculate the relative efficiency as
$$e(\hat{\theta}_1; \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}.$$
4. Conclusion: If $e(\hat{\theta}_1, \hat{\theta}_2) < 1$, $\hat{\theta}_2$ is more efficient than $\hat{\theta}_1$, and if $e(\hat{\theta}_1, \hat{\theta}_2) > 1$, then, $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$. Among the unbiased estimators, the more efficient estimator is preferable.

Example 5.5.3

Let $X_1, \ldots, X_n, n > 3$, be a random sample from a population with a true mean μ and variance σ^2. Consider the following three estimators of μ:

$$\hat{\theta}_1 = \frac{1}{3}(X_1 + X_2 + X_3),$$

$$\hat{\theta}_2 = \frac{1}{8}X_1 + \frac{3}{4(n-2)}(X_2 + \cdots + X_{n-1}) + \frac{1}{8}X_n,$$

and

$$\hat{\theta}_3 = \overline{X}.$$

(a) Show that each of the three estimators is unbiased.

(b) Find $e(\hat{\theta}_2, \hat{\theta}_1)$, $e(\hat{\theta}_3, \hat{\theta}_1)$, and $e(\hat{\theta}_3, \hat{\theta}_2)$. Which of the three estimators is more efficient?

Solution

(a) *Given* $E(X_i) = \mu, i = 1, 2, \ldots, n$. *Then,*

$$E(\hat{\theta}_1) = \frac{1}{3}[E(X_1) + E(X_2) + E(X_3)] = \frac{3\mu}{3} = \mu$$

$$E(\hat{\theta}_2) = \frac{1}{8}E(X_1) + \frac{3}{4(n-2)}\left(E(X_2) + \cdots + E\left(X_{n-1}\right)\right) + \frac{1}{8}E(X_n)$$

$$= \frac{1}{8}\mu + \frac{3}{4(n-2)}(n-2)\mu + \frac{1}{8}\mu = \mu$$

$$E(\hat{\theta}_3) = E(\overline{X}) = \mu.$$

Hence, $\hat{\theta}_1, \hat{\theta}_2$, *and* $\hat{\theta}_3$ *are unbiased estimators of* μ.

(b) *Computing the variances, we have*

$$Var(\hat{\theta}_1) = \frac{1}{9}(Var(X_1) + Var(X_2) + Var(X_3))$$

$$= \frac{1}{9}3\sigma^2 = \frac{\sigma^2}{3}.$$

$$Var(\hat{\theta}_2) = \frac{\sigma^2}{64} + \frac{9(n-2)\sigma^2}{16(n-2)^2} + \frac{\sigma^2}{64}$$

$$= \frac{2\sigma^2}{64} + \frac{9\sigma^2}{16(n-2)} = \frac{n+16}{32(n-2)}\sigma^2.$$

$$Var(\hat{\theta}_3) = \frac{\sigma^2}{n}.$$

The relative efficiencies are

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} = \frac{\sigma^2(n+16)/32(n-2)}{\sigma^2/3}$$

$$= \frac{3(n+16)}{32(n-2)} < 1 \text{ for } n > 3.$$

Thus, for $n \geq 4$, $\hat{\theta}_2$ *is more efficient than* $\hat{\theta}_1$.

$$e(\hat{\theta}_3, \hat{\theta}_1) = \frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_3)} = \frac{\sigma^2/3}{\sigma^2/n} = \frac{n}{3} > 1 \text{ for } n \geq 4.$$

Hence, for $n > 3$, $\hat{\theta}_3$ is more efficient than $\hat{\theta}_1$.

$$e(\hat{\theta}_3, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_3)} = \frac{\frac{n+16}{32(n-2)}\sigma^2}{\sigma^2/n}$$

$$= \frac{n^2 + 16n}{32\,(n-2)} > 1 \text{ for } n \geq 4.$$

Therefore, even though both $\hat{\theta}_3\hat{\theta}_2$ are based on all the n observations, for $n > 3$, the sample mean $\hat{\theta}_3$ is more efficient than $\hat{\theta}_2$.

■

It is reasonable to compare estimators on the basis of variance alone if they are both unbiased. To facilitate the cases where the estimators are biased, we use the mean square error (MSE) in the definition of relative efficiency.

Definition 5.5.3 *An estimator $\hat{\theta}_1$ is* **more efficient** *than $\hat{\theta}_2$ if*

$$MSE\hat{\theta}_1 \leq MSE\hat{\theta}_2$$

with strict inequality for some θ. Also, the **relative efficiency** *of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is*

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{E\left[(\hat{\theta}_2 - \theta)^2\right]}{E\left[(\hat{\theta}_1 - \theta)^2\right]} = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)}.$$

Example 5.5.4

Let $X_1, \ldots, X_n, n \geq 2$ be a random sample from a normal population with a true mean μ and variance σ^2. Consider the following two estimators of σ^2: $\hat{\theta}_1 = S^2$, and $\hat{\theta}_2 = S'^2$. Find $e(\hat{\theta}_1, \hat{\theta}_2)$.

Solution

Because $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2\,(n-1)$, $E(S^2) = \sigma^2$, and $MSE(S^2) = Var(S^2)$. Also, $2(n-1) = Var\left(\frac{(n-1)S^2}{\sigma^2}\right) = \frac{(n-1)^2}{\sigma^4} Var(S^2)$.

Thus,

$$MSE(\hat{\theta}_1) = \frac{2}{n-1}\sigma^4.$$

Also, it can be shown that

$$MSE(S'^2) = \frac{(2n-1)}{n^2}\sigma^2.$$

Thus, the relative efficiency of $\hat{\theta}_1$ *with respect to* $\hat{\theta}_2$ *is*

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)} = \frac{MSE(S'^2)}{MSE(S^2)}$$

$$= \frac{\frac{(2n-1)}{n^2}\sigma^2}{\frac{2}{(n-1)}\sigma^2} = \frac{(2n-1)(n-1)}{2n^2}.$$

For $n \geq 2$, *it can be seen that* $e(\hat{\theta}_1, \hat{\theta}_2) < 1$. *Hence,* S'^2 *is relatively more efficient than* S^2.

We have seen that it is possible that one unbiased estimator is more efficient than another. This leads to the possibility of having one unbiased estimator more efficient than all the other unbiased estimators. This directs us to the following definition.

Definition 5.5.4 *An unbiased estimator* $\hat{\theta}_0$, *is said to be a* **uniformly minimum variance unbiased estimator (UMVUE)** *for the parameter* θ *if, for any other unbiased estimator* $\hat{\theta}$

$$Var(\hat{\theta}_0) \leq Var(\hat{\theta}),$$

for all possible values of θ.

It is not always easy to find an UMVUE for a parameter. However, the following result gives a lower bound for the variance of any unbiased estimator.

CRAMÉR–RAO INEQUALITY

Theorem 5.5.3 *Let* $X_1, \ldots, X_n$ *be a random sample from a population with pdf (or pf)* $f_\theta(x)$ *that depends on a parameter* θ. *If* $\hat{\theta}$ *is an unbiased estimator of* θ, *then, under very general conditions, the following inequality is true:*

$$Var(\hat{\theta}) \geq \frac{1}{nE\left[\left(\frac{\partial \ln f_\theta(x)}{\partial \theta}\right)^2\right]}.$$

If $\hat{\theta}$ *is an unbiased estimator of* $\psi(\theta)$, *then*

$$Var(\hat{\theta}) \geq \frac{\left(\frac{\partial \psi(\theta)}{\partial \theta}\right)^2}{nE\left[\frac{\partial}{\partial \theta} \ln f_\theta(x)\right]^2}.$$

If $L(\theta)$ is the likelihood function, we can rewrite the Cramér–Rao inequality in the form

$$Var(\hat{\theta}) \geq \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]}$$

From the Cramér–Rao inequality, we can obtain the following result.

EFFICIENT ESTIMATOR

Theorem 5.5.4 *If $\hat{\theta}$ is an unbiased estimator of θ and if*

$$Var(\hat{\theta}) = \frac{1}{nE\left[\left(\frac{\partial \ln f_\theta(x)}{\partial \theta}\right)^2\right]},$$

then $\hat{\theta}$ is a uniformly minimum variance unbiased estimator (UMVUE) of θ. Sometimes $\hat{\theta}$ is also referred to as an efficient estimator.

Note that if the function $f(.)$ is sufficiently smooth, it can be shown that

$$E\left(\frac{\partial \ln f_\theta(x)}{\partial \theta}\right)^2 = -E\left(\frac{\partial^2 \ln f_\theta(x)}{\partial \theta^2}\right) = Var[\ln f_\theta(x)]\,.$$

Hence, the Cramér–Rao inequality in this case can be rewritten as

$$Var(\hat{\theta}) \geq \frac{1}{-nE\left(\frac{\partial^2 \ln f_\theta(x)}{\partial \theta^2}\right)} = \frac{1}{nVar\left[\frac{\partial}{\partial \theta} \ln f_\theta(x)\right]}.$$

Now, we will give a procedure to apply the Cramér–Rao inequality.

CRAMÉR–RAO PROCEDURE TO TEST FOR EFFICIENCY

1. For the pdf (or pf), find $\frac{\partial \ln f(x)}{\partial \theta}$ and $\frac{\partial^2 \ln f(x)}{\partial \theta^2}$.
2. Calculate $(1/n)\,E\left[-\frac{\partial^2 \ln f(x)}{\partial \theta^2}\right]$ if $f(x)$ is smooth, or else calculate $\left[1/nE\left[\left(\frac{\partial \ln f(x)}{\partial \theta}\right)^2\right]\right]$.
3. Calculate $Var(\hat{\theta})$.
4. If the result of step 2 is equal to the result of step 3, then, $\hat{\theta}$ is efficient for θ.

Example 5.5.5

Let $X_1, \ldots, X_n$ be a random sample from an $N(\mu, \sigma^2)$ population with density function $f(x)$. Show that $\overline{X}$ is an efficient estimator for μ.

Solution

To calculate the Cramér–Rao lower bound, we have

$$\ln f(x) = c - \frac{(x-\mu)^2}{2\sigma^2},$$

where c is a constant not involving μ. Then

$$\frac{\partial \ln f(x)}{\partial \mu} = \frac{x-\mu}{\sigma^2}$$

and

$$\frac{\partial^2 \ln f(x)}{\partial \mu^2} = -\frac{1}{\sigma^2}$$

or

$$\frac{1}{nE\left[-\frac{\partial^2 \ln f(x)}{\partial \theta^2}\right]} = \frac{1}{nE\left(\frac{1}{\sigma^2}\right)} = \frac{\sigma^2}{n} = Var(\overline{X}).$$

Therefore, $\overline{X}$ is an efficient estimator of μ. That is, $\overline{X}$ is an UMVUE of μ.

Example 5.5.6

Suppose $p(x)$ is the Poisson distribution with parameter λ. Show that the sample mean $\overline{X}_n$ is an efficient estimator for λ.

Solution

Here the density function is given by $p(x) = \lambda^x \frac{e^{-\lambda}}{x!}$. Taking logarithms,

$$\ln p(x) = x \ln \lambda - \lambda - \ln(x!)$$

$$\frac{\partial \ln p(x)}{\partial \lambda} = \frac{x}{\lambda} - 1,$$

and

$$\frac{\partial^2 \ln p(x)}{\partial \lambda^2} = -\frac{x}{\lambda^2}$$

Therefore, using the fact that the expected value of a Poisson r.v. is λ,

$$\frac{1}{nE\left[-\frac{\partial^2 \ln f(x)}{\partial \lambda^2}\right]} = \frac{1}{nE\left(\frac{X}{\lambda^2}\right)} = \frac{\lambda}{n} = Var(\overline{X}).$$

Hence, $\overline{X}$ is an efficient estimator of λ.

Example 5.5.7

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli trial with probability of success p. Show that the maximum likelihood estimator is also an efficient estimator.

Solution

Note that the MLE of p is $\hat{p} = (1/n)\sum_{i=1}^{n} X_i = X/n$, the fraction of successes in the total number of trials, n. Because we can view n Bernoulli trials as being a single observation from a binomial distribution with parameters n and p, the likelihood function is

$$L(p) = \binom{n}{x} p^x (1-p)^x.$$

Then,

$$\ln L(p) = \ln \binom{n}{x} + x \ln p + (n-x)\ln(1-p).$$

Now

$$\frac{\partial \ln L(p)}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p} = \frac{x-np}{p(1-p)}.$$

Hence,

$$\begin{aligned} E\left[\left(\frac{\partial \ln L(p)}{\partial p}\right)^2\right] &= E\left[\left(\frac{x-np}{p(1-p)}\right)^2\right] \\ &= \frac{Var(x)}{[p(1-p)]^2} \\ &= \frac{np(1-p)}{[p(1-p)]^2}\frac{n}{p(1-p)}. \end{aligned}$$

Therefore, the Cramér– Rao bound is

$$\frac{1}{E\left[\left(\frac{\partial \ln L(p)}{\partial p}\right)^2\right]} = \frac{p(1-p)}{n}.$$

Now

$$\begin{aligned} Var(\hat{p}) &= Var\left(\frac{X}{n}\right) \\ &= \frac{1}{n^2} Var(x) \\ &= \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}. \end{aligned}$$

Because the variance of the estimator is equal to the Cramér–Rao lower bound, we conclude that $\hat{p} = \frac{X}{n}$ is an efficient estimator of p.

■

It is important to note that an UMVUE may not exist for a given problem. Even when an UMVUE exists, it is not necessary that it have a variance equal to the Cramér–Rao lower bound. The term

$I(\theta) = E\left[\left(\frac{\partial \ln f(x)}{\partial \theta}\right)^2\right]$ is called the *Fisher information*. In fact, for a random sample of size n with likelihood function $L(\theta)$, the Fisher information is defined as $I_n(\theta) = E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]$. It can be shown that the Fisher information in a sample of size n is n times the Fisher information in one observation. That is, $I_n(\theta) = nI(\theta)$.

5.5.3 **Minimal Sufficiency and Minimum-Variance Unbiased Estimation**

In the study of statistics, it is desirable to reduce the data contained in the sample as much as possible without losing relevant information. Our objective is to find minimal sufficient statistics and use them to develop uniformly minimum variance unbiased estimators (UMVUEs) for true parameters. Whenever sufficient statistics exist, then a statistician with those summary measures is as well off as the statistician with the entire sample, for point estimation purposes. Minimal sufficient statistics are those that are sufficient for the parameters and are functions of every other set of sufficient statistics for those same parameters.

Definition 5.5.5 *A sufficient statistic $T(X)$ is called a* **minimal sufficient statistic** *if for any other statistic $T'(X)$, $T(X)$ is a function of $T'(X)$. That is,*

$$T(X) = g\left(T'(X)\right).$$

Using this definition, it is difficult to determine whether a set of statistics is, in fact, minimal sufficient. Now we will present a method due to Lehmann and Scheffé that will be of great help in finding a minimal sufficient statistic.

We can summarize the Lehmann and Scheffé method to find a minimal sufficient statistic as follows. Let $X_1, \ldots, X_n$ be a random sample with pdf or pmf $f(x)$ that depends on a parameter θ. Let $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ be two different sets of values of $(X_1, \ldots, X_n)$. Let

$$\frac{L(\theta; x_1, \ldots, x_n)}{L(\theta; y_1, \ldots, y_n)}$$

be the ratio of the likelihoods evaluated at these two points. Suppose it is possible to find a function $g(x_1, \ldots, x_n)$ such that this ratio will be free of the unknown parameter θ if and only if $g(x_1, \ldots, x_n) = g(y_1, \ldots, y_n)$. If such a function g can be found, then $g(X_1, \ldots, X_n)$ is a minimal sufficient statistic for θ.

Example 5.5.8

Let $X_1, \ldots, X_n$ be a random sample from the Bernoulli distribution where $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$, with p unknown. Find a minimal sufficient statistic for p.

Solution

The ratio of the likelihoods is

$$\frac{L(x_1,\ldots,x_n)}{L(y_1,\ldots,y_n)} = \frac{p(x_1,\ldots,x_n)}{p(y_1,\ldots,y_n)} = \frac{p^{\sum x_i}(1-p)^{n-\sum x_i}}{p^{\sum y_i}(1-p)^{n-\sum y_i}}$$
$$= \left(\frac{p}{1-p}\right)^{\sum x_i - \sum y_i}.$$

This ratio is to be independent of p, if and only if

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0$$

which implies

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i.$$

Therefore,

$$g(X_1,\ldots,X_n) = \sum_{i=1}^{n} X_i$$

is a minimal sufficient statistic for p.

■

Example 5.5.9

Let $X_1,\ldots,X_n$ be a random sample from a $U(0,\theta)$ distribution. Find a minimal sufficient statistic for θ.

Solution

The likelihood function is

$$L = \begin{cases} \dfrac{1}{\theta^n}, & \text{if } \max(x_1,\ldots,x_n) \le \theta \\ 0, & \text{otherwise}. \end{cases}$$

Denote by $x_{\max} = \max(x_1,\ldots,x_n)$, and $y_{\max} = \max(y_1,\ldots,y_n)$. Then, the ratio of the likelihood functions is

$$\frac{L(x_1,\ldots,x_n)}{L(y_1,\ldots,y_n)} = \begin{cases} 1, & \text{if } \max(x_{\max}, y_{\max}) \le \theta, \\ 0, & \text{if } y_{\max} < x_{\max}, \text{ and } y_{\max} \le \theta \le x_{\max}, \\ \text{undefined}, & \text{elsewhere}. \end{cases}$$

Thus, the ratio will not depend on θ if and only if $x_{\max} = y_{\max}$. Therefore, a minimal sufficient statistic for θ is $X_{(n)}$, the largest order statistic.

■

It is important to note that although we often can find a single statistic that is minimal sufficient for one parameter, this need not be the case (see Exercise 5.5.1). For most of the density functions that we consider, any unbiased estimator that is a function of a minimal sufficient statistic will be a ***uniformly minimum variance unbiased estimator*** (UMVUE), that is, it will posses the smallest variance possible among unbiased estimators.

Example 5.5.10

Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with known mean $\mu = \mu_0$ and unknown variance σ^2. Show that $\sum_{i=1}^{n} (X_i - \mu_0)^2$ is the minimal sufficient statistic for σ^2. Use this statistic to find an MVUE of σ^2.

Solution

The ratio of the likelihoods is

$$\frac{L(x_1, \ldots, x_n)}{L(y_1, \ldots, y_n)} = \frac{\exp\left[-\sum (x_i - \mu_0)^2 / 2\sigma^2\right]}{\exp\left[-\sum (y_i - \mu_0)^2 / 2\sigma^2\right]}$$

$$= \exp\left[\frac{1}{2\sigma^2}\left\{\sum (y_i - \mu_0)^2 - \sum (x_i - \mu_0)^2\right\}\right].$$

In order for this ratio to be free of σ^2, we need

$$\sum (y_i - \mu_0)^2 = \sum (x_i - \mu_0)^2.$$

Hence, $\sum (X_i - \mu_0)^2$ is minimal sufficient for σ^2.
Because $E(X_i - \mu_0)^2 = \sigma^2$, we can see that $(1/n) \sum (X_i - \mu_0)^2$ is an unbiased estimator of σ^2. Because this is a function of a minimal sufficient statistic, $(1/n) \sum_{i=1}^{n} (X_i - \mu_0)^2$ is an MVUE of σ^2.

EXERCISES 5.5

5.5.1. Show that the maximum likelihood estimator for p, Y_n/n in a binomial distribution is consistent.

5.5.2. Show that Y_n, the nth-order statistic from a $U(0, \theta)$ distribution, is a consistent estimator for θ.

5.5.3. Let $X_1, \ldots, X_n$ be a random sample with $EX_i = \mu_i$, $EX_i^2 = \mu_2'$, and $EX_i^4 = \mu_4'$, all finite. Show that $S^2 = (1/n) \sum_{i=1}^{n} (X_i - \overline{X})^2$ is a consistent estimator of $\sigma^2 = Var(X_i)$.

5.5.4. Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f(x) = \begin{cases} \alpha x^{\alpha - 1}, & \text{for } 0 < x < 1; \ \alpha > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Is the method of moments estimator for α consistent?

5.5.5. Let $X_1, \ldots, X_n$ be a random sample from an exponential population with parameter θ. Show that $\overline{X}$ is a consistent estimator of θ.

5.5.6. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent random samples from populations with means μ_1 and μ_2 variances σ_1^2 and σ_2^2, respectively. Show that the difference $\overline{X} - \overline{Y}$ is a consistent estimator of $\mu_1 - \mu_2$.

5.5.7. Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f(x) = \begin{cases} \frac{1}{\alpha} x^{(1-\alpha)/\alpha}, & \text{for } 0 < x < 1; \ \alpha > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Show that the maximum likelihood estimator of α is $\hat{\alpha} = -(1/n) \sum_{i=1}^{n} \ln X_i$.
(b) Is $\hat{\alpha}$ of part (a) an unbiased estimator of α?
(c) Is $\hat{\alpha}$ of part (a) a consistent estimator of α?

5.5.8. Let $X_1, \ldots, X_n$ be a random sample from a Rayleigh distribution with pdf

$$f(x) = \begin{cases} \frac{x}{\alpha} e^{-x^2/(2\alpha)}, & \text{for } x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Determine the maximum likelihood estimator $\hat{\alpha}$ of α.
(b) Is $\hat{\alpha}$ of part (a) an unbiased estimator of α?
(c) Is $\hat{\alpha}$ of part (a) a consistent estimator of α?

5.5.9. Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution on the interval $(\theta, \theta + 1)$. Let

$$\theta_1 = \overline{X} - \frac{1}{2}, \theta_2 = X_{(n)} - \frac{n}{n+1},$$

where $X_{(n)}$ is the nth order statistic. Find the efficiency of $\hat{\theta}_2$ relative to $\hat{\theta}_1$.

5.5.10. Let $X_1, \ldots, X_n$ be a random sample from an $N(\mu, \sigma^2)$ population. Let $\hat{\theta}_1$ be the sample mean and $\hat{\theta}_2$ be the sample median. It is known that $Var(\hat{\theta}_2) = (1.2533)^2(\sigma^2/n)$. Find the efficiency of $\hat{\theta}_2$ relative to $\hat{\theta}_1$.

5.5.11. Let $X_1, \ldots, X_n$ be a random sample from an exponential population with parameter θ. Show that $\overline{X}$ is efficient for θ.

5.5.12. Let $X_1, \ldots, X_n$ be a random sample from an $N(\mu, \sigma^2)$ population. Show that

$$MSE\left(S'^2\right) = \frac{2(n-1)}{n^2}\sigma^4,$$

where $S'^2 = (1/n) \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$.

5.5.13. Prove

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

making suitable assumptions.

5.5.14. Let $X_1, \ldots, X_n$ be a random sample from an $N(\mu, \sigma^2)$ population.

(a) Show that the sample variance S^2 is an UMVUE for σ^2 when the value of μ is not known.

(b) Show that the variance of S^2 is greater than the Cramér–Rao lower bound.

5.5.15. Let $X_1, \ldots, X_n$ be a random sample from a $U(0, \theta)$ distribution. Let $X_{(n)}$ be the nth order statistic.

(a) Show that $\hat{\theta}_1 = X_{(n)}$, $\hat{\theta}_2 = 2\overline{X}$, and $\hat{\theta}_3 = \frac{n+1}{n} X_{(n)}$ are unbiased estimators of θ.

(b) Find the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$.

(c) Find the efficiency of $\hat{\theta}_2$ relative to $\hat{\theta}_3$.

5.5.16. Let $X_1, \ldots, X_n$, $(n \geq 2)$ be a random sample from a distribution with pdf

$$f(x) = \frac{1}{\pi\left[1+(x-\theta)^2\right]}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

Show that the Cramér–Rao lower bound for a UBE of θ is $2/n$.

5.5.17. Let $X_1, \ldots, X_n$, $n > 4$, be a random sample from a population with a mean μ and variance σ^2. Consider the following three estimators of μ:

$$\hat{\theta}_1 = \frac{1}{9}(X_1 + 2X_2 + 5X_3 + X_4),$$

$$\hat{\theta}_2 = \frac{2}{5}X_1 + \frac{1}{5}X_2 + \frac{1}{5(n-3)}(X_3 + \ldots + X_{n-1}) + \frac{1}{5}X_n,$$

and $\hat{\theta}_3 = \overline{X}$.

(a) Show that each of the three estimators is unbiased.

(b) Find $e(\hat{\theta}_2, \hat{\theta}_1)$, $e(\hat{\theta}_3, \hat{\theta}_1)$, and $e(\hat{\theta}_3, \hat{\theta}_2)$.

5.5.18. Find the Cramér–Rao lower bound for the variance of an unbiased estimator of θ, based on a sample of size n for the following pdfs:

(i) $f(x, \theta) = \frac{1}{\theta^2} x e^{-x/\theta}, \quad x > 0, \quad \theta > 0.$

(ii) $f(x, \theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad \theta > 0.$

5.5.19. Let $Y_1, \ldots, Y_n$ be a random sample from the uniform distribution over the interval $(\theta - 1, \theta + 1)$. Show that the order statistics $X_1 = \min(Y_i)$ and $X_n = \max(Y_i)$ are jointly sufficient for θ. Also, show that X_1 and X_n are jointly minimal for θ.

5.5.20. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and known variance σ^2. Find the maximum likelihood estimator of μ and show that it is a function of a minimal sufficient statistic.

5.5.21. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. Show that $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$ are jointly minimal sufficient for μ and σ^2. Also show that $\overline{X}$ and S^2 are UMVUEs for μ and σ^2.

5.5.22. Let $X_1, \ldots, X_n$ be a random sample from the Weibull density

$$f(x) = \begin{cases} \left(\frac{2x}{\alpha}\right) e^{-x^2/\alpha}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find an UMVUE for α.

5.5.23. Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ. Find a minimal sufficient statistic for λ.

5.5.24. Let $X_1, \ldots, X_n$ be a random sample from a gamma distribution with parameters α and β, both unknown. Find minimal sufficient statistics for the parameters α and β.

5.5.25. Let $X_1, \ldots, X_n$ be a random sample from a distribution with density function

$$f(x) = \begin{cases} e^{x-\beta}, & x \geq \beta \\ 0, & \text{otherwise.} \end{cases}$$

Find an UMVUE for β.

5.5.26. Let $X_1, \ldots, X_n$ be a random sample from the exponential distribution with pdf

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Show that $\overline{X}$ is an UMVUE for β. Also show that $\left(\frac{n}{n+1}\right)\overline{X}^2$ is an MVUE for β^2.

5.5.27. Let $X_1, \ldots, X_n$ be a random sample from a Rayleigh distribution with pdf

$$f(x) = \begin{cases} \frac{2x}{\beta} e^{-x^2/\beta}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find an UMVUE for β.

5.6 CHAPTER SUMMARY

In this chapter we have discussed the basic concepts of point estimation. Two methods of finding point estimators were described—the method of moments and the method of maximum likelihood. We have seen that the maximum likelihood estimators possess the invariance property, which states that if $\hat{\theta}$ is a maximum likelihood estimator of the parameter θ, then $h(\hat{\theta})$ is a maximum likelihood estimator

for $h(\theta)$. Some desirable properties of the point estimators that we have discussed are unbiasedness, consistency, efficiency, and sufficiency. Unbiasedness means that the expected value of the sample statistic (the mean of its probability distribution) should be equal to the parameter. Unbiasedness guards against consistently producing under- or overestimates of the parameter in repeated sampling. If the estimator is consistent, then, as the sample size increases, the estimator can be expected to get closer and closer to the population parameter. Efficient estimators have the lowest variance among all other estimators. A sufficient estimator is a "good" estimator of the population parameter θ in the sense that it depends on fewer data values.

We will now list some of the key definitions introduced in this chapter.

- Method of moments
- Likelihood function
- Maximum likelihood equations
- Unbiased estimator
- Mean square error
- Minimum variance unbiased estimator
- Consistent estimator
- Efficiency
- Uniformly minimum variance unbiased estimator
- Efficient estimator
- Sufficient estimator
- Jointly sufficient
- Minimal sufficient statistic

In this chapter, we have also learned the following important concepts and procedures.

- The method of moments procedure
- Procedure to find MLE
- Procedure to test for consistency
- Procedure to test relative efficiency
- Cramér–Rao procedure to test for efficiency
- Procedure to verify sufficiency

5.7 COMPUTER EXAMPLES

Because in the earlier chapters we have already given steps to obtain summary statistics such as the mean and variance using SPSS and SAS, we could use those commands to obtain point estimates as we will do with Minitab. Therefore, we will not give separate subsections for SPSS and SAS procedures. The following examples illustrate Minitab procedures.

Example 5.7.1

Generate 50 sample points from an $N(4, 4)$ distribution and find the descriptive statistics. Obtain an unbiased and sufficient estimate of μ.

Solution

Because we know that the sample mean $\bar{x}$ is an unbiased and sufficient estimate of the population mean μ, we only need to find the sample mean of the generated data.

Calc > Random Data > Normal . . . > Type ***50*** in ***Generate __ rows of data > Store in column(s):*** type ***C1***
> type in ***Mean: 4.0*** and in ***Standard deviation: 2.0 >*** click ***OK***.

The following is one possible output.

C1						
4.76039	5.07819	4.85263	4.08032	6.77772	4.21677	1.51811
5.16925	3.68845	6.40513	6.13801	7.20015	2.41415	3.50008
3.25593	2.66181	1.01352	5.82506	6.04212	5.22235	5.29924
2.80955	4.19032	4.65449	3.48680	6.39083	6.56357	1.32281
2.43494	2.01465	4.02358	8.22997	2.44516	0.39563	3.78948
1.76723	3.15460	4.81882	0.36250	0.85002	14.47052	0.79586
2.86329	5.97599	7.75170	7.10011	6.61681	0.97982	4.01400
5.38503						

Now follow the procedure to obtain the descriptive statistics from Example 1.8.3 to obtain

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C1	50	4.116	4.135	4.115	2.047	0.289
Variable	Minimum	Maximum	Q1	Q3		
C1	0.362	8.230	2.443	5.863		

We can see that the unbiased and sufficient estimate of the mean μ for these data is $\bar{x} = 4.116$.

Example 5.7.2

Generate 35 samples from a $U(0, 5)$ distribution and using the descriptive statistics command, find the maximum likelihood estimate for this data.

Solution

We know that for a random sample $X_1, \ldots, X_n$ from $U(0, \theta)$, the MLE, $\hat{\theta} = \max(X_i) = X_{(n)}$, the nth order statistic. We can use the following steps to obtain the estimate.

Calc > Random Data > Uniform. . . > Type ***35*** in ***Generate __ rows of data > Store in column(s):*** type ***C1 >*** type in ***Lower end point: 0.0*** and in ***Upper end point: 5.0 >*** click ***OK***

One possible output is given below.

C1						
4.32848	4.79402	0.34515	0.08428	1.93000	0.27878	3.12992
0.07934	3.12453	1.69073	3.44003	0.47447	2.28072	0.49205
2.92537	2.39721	4.84440	1.79129	4.38718	3.60697	0.94159
4.20844	3.75506	4.56626	3.50280	1.95689	0.56969	1.02543
3.25272	4.61083	3.06527	2.34003	0.40877	2.52708	1.44525

Now follow the procedure to obtain the descriptive statistics from Example 1.8.3 to obtain

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C1	35	2.417	2.397	2.413	1.541	0.260

Variable	Minimum	Maximum	Q1	Q3
C1	0.079	4.844	0.942	3.607

Therefore the MLE $\hat{\theta} = 4.884$.

■

For the previous example, it should be noted that because we are generating random data, each time we follow this procedure, we will be getting different answers. When we have a particular data set, enter the data in **C1** and just use the procedure to find the descriptive statistics. For other distributions, click the appropriate distribution in **Random Data**.

PROJECTS FOR CHAPTER 5

5A. Asymptotic Properties

In general, we do not have a single sample with one estimator of the unknown parameter θ. Rather, we will have a general formula that defines an estimator for any sample size. This gives a sequence of estimators of θ:

$$\hat{\theta} = h_n(X_1, \ldots, X_n), \quad n = 1, 2, \ldots..$$

In this case, we can define the following asymptotic properties:

(i) The sequence of estimators $\hat{\theta}_n$ is said to be *asymptotically unbiased* for θ if $bias(\hat{\theta}_n) \to 0$ as $n \to \infty$.

(ii) Suppose $\hat{\theta}_n$ and $\hat{y}_n$ are two sequences of estimators that are asymptotically unbiased for θ. The *asymptotic relative efficiency* of $\hat{\theta}_n$ to $\hat{y}_n$ is defined by

$$\lim_n \frac{Var(\hat{\theta}_n)}{Var(\hat{y}_n)}.$$

(a) Show that $\hat{\theta}_n$ is asymptotically unbiased if and only if

$$E(\hat{\theta}_n) \to \theta \text{ as } n \to \infty.$$

(b) Let $X_1, \ldots, X_n$ be a random sample from a distribution with unknown mean μ and variance σ^2. It is known that the method of moments estimators for μ and σ^2 are, respectively, the sample mean $\overline{X}$ and $S'^2_n = (1/n)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = ((n-1)/n)\,S_n^2$, where S_n^2 is the sample variance.

- **(i)** Show that S'^2_n is an asymptotically unbiased estimator of σ^2.
- **(ii)** Show that the asymptotic relative efficiency of S'^2_n to S_n^2 is 1.
- **(iii)** Show that $MSE\left(S'^2_n\right) < MSE\left(S_n^2\right)$. Thus, S_n^2 is unbiased but S'^2_n has a smaller mean square error. However, it should be noted that the difference is very small and approaches zero as n becomes large.

5B. **Robust Estimation**

The estimators derived in this chapter are for particular parameters of a presumed underlying family of distributions. However, if the choice of the underlying family of distributions is based on past experience, there is a possibility that the true population will be slightly different from the model used to derive the estimators. Formally, a statistical procedure is *robust* if its behavior is relatively insensitive to deviations from the assumptions on which it is based. If the behavior of an estimator is taken as its variance, a given estimator may have minimum variance for the distribution used, but it may not be very good for the actual distribution. Hence, it is desirable for the derived estimators to have small variance over a range of distributions. We call such estimators *robust estimators*. The following illustrates how the variance of an estimator can be affected by deviations from the presumed underlying population model.

Consider estimating the mean of a standard normal distribution. Let $X_1, \ldots, X_n$ be a random sample from a standard normal distribution. Suppose the population actually follows a contaminated normal distribution. That is, for $0 \le \delta \le 1$, $100\,(1-\delta)\,\%$ of the observations come from an $N(0, 1)$ distribution and the remaining $100\delta\%$ of observations come from an $N(0, 5)$ distribution. We already know that the minimum variance unbiased estimator of the mean μ of an uncontaminated normal distribution is the sample mean. A less effective alternative would be the sample median.

(a) Conduct a simulation study with sample size n that takes, say, 5000 random samples of 100 observations each. Find the mean and median. Also find the sample variance of each. For various values of δ, say 0.0, 0.01, 0.05, 0.1, 0.2, 0.3, and 0.4, create a table of variances of sample mean and sample variance. Compare the variances as the value of δ increases.

(b) The aim of robust estimation is to derive estimators with variance near that of the sample mean when the distribution is standard normal while having the variance remain relatively stable as δ increases. One such estimator is the $\alpha - trimmedmean$. Let $0 \le \alpha \le 0.5$, and define $k = [n\alpha]$, where $[x]$ is the greatest integer that is less than or equal to x. For the ordered sample, discard the k highest and lowest observations and find the mean of the remaining $n-k$ observations. That is, let $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ be the ordered sample, and define

$$\overline{X_\alpha} = \frac{X_{(1+k)} \le X_{(2+k)} \le \ldots \le X_{(n+k)}}{n-2k}.$$

For the values of δ and the samples in part (a), compute the mean and the 0.05-, 0.1-, 0.25-, and 0.5-trimmed means. Discuss the robustness.

5C. **Numerical Unbiasedness and Consistency**

(a) Run the simulation of a normal experiment with increasing sample size. Numerically show the unbiased and consistent properties of the sample mean. Run the experiment at least up until $n = 1000$.

(b) Repeat the experiment of part (a), now with an exponential distribution.

5D. **Averaged Squared Errors (ASEs)**

Generate 25 samples of size 40 from a normal population with $\mu = 10$, and $\sigma^2 = 4$. For each of the 25 samples:

(a) Compute: $\overline{x}$, $s^2 = \frac{\sum_{i=1}^{40}(x_i - \overline{x})^2}{39}$, $s_1^2 = \frac{\sum_{i=1}^{40}(x_i - \overline{x})^2}{40}$, and $s_2^2 = \frac{\sum_{i=1}^{40}(x_i - \overline{x})^2}{41}$.

(b) Compute the average squared error (ASE) for each of the estimates s^2, s_1^2, s_2^2 as follows.

Let $K^{s^2} = \left[\left[\sum_{i=1}^{K}(x_i - \overline{x})^2\right]/39\right]$ for $K = 1, 2, \ldots, 25$; and K^{s^2} be the sample variance for the Kth sample. Then, the average squared error is

$$ASE = \frac{\sum_{i=1}^{25}\left(K^{s^2} - \sigma^2\right)^2}{25}.$$

Repeat this procedure for the other two estimators. Compare the three ASEs and check which has the least ASE.

(c) Repeat (a) and (b) with a sample size of 15.

5E. **Alternate Method of Estimating the Mean and Variance**

(a) Consider the following alternative method of estimating μ and σ^2. We sample sequentially, and at each stage we compute the estimates of μ and σ^2 as follows.

Let $X_1, \ldots, X_n, X_{n+1}$ be the sample values.

Compute

$$\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n}, \overline{X}_{n+1} = \frac{\sum_{i=1}^{n+1} X_i}{n+1}, S_n^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2}{n-1}, \text{ and}$$

$$S_{n+1}^2 = \frac{\sum_{i=1}^{n+1}\left(X_i - \overline{X_n}\right)^2}{n}.$$

The sequential procedure is stopped when

$$\left|S_n^2 - S_{n+1}^2\right| \leq 0.01.$$

This will also determine the sample size.

(b) Compare the sample sizes and estimates in 5D and 5E.

5F. Newton–Raphson in One Dimension

For a given function $g(x)$, suppose we need to solve $g(\theta) = 0$. Using the first-order Taylor expansion, $g(\theta) \approx g(x) + (\theta - x)g'(x)$, where $g'(x) = \frac{dg}{dx}$, and setting $g(\theta) = 0$, we get $\theta \approx x - \frac{g(x)}{g'(x)}$. Thus, starting with an initial guess solution x, the guess is updated by θ using the previous formula. This derivation is the basis for the Newton–Raphson iterative method for obtaining the solution of $g(\theta) = 0$. This is given by

$$\theta_{(n+1)} = \theta_n - \frac{g(\theta_n)}{g'(\theta_n)}, \quad n \geq 0,$$

where θ_n is the value of θ at the nth iteration, starting with the initial guess, θ_0. For a good approximation of the solution, the choice of θ_0 is important. The convergence of this algorithm cannot be guaranteed.

For the MLE, we want to find a solution of

$$g(\theta) = \frac{dL}{d\theta} = 0,$$

where $L = L(\theta)$ is the likelihood function of the random sample $X_1, \ldots, X_n$. An iterative algorithm for finding the MLE can be given by

$$\theta_{(n+1)} = \theta_n - \frac{\frac{dL}{d\theta}(\theta_n)}{\frac{d^2L}{d\theta^2}(\theta_n)}, \quad n \geq 0.$$

Write a computer program to find the MLE of α for a gamma distribution with parameters α and β.

5G. The Empirical Distribution Function

The estimators in this chapter yield a single real value (point estimate) for each parameter. In Chapter 6, we will learn about so-called interval estimates. In this project, we use an estimation procedure that estimates the whole distribution function, F, of a random variable X. We now define the empirical distribution.

The *empirical distribution function* for a random sample $X_1, \ldots, X_n$ from a distribution F is the function defined by

$$F_n(x) = \frac{1}{n}\#\{i,\ 1 \leq i \leq n : X_i \leq x\}.$$

It can be shown that $nF_n(x)$ is a binomial random variable with

$$E[F_n(x)] = F(x) \text{ and } Var[F_n(x)] = \frac{1}{n}F(x)[1 - F(x)].$$

Also, by the strong law of large numbers, for each real number x,

$$\lim_{n\to\infty} F_n(x) = F(x) \text{ with probability } 1.$$

One of the tests to determine whether a random sample comes from a specific distribution is the Kolmogorov–Smirnov (K-S) test. The K-S test is based on the maximum distance between the empirical distribution function and the actual cumulative distribution function of this specific distribution (such as, say, the normal distribution).

Using the method of Project 4A (or using any statistical software), generate 100 sample points from a normal distribution with mean 2 and variance 9. Graph the empirical distribution function for this sample. Compare this graph with the graph of the $N(2, 9)$ distribution.