CHAPTER 8

Parametric Point Estimation

8.1 INTRODUCTION

In this chapter we study the theory of point estimation. Suppose, for example, that a random variable X is known to have a normal distribution $\mathcal{N}(\mu, \sigma^2)$, but we do not know one of the parameters, say μ. Suppose further that a sample $X_1, X_2, \dots, X_n$ is taken on X. The problem of point estimation is to pick a (one-dimensional) statistic $T(X_1, X_2, \dots, X_n)$ that best estimates the parameter μ. The numerical value of T when the realization is $x_1, x_2, \dots, x_n$ is frequently called an *estimate* of μ, while the statistic T is called an *estimator* of μ. If both μ and σ^2 are unknown, we seek a joint statistic $T = (U, V)$ as an estimator of (μ, σ^2).

In Section 8.2 we formally describe the problem of parametric point estimation. Since the class of all estimators in most problems is too large, it is not possible to find the "best" estimator in this class. One narrows the search somewhat by requiring that the estimators have some specified desirable properties. We describe some of these and also outline some criteria for comparing estimators.

Section 8.3 deals, in detail, with some important properties of statistics, such as sufficiency, completeness, and ancillarity. We use these properties in later sections to facilitate our search for optimal estimators. Sufficiency, completeness, and ancillarity also have applications in other branches of statistical inference, such as testing of hypotheses and nonparametric theory.

In Section 8.4 we investigate the criterion of unbiased estimation and study methods for obtaining optimal estimators in the class of unbiased estimators. In Section 8.5 we derive two lower bounds for variance of an unbiased estimator. These bounds can sometimes help in obtaining the "best" unbiased estimator.

In Section 8.6 we describe one of the oldest methods of estimation, and in Section 8.7 we study the method of maximum likelihood estimation and its large-sample properties. Section 8.8 is devoted to Bayes and minimax estimation, and Section 8.9 deals with equivariant estimation.

8.2 PROBLEM OF POINT ESTIMATION

Let $\mathbf{X}$ be an RV defined on a probability space $(\Omega, \mathcal{S}, P)$. Suppose that the DF F of $\mathbf{X}$ depends on a certain number of parameters, and suppose further that the functional form of F is known except perhaps for a finite number of these parameters. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ be the unknown parameter associated with F.

Definition 1. The set of all admissible values of the parameters of a DF F is called the *parameter space*.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be an RV with DF $F_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ is a vector of unknown parameters, $\boldsymbol{\theta} \in \Theta$. Let ψ be a real-valued function on Θ. In this chapter we investigate the problem of approximating $\psi(\boldsymbol{\theta})$ on the basis of the observed value $\mathbf{x}$ of $\mathbf{X}$.

Definition 2. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n) \sim P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$. A statistic $\delta(\mathbf{X})$ is said to be a *(point) estimator* of ψ if $\delta : \mathfrak{X} \to \Theta$ where $\mathfrak{X}$ is the space of values of $\mathbf{X}$.

The problem of point estimation is to find an estimator δ for the unknown parametric function $\psi(\boldsymbol{\theta})$ that has some nice properties. The value $\delta(\mathbf{x})$ of $\delta(\mathbf{X})$ for the data $\mathbf{x}$ is called the estimate of $\psi(\boldsymbol{\theta})$.

In most problems $X_1, X_2, \ldots, X_n$ are iid RVs with common DF $F_{\boldsymbol{\theta}}$.

Example 1. Let $X_1, X_2, \ldots, X_n$ be iid $G(1, \theta)$, where $\Theta = \{\theta > 0\}$ and θ is to be estimated. Then $\mathfrak{X} = R_n$ and any map $\delta : \mathfrak{X} \to (0, \infty)$ is an estimator of θ. Some typical estimators of θ are $\bar{X} = n^{-1} \sum_{j=1}^{n} X_j$, and $\{2/[n(n+1)]\} \sum_{j=1}^{n} j X_j$.

Example 2. Let $X_1, X_2, \ldots, X_n$ be iid $b(1, p)$ RVs where $p \in [0, 1]$. Then $\bar{X}$ is an estimator of p and so also are $\delta_1(\mathbf{X}) = X_1$, $\delta_2(\mathbf{X}) = (X_1 + X_n)/2$, and $\delta_3(\mathbf{X}) = \sum_{j=1}^{n} a_j X_j$, where $0 \le a_j \le 1$, $\sum_{j=1}^{n} a_j = 1$.

It is clear that in any given problem of estimation we may have a large, often an infinite class of appropriate estimators to choose from. Clearly, we would like the estimator δ to be close to $\psi(\boldsymbol{\theta})$, and since δ is a statistic, the usual measure of closeness $|\delta(\mathbf{X}) - \psi(\boldsymbol{\theta})|$ is also an RV, we interpret "δ close to ψ" to mean "close on the average." Examples of such measures of closeness are

$$P_{\boldsymbol{\theta}}\{|\delta(\mathbf{X}) - \psi(\boldsymbol{\theta})| < \varepsilon\} \tag{1}$$

for some $\varepsilon > 0$, and

$$E_{\boldsymbol{\theta}}|\delta(\mathbf{X}) - \psi(\boldsymbol{\theta})|^r \tag{2}$$

for some $r > 0$. Obviously, we want (1) to be large but (2) to be small. For $r = 2$, the quantity defined in (2) is called *mean square error* and we denote it by

$$\mathrm{MSE}_{\boldsymbol{\theta}}(\delta) = E_{\boldsymbol{\theta}}\{\delta(\mathbf{X}) - \psi(\boldsymbol{\theta})\}^2. \tag{3}$$

Among all estimators for ψ, we would like to choose one, say δ_0, such that

$$P_{\boldsymbol{\theta}}\{|\delta_0(\mathbf{X}) - \psi(\boldsymbol{\theta})| < \varepsilon\} \geq P_{\boldsymbol{\theta}}\{|\delta(\mathbf{X}) - \psi(\boldsymbol{\theta})| < \varepsilon\} \tag{4}$$

for all δ, all $\varepsilon > 0$, and all $\boldsymbol{\theta}$. For (2), the requirement is to choose δ_0 such that

$$\text{MSE}_{\boldsymbol{\theta}}(\delta_0) \leq \text{MSE}_{\boldsymbol{\theta}}(\delta) \tag{5}$$

for all δ and all $\boldsymbol{\theta} \in \Theta$. Estimators satisfying (4) or (5) do not generally exist.

We note that

$$\begin{aligned} \text{MSE}_{\boldsymbol{\theta}}(\delta) &= E_{\boldsymbol{\theta}}[\delta(\mathbf{X}) - E_{\boldsymbol{\theta}}\delta(\mathbf{X})]^2 + [E_{\boldsymbol{\theta}}\delta(\mathbf{X}) - \psi(\boldsymbol{\theta})]^2 \\ &= \text{var}_{\boldsymbol{\theta}}\, \delta(\mathbf{X}) + \{b(\delta, \psi)\}^2, \end{aligned} \tag{6}$$

where

$$b(\delta, \psi) = E_{\boldsymbol{\theta}}\delta(\mathbf{X}) - \psi(\boldsymbol{\theta}), \tag{7}$$

is called the *bias* of δ. An estimator that has small MSE has small bias and variance. To control MSE, we need to control both variance and bias.

One approach is to restrict attention to estimators which have zero bias, that is,

$$E_{\boldsymbol{\theta}}\delta(\mathbf{X}) = \psi(\boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta} \in \Theta. \tag{8}$$

The condition of *unbiasedness* (8) ensures that on average, the estimator δ has no systematic error; it neither over- nor underestimates ψ on average. If we restrict attention to the class of unbiased estimators, we need to find an estimator δ_0 in this class such that δ_0 has the least variance for all $\boldsymbol{\theta} \in \Theta$. The theory of unbiased estimation is developed in Section 8.4.

Another approach is to replace $|\delta - \psi|^r$ in (2) by a more general function. Let $L(\boldsymbol{\theta}, \delta)$ measure the loss in estimating ψ by δ. Assume that L, the *loss function*, satisfies $L(\boldsymbol{\theta}, \delta) \geq 0$ for all $\boldsymbol{\theta}$ and δ, and $L(\boldsymbol{\theta}, \psi(\boldsymbol{\theta})) = 0$ for all $\boldsymbol{\theta}$. Measure average loss by the *risk function*

$$R(\boldsymbol{\theta}, \delta) = E_{\theta}L(\boldsymbol{\theta}, \delta(\mathbf{X})). \tag{9}$$

Instead of seeking an estimator that minimizes R, the risk, uniformly in θ, we minimize

$$\int R(\boldsymbol{\theta}, \delta)\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{10}$$

for some weight function π on Θ and minimize

$$\sup_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \delta). \tag{11}$$

The estimator that minimizes the average risk defined in (10) leads to the Bayes estimator, and the estimator that minimizes (11) leads to the minimax estimator. Bayes and minimax estimation are discussed in Section 8.8.

Sometimes there are symmetries in the problem which may be used to restrict attention to estimators that exhibit the same symmetry. Consider, for example, an experiment in which the length of life of a light bulb is measured. Then an estimator obtained from the measurements expressed in hours and minutes must agree with an estimator obtained from the measurements expressed in minutes. If $\mathbf{X}$ represents measurements in original units (hours) and $\mathbf{Y}$ represents corresponding measurements in transformed units (minutes), $\mathbf{Y} = c\mathbf{X}$ (here $c = 60$). If $\delta(\mathbf{X})$ is an estimator of the true mean, we would expect $\delta(\mathbf{Y})$, the estimator of the true mean, to correspond to $\delta(\mathbf{X})$ according to the relation $\delta(\mathbf{Y}) = c\delta(\mathbf{X})$. That is, $\delta(c\mathbf{X}) = c\delta(\mathbf{X})$ for all $c > 0$. This is an example of an *equivariant estimator*, a topic under extensive discussion in Section 8.9.

Finally, we consider some large-sample properties of estimators. As the sample size $n \to \infty$, the data $\mathbf{x}$ are practically the whole population, and we should expect $\delta(\mathbf{X})$ to approach $\psi(\boldsymbol{\theta})$ in some sense. For example, if $\delta(\mathbf{X}) = \overline{X}$, $\psi(\theta) = E_\theta X_1$, and $X_1, X_2, \dots, X_n$ are iid RVs with finite mean, the strong law of large numbers tells us that $\overline{X} \to E_\theta X_1$ with probability 1. This property of a sequence of estimators is called *consistency*.

Definition 3. Let $X_1, X_2, \dots$ be a sequence of iid RVs with common DF $F_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$. A sequence of point estimators $T_n(X_1, X_2, \dots, X_n) = T_n$ will be called *consistent* for $\psi(\boldsymbol{\theta})$ if

$$T_n \xrightarrow{P} \psi(\boldsymbol{\theta}) \qquad \text{as } n \to \infty$$

for each fixed $\boldsymbol{\theta} \in \Theta$.

Remark 1. Recall that $T_n \xrightarrow{P} \psi(\boldsymbol{\theta})$ if and only if $P\{|T_n - \psi(\boldsymbol{\theta})| > \varepsilon\} \to 0$ as $n \to \infty$ for every $\varepsilon > 0$. One can similarly define *strong consistency* of a sequence of estimators T_n if $T_n \xrightarrow{\text{a.s.}} \psi(\boldsymbol{\theta})$. Sometimes, one speaks of *consistency in the rth mean* when $T_n \xrightarrow{r} \psi(\boldsymbol{\theta})$. In what follows, *consistency* will mean *weak consistency* of T_n for $\psi(\boldsymbol{\theta})$, that is, $T_n \xrightarrow{P} \psi(\boldsymbol{\theta})$.

It is important to remember that consistency is a large-sample property. Moreover, we speak of consistency of a sequence of estimators rather than one point estimator.

Example 3. Let $X_1, X_2, \dots$ be iid $b(1, p)$ RVs. Then $EX_1 = p$ and it follows by the WLLN that

$$\frac{\sum_1^n X_i}{n} \xrightarrow{P} p.$$

Thus $\overline{X}$ is consistent for p. Also, $(\sum_1^n X_i + 1)/(n+2) \xrightarrow{P} p$, so that a consistent estimator need not be unique. Indeed, if $T_n \xrightarrow{P} p$ and $c_n \to 0$ as $n \to \infty$, then $T_n + c_n \xrightarrow{P} p$.

Theorem 1. If $X_1, X_2 \ldots$ are iid RVs with common law $\mathcal{L}(X)$, and $E|X|^p < \infty$ for some positive integer p, then

$$\frac{\sum_1^n X_i^k}{n} \xrightarrow{P} EX^k \qquad \text{for } 1 \le k \le p,$$

and $n^{-1}\sum_1^n X_i^k$ is consistent for EX^k, $1 \le k \le p$. Moreover, if c_n is any sequence of constants such that $c_n \to 0$ as $n \to \infty$, then $(n^{-1}\sum X_i^k + c_n)$ is also consistent for EX^k, $1 \le k \le p$. Also, if $c_n \to 1$ as $n \to \infty$, then $\left(c_n n^{-1}\sum X_i^k\right)$ is consistent for EX^k. This is simply a restatement of the WLLN for iid RVs.

Example 4. Let $X_1, X_2, \ldots$ be iid $\mathcal{N}(\mu, \sigma^2)$ RVs. If S^2 is the sample variance, we know that $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$. Thus $E(S^2/\sigma^2) = 1$ and $\text{var}(S^2/\sigma^2) = 2/(n-1)$. It follows that

$$P\{|S^2 - \sigma^2| > \varepsilon\} \le \frac{\text{var}(S^2)}{\varepsilon^2} = \frac{2\sigma^4}{(n-1)\varepsilon^2} \to 0 \qquad \text{as } n \to \infty.$$

Thus $S^2 \xrightarrow{P} \sigma^2$. Actually, this result holds for any sequence of iid RVs with $E|X|^2 < \infty$ and can be obtained from Theorem 1.

Example 4 is a particular case of the following theorem.

Theorem 2. If T_n is a sequence of estimators such that $ET_n \to \psi(\boldsymbol{\theta})$ and $\text{var}(T_n) \to 0$ as $n \to \infty$, then T_n is consistent for $\psi(\boldsymbol{\theta})$.

Proof. We have

$$\begin{aligned} P\{1T_n - \psi(\boldsymbol{\theta})| > \varepsilon\} &\le \varepsilon^{-2}E[T_n - ET_n + ET_n - \psi(\boldsymbol{\theta})]^2 \\ &= \varepsilon^{-2}\{\text{var}(T_n) + [ET_n - \psi(\boldsymbol{\theta})]^2\} \to 0 \qquad \text{as } n \to \infty. \end{aligned}$$

Other large-sample properties of estimators are asymptotic unbiasedness, asymptotic normality, and asymptotic efficiency. A sequence of estimators $\{T_n\}$ is *asymptotically unbiased* for $\psi(\boldsymbol{\theta})$ if

$$\lim_{n\to\infty} E_{\boldsymbol{\theta}} T_n(\mathbf{X}) = \psi(\boldsymbol{\theta})$$

for all $\boldsymbol{\theta}$. A consistent sequence of estimators $\{T_n\}$ is said to be *consistent asymptotically normal* (CAN) for $\psi(\boldsymbol{\theta})$ if $T_n \sim AN(\psi(\boldsymbol{\theta}), v(\boldsymbol{\theta})/n)$ for all $\boldsymbol{\theta} \in \Theta$. If

$v(\theta) = 1/I(\theta)$, where $I(\theta)$ is the Fisher information (Section 8.7), then $\{T_n\}$ is known as a *best asymptotically normal* (BAN) estimator.

Example 5. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\theta, 1)$ RVs. Then $T_n = \sum_{i=1}^n X_i/(n+1)$ is asymptotically unbiased for θ and BAN estimator for θ with $v(\theta) = 1$.

In Section 8.7 we consider large-sample properties of maximum likelihood estimators, and in Section 8.5 asymptotic efficiency is introduced.

PROBLEMS 8.2

1. Suppose that T_n is a sequence of estimators for parameter θ that satisfies the conditions of Theorem 2. Then $T_n \xrightarrow{2} \theta$, that is, T_n is squared-error consistent for θ. If T_n is consistent for θ and $|T_n - \theta| \le A < \infty$ for all θ and all $(x_1, x_2, \ldots, x_n) \in \mathcal{R}_n$, show that $T_n \xrightarrow{2} \theta$. If, however, $|T_n - \theta| \le A_n < \infty$, show that T_n may not be squared-error consistent for θ.

2. Let $X_1, X_2, \ldots, X_n$ be a sample from $U[0, \theta], \theta \in \Theta = (0, \infty)$. Let $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$. Show that $X_{(n)} \xrightarrow{P} \theta$. Write $Y_n = 2\overline{X}$. Is Y_n consistent for θ?

3. Let $X_1, X_2, \ldots, X_n$ be iid RVs with $EX_i = \mu$ and $E|X_i|^2 < \infty$. Show that $T(X_1, X_2, \ldots, X_n) = 2[n(n+1)]^{-1} \sum_{i=1}^n iX_i$ is a consistent estimator for μ.

4. Let $X_1, X_2, \ldots, X_n$ be a sample from $U[0, \theta]$. Show that $T(X_1, X_2, \ldots, X_n) = (\prod_{i=1}^n X_i)^{1/n}$ is a consistent estimator for θe^{-1}.

5. In Problem 2, show that $T(\mathbf{X}) = X_{(n)}$ is asymptotically biased for θ and is not BAN. [Show that $n(\theta - X_{(n)}) \xrightarrow{L} G(1, \theta)$.]

6. In Problem 5, consider the class of estimators $T(\mathbf{X}) = cX_{(n)}, c > 0$. Show that the estimator $T_0(\mathbf{X}) = (n+2)X_{(n)}/(n+1)$ in this class has the least MSE.

7. Let $X_1, X_2, \ldots, X_n$ be iid with PDF $f_\theta(x) = \exp\{-(x - \theta)\}, x > \theta$. Consider the class of estimators $T(\mathbf{X}) = X_{(1)} + b, b \in \mathcal{R}$. Show that the estimator that has the smallest MSE in this class is given by $T(\mathbf{X}) = X_{(1)} - 1/n$.

8.3 SUFFICIENCY, COMPLETENESS, AND ANCILLARITY

After the completion of any experiment, the job of a statistician is to interpret the data she has collected and to draw some statistically valid conclusions about the population under investigation. In adddition to being costly to store, the raw data by themselves are not suitable for this purpose. Therefore, the statistician would like to condense the data by computing some statistics from them and to base her analysis on these statistics, provided that there is "no loss of information" in doing so. In many problems of statistical inference a function of the observations contains as much information about the unknown parameter as do all the observed values. The following example illustrates this point.

Example 1. Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(\mu, 1)$, where μ is unknown. Suppose that we transform variables $X_1, X_2, \ldots, X_n$ to $Y_1, Y_2, \ldots, Y_n$ with the help of an orthogonal transformation so that Y_1 is $\mathcal{N}(\sqrt{n}\,\mu, 1)$, $Y_2, \ldots, Y_n$ are iid $\mathcal{N}(0, 1)$, and $Y_1, Y_2, \ldots, Y_n$ are independent. (Take $y_1 = \sqrt{n}\,\overline{x}$, and for $k = 2, \ldots, n$, $y_k = [(k-1)x_k - (x_1 + \cdots + x_{k-1})]/\sqrt{k(k-1)}$.) To estimate μ we can use either the observed values of $X_1, X_2, \ldots, X_n$ or simply the observed value of $Y_1 = \sqrt{n}\,\overline{X}$. The RVs $Y_2, Y_3, \ldots, Y_n$ provide no information about μ. Clearly, Y_1 is preferable since one need not keep a record of all the observations; it suffices to accumulate the observations and compute y_1. Any analysis of the data based on y_1 is just as effective as any analysis that could be based on x_i's. We note that Y_1 takes values in $\mathcal{R}_1$, whereas $(X_1, X_2, \ldots, X_n)$ takes values in $\mathcal{R}_n$.

A rigorous definition of the concept involved in the discussion above requires the notion of a conditional distribution and is beyond the scope of this book. In view of the discussion of conditional probability distributions in Section 4.2, the following definition will suffice for our purposes.

Definition 1. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a sample from $\{F_\theta : \theta \in \Theta\}$. A statistic $T = T(\mathbf{X})$ is *sufficient* for θ or for the family of distributions $\{F_\theta : \theta \in \Theta\}$ if and only if the conditional distribution of X, given $T = t$, does not depend on θ (except perhaps for a null set A, $P_\theta\{T \in A\} = 0$ for all θ).

Remark 1. The outcome $X_1, X_2, \ldots, X_n$ is always sufficient, but we will exclude this trivial statistic from consideration. According to Definition 1, if T is sufficient for θ, we need only concentrate on T since it exhausts all the information that the sample has about θ. In practice, there will be several sufficient statistics for a family of distributions, and the question arises as to which of these should be used in a given problem. We will return to this topic in more detail later in this section.

Example 2. We show that the statistic Y_1 in Example 1 is sufficient for μ. By construction $Y_2, \ldots, Y_n$ are iid $\mathcal{N}(0, 1)$ RVs that are independent of Y_1. Hence the conditional distribution of $Y_2, \ldots, Y_n$, given $Y_1 = \sqrt{n}\,\overline{X}$, is the same as the unconditional distribution of $(Y_2, \ldots, Y_n)$, which is multivariate normal with mean $(0, 0, \ldots, 0)$ and dispersion matrix $\mathbf{I}_{n-1}$. Since this distribution is independent of μ, the conditional distribution of $(Y_1, Y_2, \ldots, Y_n)$, and hence $(X_1, X_2, \ldots, X_n)$, given $Y_1 = y_1$, is also independent of μ and Y_1 is sufficient.

Example 3. Let $X_1, X_2, \ldots, X_n$ be iid $b(1, p)$ RVs. Intuitively, if a loaded coin is tossed with probability p of heads n times, it seems unnecessary to know which toss resulted in a head. To estimate p, it should be sufficient to know the number of heads in n trials. We show that this is consistent with our definition. Let $T(X_1, X_2,$

$\ldots, X_n) = \sum_{i=1}^n X_i$. Then

$$P\left\{X_1 = x_1, \ldots, X_n = x_n \,\middle|\, \sum_{i=1}^n X_i = t\right\} = \frac{P\{X_1 = x_1, \ldots, X_n = x_n, T = t\}}{\binom{n}{t} p^t (1-p)^{n-t}},$$

if $\sum_1^n x_i = t$, and $= 0$ otherwise. Thus, for $\sum_1^n x_i = t$, we have

$$P\{X_1 = x_1, \ldots, X_n = x_n \mid T = t\} = \frac{p^{\sum_1^n x_i}(1-p)^{n-\sum x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}},$$

which is independent of p. It is therefore sufficient to concentrate on $\sum_1^n X_i$.

Example 4. Let X_1, X_2 be iid $P(\lambda)$ RVs. Then $X_1 + X_2$ is sufficient for λ, for

$$P\{X_1 = x_1, X_2 = x_2 \mid X_1 + X_2 = t\}$$
$$= \begin{cases} \dfrac{P\{X_1 = x_1, X_2 = t - x_1\}}{P\{X_1 + X_2 = t\}} & \text{if } t = x_1 + x_2, x_i = 0, 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, for $x_i = 0, 1, 2, \ldots, i = 1, 2, x_1 + x_2 = t$, we have

$$P\{X_1 = x_1, X_2 = x_2 \mid X_1 + X_2 = t\} = \binom{t}{x_1}\left(\frac{1}{2}\right)^t,$$

which is independent of λ.

Not every statistic is sufficient.

Example 5. Let X_1, X_2 be iid $P(\lambda)$ RVs, and consider the statistic $T = X_1 + 2X_2$. We have

$$\begin{aligned} P\{X_1 = 0, X_2 = 1 \mid X_1 + 2X_2 = 2\} &= \frac{P\{X_1 = 0, X_2 = 1\}}{P\{X_1 + 2X_2 = 2\}} \\ &= \frac{e^{-\lambda}(\lambda e^{-\lambda})}{P\{X_1 = 0, X_2 = 1\} + P\{X_1 = 2, X_2 = 0\}} \\ &= \frac{\lambda e^{-2\lambda}}{\lambda e^{-2\lambda} + (\lambda^2/2)e^{-2\lambda}} = \frac{1}{1 + (\lambda/2)}, \end{aligned}$$

and we see that $X_1 + 2X_2$ is not sufficient for λ.

Definition 1 is not a constructive definition since it requires that we first guess a statistic T and then check to see whether T is sufficient. Moreover, the procedure for checking that T is sufficient is quite time consuming. We now give a criterion for determining sufficient statistics.

Theorem 1 (Factorization Criterion). Let $X_1, X_2, \ldots, X_n$ be discrete RVs with PMF $p_\theta(x_1, x_2, \ldots, x_n)$, $\theta \in \Theta$. Then $T(X_1, X_2, \ldots, X_n)$ is sufficient for θ if and only if we can write

$$(1) \qquad p_\theta(x_1, x_2, \ldots, x_n) = h(x_1, x_2, \ldots, x_n) g_\theta(T(x_1, x_2, \ldots, x_n)),$$

where h is a nonnegative function of $x_1, x_2, \ldots, x_n$ only and does not depend on θ, and g_θ is a nonnegative nonconstant function of θ and $T(x_1, x_2, \ldots, x_n)$ only. The statistic $T(X_1, \ldots, X_n)$ and parameter θ may be multidimensional.

Proof. Let T be sufficient for θ. Then $P\{\mathbf{X} = \mathbf{x} \mid T = t\}$ is independent of θ, and we may write

$$\begin{aligned} P_\theta\{\mathbf{X} = \mathbf{x}\} &= P_\theta\{\mathbf{X} = \mathbf{x}, T(X_1, X_2, \ldots, X_n) = t\} \\ &= P_\theta\{T = t\} P\{\mathbf{X} = \mathbf{x} \mid T = t\}, \end{aligned}$$

provided that $P\{\mathbf{X} = \mathbf{x} \mid T = t\}$ is well defined.

For values of $\mathbf{x}$ for which $P_\theta\{\mathbf{X} = \mathbf{x}\} = 0$ for all θ, let us define $h(x_1, x_2, \ldots, x_n) = 0$, and for $\mathbf{x}$ for which $P_\theta\{\mathbf{X} = \mathbf{x}\} > 0$ for some θ, we define

$$h(x_1, x_2, \ldots, x_n) = P\{X_1 = x_1, \ldots, X_n = x_n \mid T = t\}$$

and define

$$g_\theta(T(x_1, x_2, \ldots, x_n)) = P_\theta\{T(x_1, \ldots, x_n) = t\}.$$

Thus we see that (1) holds.

Conversely, suppose that (1) holds. Then for fixed t_0 we have

$$\begin{aligned} P_\theta\{T = t_0\} &= \sum_{\mathbf{x}:\, T(\mathbf{x}) = t_0} P_\theta\{\mathbf{X} = \mathbf{x}\} \\ &= \sum_{\mathbf{x}:\, T(\mathbf{x}) = t_0} g_\theta(T(\mathbf{x})) h(\mathbf{x}) \\ &= g_\theta(t_0) \sum_{T(\mathbf{x}) = t_0} h(\mathbf{x}). \end{aligned}$$

Suppose that $P_\theta\{T = t_0\} > 0$ for some $\theta > 0$. Then

$$P_\theta\{\mathbf{X} = \mathbf{x} \mid T = t_0\} = \frac{P_\theta\{\mathbf{X} = \mathbf{x}, T(\mathbf{x}) = t_0\}}{P_\theta\{T(\mathbf{x}) = t_0\}} = \begin{cases} 0 & \text{if } T(\mathbf{x}) \neq t_0, \\ \dfrac{P_\theta\{\mathbf{X} = \mathbf{x}\}}{P_\theta\{T(\mathbf{x}) = t_0\}} & \text{if } T(\mathbf{x}) = t_0. \end{cases}$$

Thus, if $T(\mathbf{x}) = t_0$, then

$$\frac{P_\theta\{\mathbf{X} = \mathbf{x}\}}{P_\theta\{T(\mathbf{x}) = t_0\}} = \frac{g_\theta(t_0)h(\mathbf{x})}{g_\theta(t_0)\sum_{T(\mathbf{x})=t_0} h(\mathbf{x})},$$

which is free of θ, as asserted. This completes the proof.

Remark 2. Theorem 1 also holds for the continuous case and, indeed, for quite arbitrary families of distributions. The general proof is beyond the scope of this book, and we refer the reader to Halmos and Savage [38] or to Lehmann [63, pp. 53–56]. We will assume that the result holds for the absolutely continuous case. We leave the reader to write the analog of (1) and to prove it, at least under the regularity conditions assumed in Theorem 4.4.2.

Remark 3. Theorem 1 (and its analog for the continuous case) holds if θ is a vector of parameters and T is a multiple RV, and we say that T is *jointly sufficient* for θ. We emphasize that even if θ is scalar, T may be multidimensional (Example 9). If θ and T are of the same dimension, and if T is sufficient for θ, it does not follow that the jth component of T is sufficient for the jth component of θ (Example 8). The converse is true under mild conditions (see Fraser [29, p. 21]).

Remark 4. If T is sufficient for θ, any one-to-one function of T is also sufficient. This follows from Theorem 1 since if $U = k(T)$ is a one-to-one function of T, then $t = k^{-1}(u)$, and we can write

$$f_\theta(\mathbf{x}) = g_\theta(t)h(\mathbf{x}) = g_\theta(k^{-1}(u))h(\mathbf{x}) = g_\theta^*(u)h(\mathbf{x}).$$

If T_1, T_2 are two distinct sufficient statistics, then

$$f_\theta(\mathbf{x}) = g_\theta(t_1)h_1(\mathbf{x}) = g_\theta(t_2)h_2(\mathbf{x}),$$

and it follows that T_1 is a function of T_2. It does not follow, however, that every function of a sufficient statistic is itself sufficient. For example, in sampling from a normal population, $\overline{X}$ is sufficient for the mean μ but $\overline{X}^2$ is not. Note that $\overline{X}$ is sufficient for μ^2.

Remark 5. As a rule, Theorem 1 cannot be used to show that a given statistic T is not sufficient. To do this, one would normally have to use the definition of sufficiency. In most cases Theorem 1 will lead to a sufficient statistic if it exists.

Remark 6. If $T(\mathbf{X})$ is sufficient for $\{F_\theta : \theta \in \Theta\}$, then T is sufficient for $\{F_\theta : \theta \in \omega\}$, where $\omega \subseteq \Theta$. This follows trivially from the definition.

Example 6. Let $X_1, X_2, \dots, X_n$ be iid $b(1, p)$ RVs. Then $T = \sum_{i=1}^n X_i$ is sufficient. We have

$$P_p\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} = p^{\sum_1^n x_i}(1-p)^{n-\sum_1^n x_i},$$

and taking

$$h(x_1, x_2, \ldots, x_n) = 1 \quad \text{and} \quad g_p(x_1, x_2, \ldots, x_n) = (1-p)^n \left(\frac{p}{1-p}\right)^{\sum_{i=1}^n x_i},$$

we see that T is sufficient. We note that $T_1(\mathbf{X}) = (X_1, X_2 + X_3 + \cdots + X_n)$ and $T_2(\mathbf{X}) = (X_1 + X_2, X_3, X_4 + X_5 + \cdots + X_n)$ are also sufficient for p, although T is preferable to T_1 or T_2.

Example 7. Let $X_1, X_2, \ldots, X_n$ be iid RVs with common PMF

$$P\{X_i = k\} = \frac{1}{N}, \qquad k = 1, 2, \ldots, N; \quad i = 1, 2, \ldots, n.$$

Then

$$\begin{aligned} P_N\{X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n\} &= \frac{1}{N^n} \qquad \text{if } 1 \le k_1, \ldots, k_n \le N, \\ &= \frac{1}{N^n}\varphi(1, \min_{1\le i\le n} k_i)\varphi(\max_{1\le i\le n} k_i, N), \end{aligned}$$

where $\varphi(a, b) = 1$ if $b \ge a$, and $= 0$ if $b < a$. It follows, by taking $g_N[\max(k_1, \ldots k_n)] = (1/N^n)\varphi(\max_{1\le i\le n} k_i, N)$ and $h = \varphi(1, \min k_i)$, that $\max(X_1, X_2, \ldots, X_n)$ is sufficient for the family of joint PMFs P_N.

Example 8. Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. The joint PDF of $(X_1, X_2, \ldots, X_n)$ is

$$\begin{aligned} f_{\mu,\sigma^2}(\mathbf{x}) &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[-\frac{\sum(x_i - \mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{\sum_1^n x_i^2}{2\sigma^2} + \frac{\mu\sum_1^n x_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right). \end{aligned}$$

It follows that the statistic

$$T(X_1, \ldots, X_n) = \left(\sum_1^n X_i, \sum_1^n X_i^2\right)$$

is jointly sufficient for the parameter (μ, σ^2). An equivalent sufficient statistic that is frequently used is $T_1(X_1, \ldots, X_n) = (\overline{X}, S^2)$. Note that $\overline{X}$ is not sufficient for μ if σ^2 is unknown, and S^2 is not sufficient for σ^2 if μ is unknown. If, however, σ^2 is known, $\overline{X}$ is sufficient for μ. If $\mu = \mu_0$ is known, $\sum_1^n (X_i - \mu_0)^2$ is sufficient for σ^2.

Example 9. Let $X_1, X_2, \dots, X_n$ be a sample from PDF

$$f_\theta(x) = \begin{cases} \dfrac{1}{\theta}, & x \in \left[-\dfrac{\theta}{2}, \dfrac{\theta}{2}\right], \quad \theta > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The joint PDF of $X_1, X_2, \dots, X_n$ is given by

$$f_\theta(x_1, x_2, \dots, x_n) = \frac{1}{\theta^n} I_A(x_1, \dots, x_n),$$

where

$$A = \left\{(x_1, x_2, \dots, x_n) : -\frac{\theta}{2} \le \min x_i \le \max x_i \le \frac{\theta}{2}\right\}.$$

It follows that $(X_{(1)}, X_{(n)})$ is sufficient for θ.

We note that the order statistic $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ is also sufficient. Note also that the parameter is one-dimensional, the statistics $(X_{(1)}, X_{(n)})$ is two-dimensional, and the order statistic is n-dimensional.

In Example 9 we saw that the order statistic is sufficient. This is not a mere coincidence. In fact, if $\mathbf{X} = (X_1, X_2, \dots, X_n)$ are exchangeable, the joint PDF of $\mathbf{X}$ is a symmetric function of its arguments. Thus

$$f_\theta(x_1, x_2, \dots, x_n) = f_\theta(x_{(1)}, x_{(2)}, \dots, x_{(n)}),$$

and it follows that the order statistic is sufficient for f_θ.

The concept of sufficiency is used frequently with another concept, called *completeness*, which we now define.

Definition 2. Let $\{f_\theta(x), \theta \in \Theta\}$ be a family of PDFs (or PMFs). We say that this family is *complete* if

$$E_\theta g(X) = 0 \qquad \text{for all } \theta \in \Theta$$

implies that

$$P_\theta\{g(X) = 0\} = 1 \qquad \text{for all } \theta \in \Theta.$$

Definition 3. A statistic $T(X)$ is said to be *complete* if the family of distributions of T is complete.

In Definition 3 X will usually be a multiple RV. The family of distributions of T is obtained from the family of distributions of $X_1, X_2, \dots, X_n$ by the usual transformation technique discussed in Section 4.4.

Example 10. Let $X_1, X_2, \ldots, X_n$ be iid $b(1, p)$ RVs. Then $T = \sum_1^n X_i$ is a sufficient statistic. We show that T is also complete; that is, the family of distributions of T, $\{b(n, p), 0 < p < 1\}$, is complete.

$$E_p g(T) = \sum_{t=0}^{n} g(t)\binom{n}{t} p^t (1-p)^{n-t} = 0 \qquad \text{for all } p \in (0, 1)$$

may be rewritten as

$$(1-p)^n \sum_{t=0}^{n} g(t)\binom{n}{t}\left(\frac{p}{1-p}\right)^t = 0 \qquad \text{for all } p \in (0, 1).$$

This is a polynomial in $p/(1-p)$. Hence the coefficients must vanish, and it follows that $g(t) = 0$ for $t = 0, 1, 2, \ldots, n$, as required.

Example 11. Let X be $\mathcal{N}(0, \theta)$. Then the family of PDFs $\{\mathcal{N}(0, \theta), \theta > 0\}$ is not complete since $EX = 0$ and $g(x) = x$ is not identically zero. Note that $T(X) = X^2$ is complete, for the PDF of $X^2 \sim \theta\chi^2(1)$ is given by

$$f(t) = \begin{cases} \dfrac{e^{-t/2\theta}}{\sqrt{2\pi\theta t}}, & t > 0, \\ 0, & \text{otherwise.} \end{cases}$$

$$E_\theta g(T) = \frac{1}{\sqrt{2\pi\theta}} \int_0^\infty g(t) t^{-1/2} e^{-t/2\theta}\, dt = 0 \qquad \text{for all } \theta > 0,$$

which holds if and only if $\int_0^\infty g(t) t^{-1/2} e^{-t/2\theta}\, dt = 0$, and using the uniqueness property of Laplace transforms, it follows that

$$g(t) t^{-1/2} = 0 \qquad \text{for all } t > 0,$$

that is, $g(t) = 0$.

The next example illustrates the existence of a sufficient statistic that is not complete.

Example 12. Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(\theta, \theta^2)$. Then $T = (\sum_1^n X_i, \sum_1^n X_i^2)$ is sufficient for θ. However, T is not complete since

$$E_\theta\left[2\left(\sum_1^n X_i\right)^2 - (n+1)\sum_1^n X_i^2\right] = 0 \qquad \text{for all } \theta,$$

and the function $g(x_1, \ldots, x_n) = 2(\sum_1^n x_i)^2 - (n+1)\sum_1^n x_j^2$ is not identically zero.

Example 13. Let $X \sim U(0, \theta)$, $\theta \in (0, \infty)$. We show that the family of PDFs of X is complete. We need to show that

$$E_\theta g(X) = \int_0^\theta \frac{1}{\theta} g(x)\,dx = 0 \qquad \text{for all } \theta > 0$$

if and only if $g(x) = 0$ for all x. In general, this result follows from Lebesgue integration theory. If g is continuous, we differentiate both sides in

$$\int_0^\theta g(x)\,dx = 0$$

to get $g(\theta) = 0$ for all $\theta > 0$.

Now let $X_1, X_2, \ldots, X_n$ be iid $U(0, \theta)$ RVs. Then the PDF of $X_{(n)}$ is given by

$$f_n(x \mid \theta) = \begin{cases} n\theta^{-n}x^{n-1}, & 0 < x < \theta, \\ 0, & \text{otherwise.} \end{cases}$$

We see by a similar argument that $X_{(n)}$ is complete, which is the same as saying that $\{f_n(x \mid \theta);\ \theta > 0\}$ is a complete family of densities. Clearly, $X_{(n)}$ is sufficient.

Example 14. Let $X_1, X_2, \ldots, X_n$ be a sample from PMF

$$P_N(x) = \begin{cases} \dfrac{1}{N}, & x = 1, 2, \ldots, N, \\ 0, & \text{otherwise.} \end{cases}$$

We first show that the family of PMFs $\{P_N, N \geq 1\}$ is complete. We have

$$E_N g(X) = \frac{1}{N} \sum_{k=1}^{N} g(k) = 0 \qquad \text{for all } N \geq 1,$$

and this happens if and only if $g(k) = 0$, $k = 1, 2, \ldots, N$. Next we consider the family of PMFs of $X_{(n)} = \max(X_1, \ldots, X_n)$. The PMF of $X_{(n)}$ is given by

$$P_N^{(n)}(x) = \frac{x^n}{N^n} - \frac{(x-1)^n}{N^n}, \qquad x = 1, 2, \ldots, N.$$

Also,

$$E_N g(X_{(n)}) = \sum_{k=1}^{N} g(k) \left[\frac{k^n}{N^n} - \frac{(k-1)^n}{N^n} \right] = 0 \qquad \text{for all } N \geq 1.$$

$$E_1 g(X_{(n)}) = g(1) = 0$$

implies that $g(1) = 0$. Again,

$$E_2 g(X_{(n)}) = \frac{g(1)}{2^n} + g(2)\left(1 - \frac{1}{2^n}\right) = 0$$

so that $g(2) = 0$.

Using an induction argument, we conclude that $g(1) = g(2) = \cdots = g(N) = 0$ and hence $g(x) = 0$. It follows that $P_N^{(n)}$ is a complete family of distributions, and $X_{(n)}$ is a complete sufficient statistic.

Now suppose that we exclude the value $N = n_0$ for some fixed $n_0 \geq 1$ from the family $\{P_N \colon N \geq 1\}$. Let us write $\mathcal{P} = \{P_N \colon N \geq 1,\ N \neq n_0\}$. Then $\mathcal{P}$ is not complete. We ask the reader to show that the class of all functions g such that $E_P g(X) = 0$ for all $P \in \mathcal{P}$ consists of functions of the form

$$g(k) = \begin{cases} 0, & k = 1, 2, \ldots, n_0 - 1, n_0 + 2, n_0 + 3, \ldots, \\ c, & k = n_0, \\ -c, & k = n_0 + 1, \end{cases}$$

where c is a constant, $c \neq 0$.

Remark 7. Completeness is a property of a family of distributions. In Remark 6 we saw that if a statistic is sufficient for a class of distributions, it is sufficient for any subclass of those distributions. Completeness works in the opposite direction. Example 14 shows that the exclusion of even one member from the family $\{P_N \colon N \geq 1\}$ destroys completeness.

The following result covers a large class of probability distributions for which a complete sufficient statistic exists.

Theorem 2. Let $\{f_\theta \colon \boldsymbol{\theta} \in \Theta\}$ be a k-parameter exponential family given by

$$(2) \qquad f_{\boldsymbol{\theta}}(\mathbf{x}) = \exp\left[\sum_{j=1}^{k} Q_j(\boldsymbol{\theta}) T_j(\mathbf{x}) + D(\boldsymbol{\theta}) + S(\mathbf{x})\right],$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta$, an interval in $\mathcal{R}_k$, $T_1, T_2, \ldots, T_k$, and S are defined on $\mathcal{R}_n$, $\mathbf{T} = (T_1, T_2, \ldots, T_k)$, and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $k \leq n$. Let $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_k)$, and suppose that the range of $\mathbf{Q}$ contains an open set in $\mathcal{R}_k$. Then

$$\mathbf{T} = (T_1(\mathbf{X}), T_2(\mathbf{X}), \ldots, T_k(\mathbf{X}))$$

is a complete sufficient statistic.

Proof. For a complete proof in a general setting, we refer the reader to Lehmann [63, pp. 142–143]. Essentially, the unicity of the Laplace transform is used on the probability distribution induced by $\mathbf{T}$. We will content ourselves here by proving the result for the $k = 1$ case when f_θ is a PMF.

Let us write $Q(\theta) = \theta$ in (2), and let $(\alpha, \beta) \subseteq \Theta$. We wish to show that

$$E_\theta g(T(\mathbf{X})) = \sum_t g(t)\, P_\theta\{T(\mathbf{X}) = t\}$$

$$(3) \qquad = \sum_t g(t) \exp[\theta t + D(\theta) + S^*(t)] = 0 \qquad \text{for all } \theta$$

implies that $g(t) = 0$.

Let us write $x^+ = x$ if $x \geq 0$, $= 0$ if $x < 0$, and $x^- = -x$ if $x < 0$, $= 0$ if $x \geq 0$. Then $g(t) = g^+(t) - g^-(t)$, and both g^+ and g^- are nonnegative functions. In terms of g^+ and g^-, (3) is the same as

$$(4) \qquad \sum_t g^+(t) e^{\theta t + S^*(t)} = \sum_t g^-(t) e^{\theta t + S^*(t)}$$

for all θ.

Let $\theta_0 \in (\alpha, \beta)$ be fixed, and write

$$(5) \qquad p^+(t) = \frac{g^+(t) e^{\theta_0 t + S^*(t)}}{\sum_t g^+(t) e^{\theta_0 t + S^*(t)}} \quad \text{and} \quad p^-(t) = \frac{g^-(t) e^{\theta_0 t + S^*(t)}}{\sum_t g^-(t) e^{\theta_0 t + S^*(t)}}.$$

Then both p^+ and p^- are PMFs, and it follows from (4) that

$$(6) \qquad \sum_t e^{\delta t} p^+(t) = \sum_t e^{\delta t} p^-(t)$$

for all $\delta \in (\alpha - \theta_0, \beta - \theta_0)$. By the uniqueness of MGFs (6) implies that

$$p^+(t) = p^-(t) \qquad \text{for all } t$$

and hence that $g^+(t) = g^-(t)$ for all t, which is equivalent to $g(t) = 0$ for all t. Since T is clearly sufficient (by the factorization criterion), it is proved that T is a complete sufficient statistic.

Example 15. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$ RVs where both μ and σ^2 are unknown. We know that the family of distributions of $\mathbf{X} = (X_1, \ldots, X_n)$ is a two-parameter exponential family with $T(X_1, \ldots, X_n) = (\sum_1^n X_i, \sum_1^n X_i^2)$. From Theorem 2 it follows that T is a complete sufficient statistic. Examples 10 and 11 fall in the domain of Theorem 2.

In Examples 6, 8, and 9 we have shown that a given family of probability distributions that admits a nontrivial sufficient statistic usually admits several sufficient statistics. Clearly, we would like to be able to choose the sufficient statistic that results in the greatest reduction of data collection. We next study the notion of a *minimal sufficient statistic*. For this purpose it is convenient to introduce the notion of a *sufficient partition*. The reader will recall that a *partition* of a space $\mathfrak{X}$ is just a collection of disjoint sets E_α such that $\sum_\alpha E_\alpha = \mathfrak{X}$. Any statistic $T(X_1, X_2, \ldots, X_n)$ induces a partition of the space of values of $(X_1, X_2, \ldots, X_n)$, that is, T induces a covering of $\mathfrak{X}$ by a family $\mathfrak{U}$ of disjoint sets $A_t = \{(x_1, x_2, \ldots, x_n) \in \mathfrak{X} : T(x_1, x_2, \ldots, x_n) = t\}$, where t belongs to the range of T. The sets A_t are called *partition sets*. Conversely, given a partition, any assignment of a number to each set so that no two partition sets have the same number assigned defines a statistic. Clearly, this function is not, in general, unique.

Definition 4. Let $\{F_\theta : \theta \in \Theta\}$ be a family of DFs, and $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a sample from F_θ. Let $\mathfrak{U}$ be a partition of the sample space induced by a statistic $T = T(X_1, X_2, \ldots, X_n)$. We say that $\mathfrak{U} = \{A_t : t$ is in the range of $T\}$ is a *sufficient partition* for θ (or the family $\{F_\theta : \theta \in \Theta\}$) if the conditional distribution of $\mathbf{X}$, given $T = t$, does not depend on θ for any A_t, provided that the conditional probability is well defined.

Example 16. Let $X_1, X_2, \ldots, X_n$ be iid $b(1, p)$ RVs. The sample space of values of $(X_1, X_2, \ldots, X_n)$ is the set of n-tuples $(x_1, x_2, \ldots, x_n)$, where each $x_i = 0$ or $= 1$ and consists of 2^n points. Let $T(X_1, X_2, \ldots, X_n) = \sum_1^n X_i$, and consider the partition $\mathfrak{U} = \{A_0, A_1, \ldots, A_n\}$, where $\mathbf{x} \in A_j$ if and only if $\sum_1^n x_i = j, 0 \le j \le n$. Each A_j contains $\binom{n}{j}$ sample points. The conditional probability

$$P_p\{\mathbf{x} \mid A_j\} = \frac{P_p\{\mathbf{x}\}}{P_p(A_j)} = \binom{n}{j}^{-1} \qquad \text{if } \mathbf{x} \in A_j,$$

and we see that $\mathfrak{U}$ is a sufficient partition.

Example 17. Let $X_1, X_2, \ldots, X_n$ be iid $U[0, \theta]$ RVs. Consider the statistic $T(\mathbf{X}) = \max_{1 \le i \le n} X_i$. The space of values of $X_1, X_2, \ldots, X_n$ is the set of points $\{\mathbf{x} : 0 \le x_i \le \theta, i = 1, 2, \ldots, n\}$. T induces a partition $\mathfrak{U}$ on this set. The sets of this partition are $A_t = \{(x_1, x_2, \ldots, x_n) : \max(x_1, \ldots, x_n) = t\}, t \in [0, \theta]$.

We have

$$f_\theta(\mathbf{x} \mid t) = \frac{f_\theta(\mathbf{x})}{f_\theta^T(t)} \qquad \text{if } \mathbf{x} \in A_t,$$

where $f_\theta^T(t)$ is the PDF of T. We have

$$f_\theta(\mathbf{x} \mid t) = \frac{1/\theta^n}{nt^{n-1}/\theta^n} = \frac{1}{nt^{n-1}} \qquad \text{if } \mathbf{x} \in A_t.$$

It follows that $\mathfrak{U} = \{A_t\}$ defines a sufficient partition.

Remark 8. Clearly, a sufficient statistic T for a family of DFs $\{F_\theta : \theta \in \Theta\}$ induces a sufficient partition; and conversely, given a sufficient partition, we can define a sufficient statistic (not necessarily uniquely) for the family.

Remark 9. Two statistics T_1, T_2 that define the same partition must be in one-to-one correspondence, that is, there exists a function h such that $T_1 = h(T_2)$ with a unique inverse, $T_2 = h^{-1}(T_1)$. It follows that if T_1 is sufficient, every one-to-one function of T_1 is also sufficient.

Let $\mathfrak{U}_1, \mathfrak{U}_2$ be two partitions of a space $\mathfrak{X}$. We say that $\mathfrak{U}_1$ is a *subpartition* of $\mathfrak{U}_2$ if every partition set in $\mathfrak{U}_2$ is a union of sets of $\mathfrak{U}_1$. We sometimes say also that $\mathfrak{U}_1$ is *finer* than $\mathfrak{U}_2$ ($\mathfrak{U}_2$ is *coarser* than $\mathfrak{U}_1$) or that $\mathfrak{U}_2$ is a *reduction* of $\mathfrak{U}_1$. In this case, a statistic T_2 that defines $\mathfrak{U}_2$ must be a function of any statistic T_1 that defines $\mathfrak{U}_1$. Clearly, this function need not have a unique inverse unless the two partitions have exactly the same partition sets.

Given a family of distributions $\{F_\theta : \theta \in \Theta\}$ for which a sufficient partition exists, we seek to find a sufficient partition $\mathfrak{U}$ that is as coarse as possible; that is, any reduction of $\mathfrak{U}$ leads to a partition that is not sufficient.

Definition 5. A partition $\mathfrak{U}$ is said to be *minimal sufficient* if

(i) $\mathfrak{U}$ is a sufficient partition, and

(ii) if C is any sufficient partition, C is a subpartition of $\mathfrak{U}$.

The question of the existence of the minimal partition was settled by Lehmann and Scheffé [62] and, in general, involves measure-theoretic considerations. However, in the cases that we consider where the sample space is either discrete or a finite-dimensional Euclidean space, and the family of distributions of $\mathbf{X}$ is defined by a family of PDFs (PMFs) $\{f_\theta, \theta \in \Theta\}$, such difficulties do not arise. The construction may be described as follows.

Two points $\mathbf{x}$ and $\mathbf{y}$ in the sample space are said to be *likelihood equivalent*, and we write $\mathbf{x} \sim \mathbf{y}$, if and only if there exists a $k(\mathbf{y}, \mathbf{x}) \neq 0$ which does not depend on θ such that $f_\theta(\mathbf{y}) = k(\mathbf{y}, \mathbf{x}) f_\theta(\mathbf{x})$. We leave the reader to check that "$\sim$" is an equivalence relation (that is, it is reflexive, symmetric, and transitive) and hence "$\sim$" defines a partition of the sample space. This partition defines the minimal sufficient partition.

Example 18. Consider Example 16 again. Then

$$\frac{f_p(\mathbf{x})}{f_p(\mathbf{y})} = p^{\sum x_i - \sum y_i} (1-p)^{-\sum x_i + \sum y_i},$$

and this ratio is independent of p if and only if

$$\sum_1^n x_i = \sum_1^n y_i,$$

so that $\mathbf{x} \sim \mathbf{y}$ if and only if $\sum_1^n x_i = \sum_1^n y_i$. It follows that the partition $\mathfrak{U} = \{A_0, A_1, \ldots, A_n\}$, where $\mathbf{x} \in A_j$ if and only if $\sum_1^n x_i = j$, introduced in Example 16, is minimal sufficient.

A rigorous proof of the assertion above is beyond the scope of this book. The basic ideas are outlined in the following theorem.

Theorem 3. The relation "$\sim$" defined above induces a minimal sufficient partition.

Proof. If T is a sufficient statistic, we have to show that $\mathbf{x} \sim \mathbf{y}$ whenever $T(\mathbf{x}) = T(\mathbf{y})$. This will imply that every set of the minimal sufficient partition is a union of sets of the form $A_t = \{T = t\}$, proving condition (ii) of Definition 5.

Sufficiency of T means that whenever $\mathbf{x} \in A_t$, then

$$f_\theta\{\mathbf{x} \mid T = t\} = \frac{f_\theta(\mathbf{x})}{f_\theta^T(t)} \qquad \text{if } \mathbf{x} \in A_t$$

is free of θ. It follows that if both $\mathbf{x}$ and $\mathbf{y} \in A_t$, then

$$\frac{f_\theta(\mathbf{x} \mid t)}{f_\theta(\mathbf{y} \mid t)} = \frac{f_\theta(\mathbf{x})}{f_\theta(\mathbf{y})}$$

is independent of θ, and hence $\mathbf{x} \sim \mathbf{y}$.

To prove the sufficiency of the minimal sufficient partition $\mathfrak{U}$, let T_1 be an RV that induces $\mathfrak{U}$. Then T_1 takes on distinct values over distinct sets of $\mathfrak{U}$ but remains constant on the same set. If $\mathbf{x} \in \{T_1 = t_1\}$, then

$$f_\theta(\mathbf{x} \mid T_1 = t_1) = \frac{f_\theta(\mathbf{x})}{P_\theta\{T_1 = t_1\}}. \tag{7}$$

Now

$$P_\theta\{T_1 = t_1\} = \int_{(\mathbf{y}:T_1(\mathbf{y})=t_1)} f_\theta(\mathbf{y})\,d\mathbf{y} \quad \text{or} \quad \sum_{(\mathbf{y}:T_1(\mathbf{y})=t_1)} f_\theta(\mathbf{y}),$$

depending on whether the joint distribution of $\mathbf{X}$ is absolutely continuous or discrete. Since $f_\theta(\mathbf{x})/f_\theta(\mathbf{y})$ is independent of θ whenever $\mathbf{x} \sim \mathbf{y}$, it follows that the ratio on the right-hand side of (7) does not depend on θ. Thus T_1 is sufficient.

Definition 6. A statistic that induces the minimal sufficient partition is called a *minimal sufficient statistic.*

In view of Theorem 3, a minimal sufficient statistic is a function of every sufficient statistic. It follows that if T_1 and T_2 are both minimal sufficient, then both must induce the same minimal sufficient partition, and hence T_1 and T_2 must be equivalent in the sense that each must be a function of the other (with probability 1).

How does one show that a statistic T is not sufficient for a family of distributions $\mathcal{P}$? Other than using the definition of sufficiency, one can sometimes use a result of Lehmann and Scheffé [62] according to which if $T_1(\mathbf{X})$ is sufficient for $\boldsymbol{\theta}$, $\boldsymbol{\theta} \in$

Θ, then $T_2(\mathbf{X})$ is also sufficient if and only if $T_1(\mathbf{X}) = g(T_2(\mathbf{X}))$ for some Borel-measurable function g and all $\mathbf{x} \in B$, where B is a Borel set with $P_{\boldsymbol{\theta}} B = 1$.

Another way to prove T nonsufficient is to show that there exist $\mathbf{x}$ for which $T(\mathbf{x}) = T(\mathbf{y})$ but $\mathbf{x}$ and $\mathbf{y}$ are not likelihood equivalent. We refer to Sampson and Spencer [96] for this and similar results.

The following important result is proved in the next section.

Theorem 4. A complete sufficient statistic is minimal sufficient.

We emphasize that the converse is not true. A minimal sufficient statistic may not be complete.

Example 19. Suppose that $X \sim U(\theta, \theta + 1)$. Then X is a minimal sufficient statistic. However, X is not complete. Take, for example, $g(x) = \sin 2\pi x$. Then

$$Eg(X) = \int_{\theta}^{\theta+1} \sin 2\pi x \, dx = \int_0^1 \sin 2\pi x \, dx = 0$$

for all θ, and it follows that X is not complete.

If $X_1, X_2, \ldots, X_n$ is a sample from $U(\theta, \theta + 1)$, then $(X_{(1)}, X_{(n)})$ is minimal sufficient for θ but not complete since

$$E_\theta(X_{(n)} - X_{(1)}) = \frac{n-1}{n+1}$$

for all θ.

Finally, we consider statistics that have distributions free of the parameter(s) $\boldsymbol{\theta}$ and seem to contain no information about $\boldsymbol{\theta}$. We will see (Example 23) that such statistics can sometimes provide useful information about $\boldsymbol{\theta}$.

Definition 7. A statistic $A(\mathbf{x})$ is said to be *ancillary* if its distribution does not depend on the underlying model parameter θ.

Example 20. Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, 1)$. Then the statistic $A(\mathbf{X}) = (n-1)S^2 = \sum_{i=1}^n (X_i - \overline{X})^2$ is ancillary since $(n-1)S^2 \sim \chi^2(n-1)$, which is free of μ. Some other ancillary statistics are

$$X_1 - \overline{X}, X_{(n)} - X_{(1)}, \quad \text{and} \quad \sum_{i=1}^n |X_i - \overline{X}|.$$

Also, $\overline{X}$, a complete sufficient statistic (hence minimal sufficient) for μ is independent of $A(\mathbf{X})$.

Example 21. Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(0, \sigma^2)$. Then $A(\mathbf{X}) = \overline{X}$ follows $\mathcal{N}(0, n^{-1}\sigma^2)$ and is not ancillary with respect to the parameter σ^2.

Example 22. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics of a random sample from the PDF $f(x - \theta)$, where $\theta \in \mathcal{R}$. Then the statistic $A(\mathbf{X}) = (X_{(2)} - X_{(1)}, \ldots X_{(n)} - X_{(1)})$ is ancillary for θ.

In Example 20 we saw that S^2 was independent of the minimal sufficient statistic $\overline{X}$. The following result due to Basu shows that it is not a mere coincidence.

Theorem 5. If $S(\mathbf{X})$ is a complete sufficient statistic for θ, then any ancillary statistic $A(\mathbf{X})$ is independent of S.

Proof. If A is ancillary, then $P_\theta\{A(\mathbf{X}) \le a\}$ is free of θ for all a. Consider the conditional probability $g_a(s) = P\{A(\mathbf{X}) \le a \mid S(\mathbf{X}) = s\}$. Clearly,

$$E_\theta\left\{g_a(S(\mathbf{X}))\right\} = P_\theta\{A(\mathbf{X}) \le a\}.$$

Thus

$$E_\theta(g_a(S) - P\{A(\mathbf{X}) \le a\}) = 0$$

for all θ. By completeness of S it follows that

$$P_\theta\{g_a(S) - P\{A \le a\} = 0\} = 1;$$

that is,

$$P_\theta\{A(\mathbf{X}) \le a \mid S(\mathbf{X}) = s\} = P\{A(\mathbf{X}) \le a\}$$

with probability 1. Hence A and S are independent.

The converse of Basu's theorem is not true. A statistic S that is independent of every ancillary statistic need not be complete (see, for example, Lehmann [60]).

The following example due to R. A. Fisher shows that if there is no sufficient statistic for θ but there exists a reasonable statistic not independent of an ancillary statistic $A(\mathbf{X})$, the recovery of information is sometimes helped by the ancillary statistic via a conditional analysis. Unfortunately, the lack of uniqueness of ancillary statistics creates problems with this conditional analysis.

Example 23. Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with mean θ, and let $Y_1, Y_2, \ldots, Y_n$ be another random sample from an exponential distribution and mean $1/\theta$. Assume that X's and Y's are independent and consider the problem of estimation of θ based on the observations $(X_1, X_2, \ldots, X_n;\ Y_1, Y_2, \ldots, Y_n)$. Let $S_1(\mathbf{x}) = \sum_{i=1}^n x_i$ and $S_2(\mathbf{y}) = \sum_{i=1}^n y_i$.

Then $\big(S_1(\mathbf{X}), S_2(\mathbf{Y})\big)$ is jointly sufficient for θ. It is easily seen that (S_1, S_2) is a minimal sufficient statistic for θ.

Consider the statistics

$$S(\mathbf{X}, \mathbf{Y}) = \left[\frac{S_1(\mathbf{X})}{S_2(\mathbf{Y})}\right]^{1/2},$$

and

$$A(\mathbf{X}, \mathbf{Y}) = S_1(\mathbf{X})S_2(\mathbf{Y}).$$

Then the joint PDF of S and A is given by

$$\frac{2}{[\Gamma(n)]^2}\exp\left[-A(\mathbf{x}, \mathbf{y})\left(\frac{S(\mathbf{x}, \mathbf{y})}{\theta} + \frac{\theta}{S(\mathbf{x}, \mathbf{y})}\right)\right]\frac{[A(\mathbf{x}, \mathbf{y})]^{2n-1}}{S(\mathbf{x}, \mathbf{y})}$$

and it is clear that S and A are not independent. The marginal distribution of A is given by the PDF

$$C(\mathbf{x}, \mathbf{y})[A(\mathbf{x}, \mathbf{y})]^{2n-1},$$

where $C(\mathbf{x}, \mathbf{y})$ is the constant of integration, which depends only on $\mathbf{x}$, $\mathbf{y}$, and n but not on θ. In fact, $C(\mathbf{x}, \mathbf{y}) = 4K_0[2A(\mathbf{x}, \mathbf{y})]/[\Gamma(n)]^2$, where K_0 is the standard form of a Bessel function (Watson [115]). Consequently A is ancillary for θ.

Clearly, the conditional PDF of S given $A = a$ is of the form

$$\frac{1}{2K_0[2a]S(\mathbf{x}, \mathbf{y})}\exp\left[-a\left(\frac{S(\mathbf{x}, \mathbf{y})}{\theta} + \frac{\theta}{S(\mathbf{x}, \mathbf{y})}\right)\right].$$

The amount of information lost by using $S(\mathbf{X}, \mathbf{Y})$ alone is the $[1/(2n + 1)]$th part of the total, and this loss of information is gained by knowledge of the ancillary statistic $A(\mathbf{X}, \mathbf{Y})$. These calculations are discussed in Example 8.5.9.

PROBLEMS 8.3

1. Find a sufficient statistic in each of the following cases based on a random sample of size n:

   (a) $X \sim B(\alpha, \beta)$ when (i) α is unknown, β known; (ii) β is unknown, α known; and (iii) α, β are both unknown.

   (b) $X \sim G(\alpha, \beta)$ when (i) α is unknown, β known; (ii) β is unknown, α known; and (iii) α, β are both unknown.

   (c) $X \sim P_{N_1, N_2}(x)$, where

$$P_{N_1,N_2}(x) = \frac{1}{N_2 - N_1}, \qquad x = N_1 + 1, N_1 + 2, \dots, N_2,$$

and $N_1, N_2 (N_1 < N_2)$ are integers, when (i) N_1 is known, N_2 unknown; (ii) N_2 known, N_1 unknown; and (iii) N_1, N_2 are both unknown.

(d) $X \sim f_\theta(x)$, where

$$f_\theta(x) = \begin{cases} e^{-x+\theta} & \text{if } < x < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

(e) $X \sim f(x; \mu, \sigma)$, where

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(\log x - \mu)^2\right], x > 0.$$

(f) $X \sim f_\theta(x)$, where

$$f_\theta(x) = P_\theta\{X = x\} = c(\theta)2^{-x/\theta}, \quad x = \theta, \theta + 1, \dots, \theta > 0,$$

and

$$c(\theta) = 2^{1-1/\theta}(2^{1/\theta} - 1).$$

(g) $X \sim P_{\theta,p}(x)$, where

$$P_{\theta,p}(x) = (1 - p)p^{x-\theta}, \qquad x = \theta,\ \theta + 1, \dots, 0 < p < 1,$$

when (i) p is known, θ unknown; (ii) p is unknown, θ known; and (iii) p, θ are both unknown.

2. Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be a sample from $\mathcal{N}(\alpha\sigma, \sigma^2)$, where α is a known real number. Show that the statistic $T(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient for σ but that the family of distributions of $T(\mathbf{X})$ is not complete.

3. Let $X_1, X_2, \dots, X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$. Then $\mathbf{X} = (X_1, X_2, \dots, X_n)$ is clearly sufficient for the family $\mathcal{N}(\mu, \sigma^2), \mu \in \mathcal{R}, \sigma > 0$. Is the family of distributions of $\mathbf{X}$ complete?

4. Let $X_1, X_2, \dots, X_n$ be a sample from $U(\theta - \frac{1}{2}, \theta + \frac{1}{2}), \theta \in \mathcal{R}$. Show that the statistic $T(X_1, \dots, X_n) = (\min X_i, \max X_i)$ is sufficient for θ but not complete.

5. If $T = g(U)$ and T is sufficient, so is U.

6. In Example 14, show that the class of all functions g for which $E_P g(X) = 0$ for all $P \in \mathcal{P}$ consists of functions of the form

$$g(k) = \begin{cases} 0, & k = 1, 2, \dots, n_0 - 1,\ n_0 + 2,\ n_0 + 3, \dots, \\ c, & k = n_0, \\ -c, & k = n_0 + 1, \end{cases}$$

where c is a constant.

7. For the class $\{F_{\theta_1}, F_{\theta_2}\}$ of two DFs where F_{θ_1} is $\mathcal{N}(0, 1)$ and F_{θ_2} is $C(1, 0)$, find a sufficient statistic.

8. Consider the class of hypergeometric probability distributions $\{P_D\colon D = 0, 1, 2, \ldots, N\}$, where

$$P_D\{X = x\} = \binom{N}{n}^{-1}\binom{D}{x}\binom{N-D}{n-x}, \qquad x = 0, 1, \ldots, \min\{n, D\}.$$

Show that it is a complete class. If $\mathcal{P} = \{P_D\colon D = 0, 1, 2, \ldots, N,\ D \neq d,\ d \text{ integral } 0 \le d \le N\}$, is $\mathcal{P}$ complete?

9. Is the family of distributions of the order statistic in sampling from a Poisson distribution complete?

10. Let $(X_1, X_2, \ldots, X_n)$ be a random vector of the discrete type. Is the statistic $T(X_1, \ldots, X_n) = (X_1, \ldots, X_{n-1})$ sufficient?

11. Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with law $\mathcal{L}(X)$. Find a minimal sufficient statistic in each of the following cases:
    (a) $X \sim P(\lambda)$.
    (b) $X \sim U[0, \theta]$.
    (c) $X \sim NB(1; p)$.
    (d) $X \sim P_N$, where $P_N\{X = k\} = 1/N$ if $k = 1, 2, \ldots, N$, and $= 0$ otherwise.
    (e) $X \sim \mathcal{N}(\mu, \sigma^2)$.
    (f) $X \sim G(\alpha, \beta)$.
    (g) $X \sim B(\alpha, \beta)$.
    (h) $X \sim f_\theta(x)$, where $f_\theta(x) = (2/\theta^2)(\theta - x)$, $0 < x < \theta$.

12. Let X_1, X_2 be a sample of size 2 from $P(\lambda)$. Show that the statistic $X_1 + \alpha X_2$, where $\alpha > 1$ is an integer, is not sufficient for λ.

13. Let $X_1, X_2, \ldots, X_n$ be a sample from the PDF

$$f_\theta(x) = \begin{cases} \dfrac{x}{\theta} e^{-x^2/2\theta} & \text{if } x > 0 \\ 0 & \text{if } x \le 0 \end{cases} \qquad \theta > 0.$$

Show that $\sum_{i=1}^n X_i^2$ is a minimal sufficient statistic for θ, but $\sum_{i=1}^n X_i$ is not sufficient.

14. Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(0, \sigma^2)$. Show that $\sum_{i=1}^n X_i^2$ is a minimal sufficient statistic but $\sum_{i=1}^n X_i$ is not sufficient for σ^2.

15. Let $X_1, X_2, \ldots, X_n$ be a sample from the PDF $f_{\alpha,\beta}(x) = \beta e^{-\beta(x-\alpha)}$ if $x > \alpha$, and $= 0$ if $x \le \alpha$. Find a minimal sufficient statistic for (α, β).

16. Let T be a minimal sufficient statistic. Show that a necessary condition for a sufficient statistic U to be complete is that U be minimal.

17. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$. Show that $(\bar{X}, S^2)$ is independent of each of $(X_{(n)} - X_{(1)})/S$, $(X_{(n)} - \bar{X})/S$, and $\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2/S^2$.

18. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\theta, 1)$. Show that a necessary and sufficient condition for $\sum_{i=1}^{n} a_i X_i$ and $\sum_{i=1}^{n} X_i$ to be independent is $\sum_{i=1}^{n} a_i = 0$.

19. Let $X_1, X_2, \ldots, X_n$ be a random sample from $f_\theta(x) = \exp[-(x - \theta)]$, $x > \theta$. Show that $X_{(1)}$ is a complete sufficient statistic which is independent of S^2.

20. Let $X_1, X_2, \ldots, X_n$ be iid RVs with common PDF $f_\theta(x) = (1/\theta)\exp(-x/\theta)$, $x > 0$, $\theta > 0$. Show that $\overline{X}$ must be independent of every scale-invariant statistic, such as $X_1/\sum_{j=1}^{n} X_j$.

21. Let T_1, T_2 be two statistics with common domain D. Then T_1 is a function of T_2 if and only if

$$\text{for all } x, y \in D, \qquad T_1(x) = T_1(y) \Longrightarrow T_2(x) = T_2(y).$$

22. Let S be the support of f_θ, $\theta \in \Theta$, and let T be a statistic such that for some $\theta_1, \theta_2 \in \Theta$, and $x, y \in S$, $x \neq y$, $T(x) = T(y)$ but $f_{\theta_1}(x) f_{\theta_2}(y) \neq f_{\theta_2}(x) f_{\theta_1}(y)$. Then show that T is not sufficient for θ.

23. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\theta, 1)$. Use the result in Problem 22 to show that $\left(\sum_1^n X_i\right)^2$ is not sufficient for θ.

24. (a) If T is complete, show that any one-to-one mapping of T is also complete.

    (b) Show with the help of an example that a complete statistic is not unique for a family of distributions.

8.4 UNBIASED ESTIMATION

In this section we focus attention on the class of unbiased estimators. We develop a criterion to check if an unbiased estimator is optimal in this class. Using sufficiency and completeness, we describe a method of constructing uniformly minimum variance unbiased estimators.

Definition 1. Let $\{F_{\boldsymbol{\theta}}, \ \boldsymbol{\theta} \in \Theta\}$, $\Theta \subseteq \mathcal{R}_k$, be a nonempty set of probability distributions. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a multiple RV with DF $F_{\boldsymbol{\theta}}$ and sample space $\mathcal{X}$. Let $\psi : \Theta \to \mathcal{R}$ be a real-valued parametric function. A Borel-measurable function $T : \mathcal{X} \to \Theta$ is said to be *unbiased* for ψ if

$$E_{\boldsymbol{\theta}} T(\mathbf{X}) = \psi(\boldsymbol{\theta}) \qquad \text{for all } \boldsymbol{\theta} \in \Theta. \tag{1}$$

Any parametric function ψ for which there exists a T satisfying (1) is called an *estimable function*. An estimator that is not unbiased is called *biased*, and the function $b(T, \psi)$, defined by

$$(2) \qquad b(T, \psi) = E_{\boldsymbol{\theta}} T(\mathbf{X}) - \psi(\boldsymbol{\theta}),$$

is called the *bias* of T.

Remark 1. Definition 1, in particular, requires that $E_{\boldsymbol{\theta}}|T| < \infty$ for all $\boldsymbol{\theta} \in \Theta$ and can be extended to the case when both ψ and T are multidimensional. In most applications we consider $\Theta \subseteq \mathcal{R}_1$, $\psi(\theta) = \theta$, and $X_1, X_2, \ldots, X_n$ are iid RVs.

Example 1. Let $X_1, X_2, \ldots, X_n$ be a random sample from some population with finite mean. Then $\overline{X}$ is unbiased for the population mean. If the population variance is finite, the sample variance S^2 is unbiased for the population variance. In general, if the kth population moment m_k exists, the kth sample moment is unbiased for m_k.

Note that S is not, in general, unbiased for σ. If $X_1, X_2, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ RVs we know that $(n-1)S^2/\sigma^2$ is $\chi^2(n-1)$. Therefore,

$$\begin{aligned} E(S\sqrt{n-1}/\sigma) &= \int_0^\infty \sqrt{x}\, \frac{1}{2^{(n-1)/2}\Gamma[(n-1)/2]} x^{(n-1)/2-1} e^{-x/2}\, dx \\ &= \sqrt{2}\,\Gamma\left(\frac{n}{2}\right)\left[\Gamma\left(\frac{n-1}{2}\right)\right]^{-1} \end{aligned}$$

and

$$E_\sigma(S) = \sigma \left\{ \sqrt{\frac{2}{n-1}}\, \Gamma\left(\frac{n}{2}\right)\left[\Gamma\left(\frac{n-1}{2}\right)\right]^{-1} \right\}.$$

The bias of S is given by

$$b(S, \sigma) = \sigma \left\{ \sqrt{\frac{2}{n-1}}\, \Gamma\left(\frac{n}{2}\right)\left[\Gamma\left(\frac{n-1}{2}\right)\right]^{-1} - 1 \right\}.$$

We note that $b(s, \sigma) \to 0$ as $n \to \infty$, so that S is asymptotically unbiased for σ.

If T is unbiased for θ, $g(T)$ is not, in general, an unbiased estimator of $g(\theta)$ unless g is a linear function.

Example 2. Unbiased estimators do not always exist. Consider an RV with PMF $b(1, p)$. Suppose that we wish to estimate $\psi(p) = p^2$. Then, in order that T be unbiased for p^2, we must have

$$p^2 = E_p T = pT(1) + (1-p)T(0), \qquad 0 \le p \le 1;$$

that is,

$$p^2 = p\{T(1) - T(0)\} + T(0)$$

must hold for all p in the interval $[0, 1]$, which is impossible. (If a convergent power series vanishes in an open interval, each of the coefficients must be 0. See also Problem 1.)

Example 3. Sometimes an unbiased estimator may be absurd. Let X be $P(\lambda)$ and $\psi(\lambda) = e^{-3\lambda}$. We show that $T(X) = (-2)^X$ is unbiased for $\psi(\lambda)$. We have

$$E_\lambda T(X) = e^{-\lambda} \sum_{x=0}^{\infty} (-2)^x \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(-2\lambda)^x}{x!} = e^{-\lambda} e^{-2\lambda} = \psi(\lambda).$$

However, $T(x) = (-2)^x > 0$ if x is even and < 0 if x is odd, which is absurd since $\psi(\lambda) > 0$.

Example 4. Let $X_1, X_2, \ldots, X_n$ be a sample from $P(\lambda)$. Then $\overline{X}$ is unbiased for λ and so also is S^2, since both the mean and the variance are equal to λ. Indeed, $\alpha\overline{X} + (1 - \alpha)S^2, 0 \le \alpha \le 1$, is unbiased for λ.

Let θ be estimable, and let T be an unbiased estimator of θ. Let T_1 be another unbiased estimator of θ, different from T. This means that there exists at least one θ such that $P_\theta\{T \neq T_1\} > 0$. In this case there exist infinitely many unbiased estimators of θ of the form $\alpha T + (1 - \alpha)T_1, 0 < \alpha < 1$. It is therefore desirable to find a procedure to differentiate among these estimators.

Definition 2. Let $\theta_0 \in \Theta$ and $\mathcal{U}(\theta_0)$ be the class of all unbiased estimators T of θ_0 such that $E_{\theta_0}T^2 < \infty$. Then $T_0 \in \mathcal{U}(\theta_0)$ is called a *locally minimum variance unbiased estimator* (LMVUE) at θ_0 if

$$(3) \qquad E_{\theta_0}(T_0 - \theta_0)^2 \le E_{\theta_0}(T - \theta_0)^2$$

holds for all $T \in \mathcal{U}(\theta_0)$.

Definition 3. Let $\mathcal{U}$ be the set of all unbiased estimators T of $\theta \in \Theta$ such that $E_\theta T^2 < \infty$ for all $\theta \in \Theta$. An estimator $T_0 \in \mathcal{U}$ is called a *uniformly minimum variance unbiased estimator* (UMVUE) of θ if

$$(4) \qquad E_\theta(T_0 - \theta)^2 \le E_\theta(T - \theta)^2$$

for all $\theta \in \Theta$ and every $T \in \mathcal{U}$.

Remark 2. Let $a_1, a_2, \ldots, a_n$ be any set of real numbers with $\sum_{i=1}^n a_i = 1$. Let $X_1, X_2, \ldots, X_n$ be independent RVs with common mean μ and variances σ_k^2, $k = 1, 2, \ldots, n$. Then $T = \sum_{i=1}^n a_i X_i$ is an unbiased estimator of μ with variance $\sum_{i=1}^n a_i^2 \sigma_i^2$ (see Theorem 4.5.6). T is called a *linear unbiased estimator* of μ. Linear unbiased estimators of μ that have minimum variance (among all linear unbiased estimators) are called *best linear unbiased estimators* (BLUEs). In Theorem 4.5.6

(Corollary 2) we have shown that if X_i are iid RVs with common variance σ^2, the BLUE of μ is $\overline{X} = n^{-1}\sum_{i=1}^n X_i$. If X_i are independent with common mean μ but different variance σ_i^2, the BLUE of μ is obtained if we choose a_i proportional to $1/\sigma_i^2$; then the minimum variance is H/n, where H is the harmonic mean of $\sigma_1^2, \dots, \sigma_n^2$ (see Example 4.5.4).

Remark 3. Sometimes the precision of an estimator T of parameter θ is measured by the *mean square error* (MSE). We say that an estimator T_0 is at least as good as any other estimator T in the sense of the MSE if

$$(5)\qquad E_\theta(T_0-\theta)^2 \le E_\theta(T-\theta)^2 \qquad \text{for all } \theta \in \Theta.$$

In general, a particular estimator will be better than another for some values of θ and worse for others. Definitions 2 and 3 are special cases of this concept if we restrict attention to unbiased estimators.

The following result gives a necessary and sufficient condition for an unbiased estimator to be a UMVUE.

Theorem 1. Let $\mathcal{U}$ be the class of all unbiased estimators T of a parameter $\theta \in \Theta$ with $E_\theta T^2 < \infty$ for all θ, and suppose that $\mathcal{U}$ is nonempty. Let $\mathcal{U}_0$ be the class of all unbiased estimators v of 0, that is,

$$\mathcal{U}_0 = \{v : E_\theta v = 0,\ E_\theta v^2 < \infty \text{ for all } \theta \in \Theta\}.$$

Then $T_0 \in \mathcal{U}$ is a UMVUE if and only if

$$(6)\qquad E_\theta(vT_0) = 0 \qquad \text{for all } \theta \text{ and all } v \in \mathcal{U}_0.$$

Proof. The conditions of the theorem guarantee the existence of $E_\theta(vT_0)$ for all θ and $v \in \mathcal{U}_0$. Suppose that $T_0 \in \mathcal{U}$ is a UMVUE and $E_{\theta_0}(v_0T_0) \ne 0$ for some θ_0 and some $v_0 \in \mathcal{U}_0$. Then $T_0 + \lambda v_0 \in \mathcal{U}$ for all real λ. If $E_{\theta_0}v_0^2 = 0$, then $E_{\theta_0}(v_0T_0) = 0$ must hold since $P_{\theta_0}\{v_0 = 0\} = 1$. Let $E_{\theta_0}v_0^2 > 0$. Choose $\lambda_0 = -E_{\theta_0}(T_0v_0)/E_{\theta_0}v_0^2$. Then

$$(7)\qquad E_{\theta_0}(T_0+\lambda_0v_0)^2 = E_{\theta_0}T_0^2 - \frac{E_{\theta_0}^2(v_0T_0)}{E_{\theta_0}v_0^2} < E_{\theta_0}T_0^2.$$

Since $T_0 + \lambda_0v_0 \in \mathcal{U}$ and $T_0 \in \mathcal{U}$, it follows from (7) that

$$(8)\qquad \operatorname{var}_{\theta_0}(T_0+\lambda_0v_0) < \operatorname{var}_{\theta_0}(T_0),$$

which is a contradiction. It follows that (6) holds.

Conversely, let (6) hold for some $T_0 \in \mathcal{U}$, all $\theta \in \Theta$ and all $v \in \mathcal{U}_0$, and let $T \in \mathcal{U}$. Then $T_0 - T \in \mathcal{U}_0$, and for every θ,

$$E_\theta\{T_0(T_0 - T)\} = 0.$$

We have

$$E_\theta T_0^2 = E_\theta(TT_0) \leq (E_\theta T_0^2)^{1/2}(E_\theta T^2)^{1/2}$$

by the Cauchy–Schwarz inequality. If $E_\theta T_0^2 = 0$, then $P(T_0 = 0) = 1$ and there is nothing to prove. Otherwise,

$$(E_\theta T_0^2)^{1/2} \leq (E_\theta T^2)^{1/2}$$

or $\text{var}_\theta(T_0) \leq \text{var}_\theta(T)$. Since T is arbitrary, the proof is complete.

Theorem 2. Let $\mathcal{U}$ be the nonempty class of unbiased estimators as defined in Theorem 1. Then there exists at most one UMVUE for θ.

Proof. If T and $T_0 \in \mathcal{U}$ are both UMVUEs, then $T - T_0 \in U_0$ and

$$E_\theta\{T_0(T - T_0)\} = 0 \qquad \text{for all } \theta \in \Theta,$$

that is, $E_\theta T_0^2 = E_\theta(TT_0)$, and it follows that

$$\text{cov}(T, T_0) = \text{var}_\theta(T_0) \qquad \text{for all } \theta.$$

Since T_0 and T are both UMVUEs, $\text{var}_\theta(T) = \text{var}_\theta(T_0)$, and it follows that the correlation coefficient between T and T_0 is 1. This implies that $P_\theta\{aT + bT_0 = 0\} = 1$ for some a, b and all $\theta \in \Theta$. Since T and T_0 are both unbiased for θ, we must have $P_\theta\{T = T_0\} = 1$ for all θ.

Remark 4. Both Theorems 1 and 2 have analogs for LMVUE's at $\theta_0 \in \Theta$, θ_0 fixed.

Theorem 3. If UMVUEs T_i exist for real functions ψ_i, $i = 1, 2$, of θ, they also exist for $\lambda\psi_i$ (λ real), as well as for $\psi_1 + \psi_2$, and are given by λT_i and $T_1 + T_2$, respectively.

Theorem 4. Let $\{T_n\}$ be a sequence of UMVUEs and T be a statistic with $E_\theta T^2 < \infty$ and such that $E_\theta\{T_n - T\}^2 \to 0$ as $n \to \infty$ for all $\theta \in \Theta$. Then T is also the UMVUE.

Proof. That T is unbiased follows from $|E_\theta T - \theta| \leq E_\theta|T - T_n| \leq E_0^{1/2}(T_n - T)^2$. For all $v \in \mathcal{U}_0$, all θ, and every $n = 1, 2, \ldots,$

$$E_\theta(T_n v) = 0$$

by Theorem 1. Therefore,

$$\begin{aligned}E_\theta(vT) &= E_\theta(vT) - E_\theta(vT_n)\\ &= E_\theta[v(T - T_n)]\end{aligned}$$

and

$$|E_\theta(vT)| \le (E_\theta v^2)^{1/2}[E_\theta(T - T_n)^2]^{1/2} \to 0 \qquad \text{as } n \to \infty$$

for all θ and all $v \in \mathcal{U}$. Thus

$$E_\theta(vT) = 0 \qquad \text{for all } v \in \mathcal{U}_0, \quad \text{all } \theta \in \Theta,$$

and by Theorem 1, T must be the UMVUE.

Example 5. Let $X_1, X_2, \dots, X_n$ be iid $P(\lambda)$. Then $\overline{X}$ is the UMVUE of λ. Surely, $\overline{X}$ is unbiased. Let g be an unbiased estimator of 0. Then $T(\mathbf{X}) = \overline{X} + g(\overline{X})$ is unbiased for θ. But $\overline{X}$ is complete. It follows that

$$E_\lambda g(\overline{X}) = 0 \qquad \text{for all } \lambda > 0 \Rightarrow g(x) = 0 \quad \text{for } x = 0, 1, 2, \dots .$$

Hence $\overline{X}$ must be the UMVUE of λ.

Example 6. Sometimes an estimator with larger variance may be preferable.

Let X be a $G(1, 1/\beta)$ RV. X is usually taken as a good model to describe the time to failure of a piece of equipment. Let $X_1, X_2, \dots, X_n$ be a sample of n observations on X. Then $\overline{X}$ is unbiased for $EX = 1/\beta$ with variance $1/(n\beta^2)$. ($\overline{X}$ is actually the UMVUE for $1/\beta$.) Now consider $X_{(1)} = \min(X_1, X_2, \dots, X_n)$. Then $nX_{(1)}$ is unbiased for $1/\beta$ with variance $1/\beta^2$, and it has a larger variance than $\overline{X}$. However, if the length of time is of importance, $nX_{(1)}$ may be preferable to $\overline{X}$, since to observe $nX_{(1)}$ one needs to wait only until the first piece of equipment fails, whereas to compute $\overline{X}$ one would have to wait until all the n observations $X_1, X_2, \dots, X_n$ are available.

Theorem 5. If a sample consists of n independent observations $X_1, X_2, \dots, X_n$ from the same distribution, the UMVUE, if it exists, is a symmetric function of the X_i's.

The proof is left as an exercise.

The converse of Theorem 5 is not true. If $X_1, X_2, \dots, X_n$ are iid $P(\lambda)$ RVs, $\lambda > 0$, both $\overline{X}$ and S^2 are unbiased for θ. But $\overline{X}$ is the UMVUE, whereas S^2 is not.

We now turn our attention to some methods for finding UMVUEs.

Theorem 6 (Blackwell [9], Rao [85]). Let $\{F_\theta : \theta \in \Theta\}$ be a family of probability DFs and h be any statistic in $\mathcal{U}$, where $\mathcal{U}$ is the (nonempty) class of all unbiased estimators of θ with $E_\theta h^2 < \infty$. Let T be a sufficient statistic for $\{F_\theta, \theta \in \Theta\}$. Then the conditional expectation $E_\theta\{h \mid T\}$ is independent of θ and is an unbiased estimator of θ. Moreover,

$$E_\theta(E\{h \mid T\} - \theta)^2 \le E_\theta(h - \theta)^2 \qquad \text{for all } \theta \in \Theta. \tag{9}$$

The equality in (9) holds if and only if $h = E\{h \mid T\}$ (that is, $P_\theta\{h = E\{h \mid T\}\} = 1$ for all θ).

Proof. We have

$$E_\theta\{E\{h \mid T\}\} = E_\theta h = \theta.$$

It is therefore sufficient to show that

$$E_\theta\{E\{h \mid T\}\}^2 \le E_\theta h^2 \qquad \text{for all } \theta \in \Theta. \tag{10}$$

But $E_\theta h^2 = E_\theta\{E\{h^2 \mid T\}\}$, so that it will be sufficient to show that

$$[E\{h \mid T\}]^2 \le E\{h^2 \mid T\}. \tag{11}$$

By the Cauchy–Schwarz inequality

$$E^2\{h \mid T\} \le E\{h^2 \mid T\}E\{1 \mid T\},$$

and (11) follows. The equality holds in (9) if and only if

$$E_\theta[E\{h \mid T\}]^2 = E_\theta h^2, \tag{12}$$

that is,

$$E_\theta[E\{h^2 \mid T\} - E^2\{h \mid T\}] = 0,$$

which is the same as

$$E_\theta\{\text{var}\{h \mid T\}\} = 0.$$

This happens if and only if $\text{var}\{h \mid T\} = 0$, that is, if and only if

$$E\{h^2 \mid T\} = E^2\{h \mid T\},$$

as will be the case if and only if h is a function of T. Thus $h = E\{h \mid T\}$ with probability 1.

Theorem 6 is applied along with completeness to yield the following result.

Theorem 7 (Lehmann–Scheffé [62]). If T is a complete sufficient statistic and there exists an unbiased estimator h of θ, there exists a unique UMVUE of θ, which is given by $E\{h \mid T\}$.

Proof. If $h_1, h_2 \in \mathcal{U}$, then $E\{h_1 \mid T\}$ and $E\{h_2 \mid T\}$ are both unbiased and

$$E_\theta[E\{h_1 \mid T\} - E\{h_2 \mid T\}] = 0 \qquad \text{for all } \theta \in \Theta.$$

Since T is a complete sufficient statistic, it follows that $E\{h_1 \mid T\} = E\{h_2 \mid T\}$. By Theorem 6 $E\{h \mid T\}$ is the UMVUE.

Remark 5. According to Theorem 6, we should restrict our search to Borel-measurable functions of a sufficient statistic (whenever it exists). According to Theorem 7, if a complete sufficient statistic T exists, all we need to do is to find a Borel-measurable function of T that is unbiased. If a complete sufficient statistic does not exist, an UMVUE may still exist (see Example 11).

***Example** 7.* Let $X_1, X_2, \dots, X_n$ be $\mathcal{N}(\theta, 1)$. X_1 is unbiased for θ. However, $\overline{X} = n^{-1} \sum_1^n X_i$ is a complete sufficient statistic, so that $E\{X_1 \mid \overline{X}\}$ is the UMVUE.

We will show that $E\{X_1 \mid \overline{X}\} = \overline{X}$. Let $Y = n\overline{X}$. Then Y is $\mathcal{N}(n\theta, n)$, X_1 is $\mathcal{N}(\theta, 1)$, and (X_1, Y) is a bivariate normal RV with variance covariance matrix $\begin{pmatrix} 1 & 1 \\ 1 & n \end{pmatrix}$. Therefore,

$$\begin{aligned} E\{X_1 \mid y\} &= EX_1 + \frac{\operatorname{cov}(X_1, Y)}{\operatorname{var}(Y)}(y - EY) \\ &= \theta + \frac{1}{n}(y - n\theta) = \frac{y}{n}, \end{aligned}$$

as asserted.

If we let $\psi(\theta) = \theta^2$, we can show similarly that $\overline{X}^2 - 1/n$ is the UMVUE for $\psi(\theta)$. Note that $\overline{X}^2 - 1/n$ may occasionally be negative, so that an UMVUE for θ^2 is not very sensible in this case.

***Example** 8.* Let $X_1, X_2, \dots, X_n$ be iid $b(1, p)$ RVs. Then $T = \sum_1^n X_i$ is a complete sufficient statistic. The UMVUE for p is clearly $\overline{X}$. To find the UMVUE for $\psi(p) = p(1 - p)$, we have $E(nT) = n^2 p$, $ET^2 = np + n(n-1)p^2$, so that $E\{nT - T^2\} = n(n-1)p(1-p)$, and it follows that $(nT - T^2)/n(n-1)$ is the UMVUE for $\psi(p) = p(1-p)$.

***Example** 9.* Let $X_1, X_2, \dots, X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$. Then $(\overline{X}, S^2)$ is a complete sufficient statistic for (μ, σ^2). $\overline{X}$ is the UMVUE for μ, and S^2 is the UMVUE for σ^2. Also, $k(n)S$ is the UMVUE for σ, where $k(n) = \sqrt{(n-1)/2}\,\Gamma[(n-1)/2]/\Gamma(n/2)$. We wish to find the UMVUE for the pth quantile $\mathfrak{z}_p$. We have

$$p = P\{X \le \mathfrak{z}_p\} = P\left\{Z \le \frac{\mathfrak{z}_p - \mu}{\sigma}\right\},$$

where Z is $\mathcal{N}(0, 1)$. Thus $\mathfrak{z}_p = \sigma z_{1-p} + \mu$, and the UMVUE is

$$T(X_1, X_2, \dots, X_n) = z_{1-p}k(n)S + \overline{X}.$$

Example 10 (Stigler [109]). We return to Example 8.3.14. We have seen that the family $\{P_N^{(n)}: N \geq 1\}$ of PMFs of $X_{(n)} = \max_{1 \leq i \leq n} X_i$ is complete and $X_{(n)}$ is sufficient for N. Now $EX_1 = (N+1)/2$, so that $T(X_1) = 2X_1 - 1$ is unbiased for N. It follows from Theorem 7 that $E\{T(X_1) \mid X_{(n)}\}$ is the UMVUE of N. We have

$$P\{X_1 = x_1 \mid X_{(n)} = y\} = \begin{cases} \dfrac{y^{n-1} - (y-1)^{n-1}}{y^n - (y-1)^n} & \text{if } x_1 = 1, 2, \dots, y-1, \\[2ex] \dfrac{y^{n-1}}{y^n - (y-1)^n} & \text{if } x_1 = y. \end{cases}$$

Thus

$$\begin{aligned} E\{T(X_1) \mid X_{(n)} = y\} &= \frac{y^{n-1} - (y-1)^{n-1}}{y^n - (y-1)^n} \sum_{x_1=1}^{y-1} (2x_1 - 1) \\ &\quad + (2y-1)\frac{y^{n-1}}{y^n - (y-1)^n} \\ &= \frac{y^{n+1} - (y-1)^{n+1}}{y^n - (y-1)^n} \end{aligned}$$

is the UMVUE of N.

If we consider the family $\mathcal{P}$ instead, we have seen (Example 8.3.14 and Problem 8.3.6) that $\mathcal{P}$ is not complete. The UMVUE for the family $\{P_N: N \geq 1\}$ is $T(X_1) = 2X_1 - 1$, which is not the UMVUE for $\mathcal{P}$. The UMVUE for $\mathcal{P}$ is, in fact, given by

$$T_1(k) = \begin{cases} 2k - 1, & k \neq n_0, \quad k \neq n_0 + 1, \\ 2n_0, & k = n_0, \quad k = n_0 + 1. \end{cases}$$

The reader is asked to check that T_1 has covariance 0 with all unbiased estimators g of 0 that are of the form described in Example 8.3.14 and Problem 8.3.6, and hence Theorem 1 implies that T_1 is the UMVUE. Actually, $T_1(X_1)$ is a complete sufficient statistic for $\mathcal{P}$. Since $E_{n_0}T_1(X_1) = n_0 + 1/n_0$, T_1 is not even unbiased for the family $\{P_N: N \geq 1\}$. The minimum variance is given by

$$\operatorname{var}_N(T_1(X_1)) = \begin{cases} \operatorname{var}_N(T(X_1)) & \text{if } N < n_0, \\[1ex] \operatorname{var}_N(T(X_1)) - \dfrac{2}{N} & \text{if } N > n_0. \end{cases}$$

The following example shows that a UMVUE may exist, whereas a minimal sufficient statistic may not.

Example 11. Let X be an RV with PMF

$$P_\theta(X=-1)=\theta \quad \text{and} \quad P_\theta(X=x)=(1-\theta)^2\theta^x,$$

$x=0,1,2,\dots$, where $0<\theta<1$. Let $\psi(\theta)=P_\theta(X=0)=(1-\theta)^2$. Then X is clearly sufficient, in fact minimal sufficient, for θ but since

$$\begin{aligned} E_\theta X &= (-1)\theta+\sum_{x=0}^{\infty} x(1-\theta)^2\theta^x \\ &= -\theta+\theta(1-\theta)^2\frac{d}{d\theta}\sum_{x=1}^{\infty}\theta^x=0, \end{aligned}$$

it follows that X is not complete for $\{P_\theta : 0<\theta<1\}$. We will use Theorem 1 to check if a UMVUE for $\psi(\theta)$ exists. Suppose that

$$E_\theta h(X)=h(-1)\theta+\sum_{x=0}^{\infty}(1-\theta)^2\theta^x h(x)=0$$

for all $0<\theta<1$. Then, for $0<\theta<1$,

$$\begin{aligned} 0 &= \theta h(-1)+\sum_{x=0}^{\infty}\theta^x h(x)-2\sum_{x=0}^{\infty}\theta^{x+1}h(x)+\sum_{x=0}^{\infty}\theta^{x+2}h(x) \\ &= h(0)+\sum_{x=0}^{\infty}\theta^{x+1}[h(x+1)-2h(x)+h(x-1)] \end{aligned}$$

which is a power series in θ.

It follows that $h(0)=0$, and for $x\ge 1$, $h(x+1)-2h(x)+h(x-1)=0$. Thus

$$\begin{aligned} &h(1)=h(-1),\ h(2)=2h(1)-h(0)=2h(-1), \\ &h(3)=2h(2)-h(1)=4h(-1)-h(-1)=3h(-1), \end{aligned}$$

and so on. Consequently, all unbiased estimators of zero are of the form $h(X)=cX$. Clearly, $T(X)=1$ if $X=0$, and $=0$ otherwise is unbiased for $\psi(\theta)$. Moreover, for all θ,

$$E\{cX\cdot T(X)\}=0,$$

so that T is UMVUE of $\psi(\theta)$.

We conclude this section with a proof of Theorem 8.3.4.

Theorem 8. (Theorem 8.3.4) A complete sufficient statistic is a minimal sufficient statistic.

Proof. Let $S(\mathbf{X})$ be a complete sufficient statistic for $\{f_\theta : \theta \in \Theta\}$ and let T be any statistic for which $E_\theta|T^2| < \infty$. Writing $h(S) = E_\theta\{T|S\}$, we see that h is the UMVUE of $E_\theta T$. Let $S_1(\mathbf{X})$ be another sufficient statistic. We show that $h(S)$ is a function of S_1. If not, then $h_1(S_1) = E_\theta\{h(S)|S_1\}$ is unbiased for $E_\theta T$ and by the Rao–Blackwell theorem,

$$\operatorname{var}_\theta h_1(S_1) \le \operatorname{var}_\theta h(S),$$

contradicting the fact that $h(S)$ is UMVUE for $E_\theta T$. It follows that $h(S)$ is a function of S_1. Since h and S_1 are arbitrary, S must be a function of every sufficient statistic and hence, minimal sufficient.

PROBLEMS 8.4

1. Let $X_1, X_2, \dots, X_n (n \ge 2)$ be a sample from $b(1, p)$. Find an unbiased estimator for $\psi(p) = p^2$.
2. Let $X_1, X_2, \dots, X_n (n \ge 2)$ be a sample from $\mathcal{N}(\mu, \sigma^2)$. Find an unbiased estimator for σ^p, where $p + n > 1$. Find a minimum MSE estimator of σ^p.
3. Let $X_1, X_2, \dots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$ RVs. Find a minimum MSE estimator of the form αS^2 for the parameter σ^2. Compare the variances of the minimum MSE estimator and the obvious estimator S^2.
4. Let $X \sim b(1, \theta^2)$. Does there exist an unbiased estimator of θ?
5. Let $X \sim P(\lambda)$. Does there exist an unbiased estimator of $\psi(\lambda) = \lambda^{-1}$?
6. Let $X_1, X_2, \dots, X_n$ be a sample from $b(1, p), 0 < p < 1$, and $0 < s < n$ be an integer. Find the UMVUE for (a) $\psi(p) = p^s$, and (b) $\psi(p) = p^s + (1-p)^{n-s}$.
7. Let $X_1, X_2, \dots, X_n$ be a sample from a population with mean θ and finite variance, and T be an estimator of θ of the form $T(X_1, X_2, \dots, X_n) = \sum_{i=1}^n \alpha_i X_i$. If T is an unbiased estimator of θ that has minimum variance and T' is another linear unbiased estimator of θ, then
$$\operatorname{cov}_\theta(T, T') = \operatorname{var}_\theta(T).$$
8. Let T_1, T_2 be two unbiased estimators having common variance $\alpha\sigma^2 (\alpha > 1)$, where σ^2 is the variance of the UMVUE. Show that the correlation coefficient between T_1 and T_2 is $\ge (2-\alpha)/\alpha$.
9. Let $X \sim NB(1; \theta)$ and $d(\theta) = P_\theta\{X = 0\}$. Let $X_1, X_2, \dots, X_n$ be a sample on X. Find the UMVUE of $d(\theta)$.

10. This example covers most discrete distributions. Let $X_1, X_2, \dots, X_n$ be a sample from the PMF

$$P_\theta\{X = x\} = \frac{\alpha(x)\theta^x}{f(\theta)}, \qquad x = 0, 1, 2, \dots,$$

where $\theta > 0$, $\alpha(x) > 0$, $f(\theta) = \sum_{x=0}^{\infty} \alpha(x)\theta^x$, $\alpha(0) = 1$, and let $T = X_1 + X_2 + \cdots + X_n$. Write

$$c(t, n) = \sum_{x_1, x_2, \dots, x_n} \prod_{i=1}^{n} \alpha(x_i)$$

$$\text{with} \quad \sum_{i=1}^{n} x_i = t.$$

Show that T is a complete sufficient statistic for θ and that the UMVUE for $d(\theta) = \theta^r$ ($r > 0$ is an integer) is given by

$$Y_r(t) = \begin{cases} 0 & \text{if } t < r. \\ \dfrac{c(t-r, n)}{c(t, n)} & \text{if } t \ge r. \end{cases}$$

(Roy and Mitra [92])

11. Let X be a hypergeometric RV with PMF

$$P_M\{X = x\} = \binom{N}{n}^{-1} \binom{M}{x} \binom{N-M}{n-x},$$

where $\max(0, M + n - N) \le x \le \min(M, n)$.

(a) Find the UMVUE for M when N is assumed to be known.

(b) Does there exist an unbiased estimator of N (M known)?

12. Let $X_1, X_2, \dots, X_n$ be iid $G(1, 1/\lambda)$ RVs $\lambda > 0$. Find the UMVUE of $P_\lambda\{X_1 \le t_0\}$, where $t_0 > 0$ is a fixed real number.

13. Let $X_1, X_2, \dots, X_n$ be a random sample from $P(\lambda)$. Let $\psi(\lambda) = \sum_{k=0}^{\infty} c_k \lambda^k$ be a parametric function. Find the UMVUE for $\psi(\lambda)$. In particular, find the UMVUE for (a) $\psi(\lambda) = 1/(1 - \lambda)$, (b) $\psi(\lambda) = \lambda^s$ for some fixed integer $s > 0$, (c) $\psi(\lambda) = P_\lambda\{X = 0\}$, and (d) $\psi(\lambda) = P_\lambda\{X = 0 \text{ or } 1\}$.

14. Let $X_1, X_2, \dots, X_n$ be a sample from the PMF

$$P_N(x) = \frac{1}{N}, \qquad x = 1, 2, \dots, N.$$

Let $\psi(N)$ be a function of N. Find the UMVUE of $\psi(N)$.

15. Let $X_1, X_2, \ldots, X_n$ be a random sample from $P(\lambda)$. Find the UMVUE of $\psi(\lambda) = P_\lambda\{X = k\}$, where k is a fixed positive integer.

16. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a sample from a bivariate normal population with parameters μ_1, μ_2, σ_1^2, σ_2^2, and ρ. Assume that $\mu_1 = \mu_2 = \mu$, and it is required to find an unbiased estimator of μ. Since a complete sufficient statistic does not exist, consider the class of all linear unbiased estimators

$$\hat{\mu}(\alpha) = \alpha\overline{X} + (1 - \alpha)\overline{Y}.$$

(a) Find the variance of $\hat{\mu}$.

(b) Choose $\alpha = \alpha_0$ to minimize $\text{var}(\hat{\mu})$, and consider the estimator

$$\hat{\mu}_0 = \alpha_0\overline{X} + (1 - \alpha_0)\overline{Y}.$$

Compute $\text{var}(\hat{\mu}_0)$. If $\sigma_1 = \sigma_2$, the BLUE of μ (in the sense of minimum variance) is

$$\hat{\mu}_1 = \frac{\overline{X} + \overline{Y}}{2}$$

irrespective of whether σ_1 and ρ are known or unknown.

(c) If $\sigma_1 \neq \sigma_2$ and ρ, σ_1, σ_2 are unknown, replace these values in α_0 by their corresponding estimators. Let

$$\hat{\alpha} = \frac{S_2^2 - S_{11}}{S_1^2 + S_2^2 - 2S_{11}}.$$

Show that

$$\hat{\mu}_2 = \overline{Y} + (\overline{X} - \overline{Y})\hat{\alpha}$$

is an unbiased estimator of μ.

17. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\theta, 1)$. Let $p = \Phi(x - \theta)$, where Φ is the DF of a $\mathcal{N}(0, 1)$ RV. Show that the UMVUE of p is given by $\Phi\left((x - \overline{x})\sqrt{n/(n-1)}\right)$.

18. Prove Theorem 5.

19. In Example 10 show that T_1 is the UMVUE for N (restricted to the family $\mathcal{P}$), and compute the minimum variance.

20. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample from a bivariate population with finite variances σ_1^2 and σ_2^2, respectively, and covariance γ. Show that

$$\text{var}(S_{11}) = \frac{1}{n}\left(\mu_{22} - \frac{n-2}{n-1}\gamma^2 + \frac{\sigma_1^2\sigma_2^2}{n-1}\right),$$

where $\mu_{22} = E[(X - EX)^2(Y - EY)^2]$. It is assumed that appropriate order moments exist.

21. Suppose that a random sample is taken on (X, Y) and it is desired to estimate γ, the unknown covariance between X and Y. Suppose that for some reason a set S of n observations is available on both X and Y, an additional $n_1 - n$ observations are available on X but the corresponding Y values are missing, and an additional $n_2 - n$ observations of Y are available for which the X values are missing. Let S_1 be the set of all $n_1 (\geq n)$ X values, and S_2, the set of all $n_2 (\geq n)$ Y values, and write

$$\hat{X} = \frac{\sum_{j\in S_1} X_j}{n_1}, \quad \hat{Y} = \frac{\sum_{j\in S_2} Y_j}{n_2}, \quad \overline{X} = \frac{\sum_{i\in S} X_i}{n}, \quad \overline{Y} = \frac{\sum_{i\in S} Y_i}{n}.$$

Show that

$$\hat{\gamma} = \frac{n_1 n_2}{n(n_1 n_2 - n_1 - n_2 + n)} \sum_{i\in S} (X_i - \hat{X})(Y_i - \hat{Y})$$

is an unbiased estimator of γ. Find the variance of $\hat{\gamma}$, and show that $\text{var}(\hat{\gamma}) \leq \text{var}(S_{11})$, where S_{11} is the usual unbiased estimator of γ based on the n observations in S. (Boas [10])

22. Let $X_1, X_2, \ldots, X_n$ be iid with common PDF $f_\theta(x) = \exp(-x + \theta)$, $x > \theta$. Let x_0 be a fixed real number. Find the UMVUE of $f_\theta(x_0)$.

23. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\mu, 1)$ RVs. Let $T(\mathbf{X}) = \sum_{i=1}^n X_i$. Show that $\varphi(x; t/n, n - 1/n)$ is the UMVUE of $\varphi(x; \mu, 1)$ where $\varphi(x; \mu, \sigma^2)$ is the PDF of a $\mathcal{N}(\mu, \sigma^2)$ RV.

24. Let $X_1, X_2, \ldots, X_n$ be iid $G(1, \theta)$ RVs. Show that the UMVUE of $f(x; \theta) = (1/\theta)\exp(-x/\theta)$, $x > 0$, is given by $h(x|t)$ the conditional PDF of X_1 given $T(X) = \sum_{i=1}^n X_i = t$, where

$$h(x|t) = (n-1)(t-x)^{n-2}/t^{n-1} \qquad \text{for } x < t \text{ and } = 0 \text{ for } x > t.$$

25. Let $X_1, X_2, \ldots, X_n$ be iid RVs with common PDF $f_\theta(x) = 1/(2\theta)$, $|x| < \theta$, and $= 0$ elsewhere. Show that $T(\mathbf{X}) = \max\{-X_{(1)}, X_{(n)}\}$ is a complete sufficient statistic for θ. Find the UMVUE of θ^r.

26. Let $X_1, X_2, \ldots, X_n$ be a random sample from the PDF

$$f_\theta(x) = \frac{1}{\sigma} \exp\left[-\frac{(x-\mu)}{\sigma}\right], \qquad x > \mu, \quad \sigma > 0$$

where $\theta = (\mu, \sigma)$.

(a) Show that $\left(X_{(1)}, \sum_{j=1}^n \left(X_j - X_{(1)}\right)\right)$ is a complete sufficient statistic for θ.

(b) Show that the UMVUEs of μ and σ are given by

$$\hat{\mu} = X_{(1)} - \frac{1}{n(n-1)} \sum_{j=1}^{n} \left(X_j - X_{(1)}\right), \ \hat{\sigma} = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - X_{(1)}\right).$$

(c) Find the UMVUE of $\psi(\mu, \sigma) = E_{\mu,\sigma} X_1$.

(d) Show that the UMVUE of $P_\theta\{X_1 \geq t\}$ is given by

$$\hat{P}(X_1 \geq t) = \frac{n-1}{n} \left\{ \left[1 - \frac{t - X_{(1)}}{\sum_1^n (X_j - X_{(1)})} \right]^+ \right\}^{n-2}$$

where $x^+ = \max(x, 0)$.

8.5 UNBIASED ESTIMATION (*CONTINUED*): LOWER BOUND FOR THE VARIANCE OF AN ESTIMATOR

In this section we consider two inequalities, each of which provides a lower bound for the variance of an estimator. These inequalities can sometimes be used to show that an unbiased estimator is the UMVUE. We first consider an inequality due to Fréchet, Cramér, and Rao (the FCR inequality).

Theorem 1 (Cramér [17], Fréchet [31], Rao [84]). Let $\Theta \subseteq \mathcal{R}$ be an open interval and suppose that the family $\{f_\theta : \theta \in \Theta\}$ satisfies the following regularity conditions:

(i) It has common support set S. Thus $S = \{x : f_\theta(\mathbf{x}) > 0\}$ does not depend on θ.

(ii) For $x \in S$ and $\theta \in \Theta$, the derivative $\dfrac{\partial}{\partial \theta} \log f_\theta(x)$ exists and is finite.

(iii) For any statistic h with $E_\theta|h(\mathbf{X})| < \infty$ for all θ, the operations of integration (summation) and differentiation with respect to θ can be interchanged in $E_\theta h(\mathbf{X})$. That is,

$$\frac{\partial}{\partial \theta} \int h(\mathbf{x}) f_\theta(\mathbf{x})\, d\mathbf{x} = \int h(\mathbf{x}) \frac{\partial}{\partial \theta} f_\theta(\mathbf{x})\, d\mathbf{x} \tag{1}$$

whenever the right-hand side of (1) is finite.

Let $T(\mathbf{X})$ be such that $\operatorname{var}_\theta T(\mathbf{X}) < \infty$ for all θ and set $\psi(\theta) = E_\theta T(\mathbf{X})$. If $I(\theta) = E_\theta \left[\dfrac{\partial}{\partial \theta} \log f_\theta(\mathbf{X}) \right]^2$ satisfies $0 < I(\theta) < \infty$, then

$$\operatorname{var}_\theta T(\mathbf{X}) \geq \frac{[\psi'(\theta)]^2}{I(\theta)}. \tag{2}$$

Proof. Since (iii) holds for $h \equiv 1$, we get

$$
\begin{aligned}
(3)\qquad 0 &= \int_S \frac{\partial}{\partial\theta} f_\theta(\mathbf{x})\,dx \\
&= \int_S \left[\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{x})\right] f_\theta(\mathbf{x})\,d\mathbf{x} \\
&= E\left[\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{X})\right].
\end{aligned}
$$

Differentiating $\psi(\theta) = E_\theta T(\mathbf{X})$ and using (1), we get

$$
\begin{aligned}
(4)\qquad \psi'(\theta) &= \int_S T(\mathbf{x})\frac{\partial}{\partial\theta} f_\theta(\mathbf{x})\,d\mathbf{x} \\
&= \int_S \left[T(\mathbf{x})\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{x})\right] f_\theta(\mathbf{x})\,d\mathbf{x} \\
&= \operatorname{cov}\left[T(\mathbf{X}), \frac{\partial}{\partial\theta}\log f_\theta(\mathbf{X})\right].
\end{aligned}
$$

Also, in view of (3), we have

$$\operatorname{var}_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{X})\right] = E_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{X})\right]^2.$$

and using Cauchy–Schwarz inequality in (4), we get

$$[\psi'(\theta)]^2 \le \operatorname{var}_\theta T(\mathbf{X}) E_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{X})\right]^2$$

which proves (2). Practically the same proof may be given when f_θ is a PMF by replacing $\int$ by Σ.

Remark 1. If, in particular, $\psi(\theta) = \theta$, then (2) reduces to

$$(5)\qquad \operatorname{var}_\theta(T(\mathbf{X})) \ge \frac{1}{I(\theta)}.$$

Remark 2. Let $X_1, X_2, \ldots, X_n$ be iid RVs with common PDF (PMF) $f_\theta(x)$. Then

$$
\begin{aligned}
I(\theta) &= E_\theta\left[\frac{\partial \log f_\theta(\mathbf{X})}{\partial\theta}\right]^2 = \sum_{i=1}^{n} E_\theta\left[\frac{\partial \log f_\theta(X_i)}{\partial\theta}\right]^2 \\
&= nE_\theta\left[\frac{\partial \log f_\theta(X_1)}{\partial\theta}\right]^2 = nI_1(\theta),
\end{aligned}
$$

where $I_1(\theta) = E_\theta[\partial \log f_\theta(X_1)/\partial\theta]^2$. In this case the inequality (2) reduces to

$$\text{var}_\theta(T(\mathbf{X})) \geq \frac{[\psi'(\theta)]^2}{nI_1(\theta)}.$$

Definition 1. The quantity

$$I_1(\theta) = E_\theta\left[\frac{\partial \log f_\theta(X_1)}{\partial\theta}\right]^2 \tag{6}$$

is called *Fisher information* in X_1 and

$$I_n(\theta) = E_\theta\left[\frac{\partial \log f_\theta(\mathbf{X})}{\partial\theta}\right]^2 = nI_1(\theta) \tag{7}$$

is known as Fisher information in the random sample $X_1, X_2, \dots, X_n$.

Remark 3. As n gets larger, the lower bound for $\text{var}_\theta(T(\mathbf{X}))$ gets smaller. Thus, as the Fisher information increases, the lower bound decreases and the "best" estimator [one for which equality holds in (2)] will have smaller variance, consequently more information about θ.

Remark 4. Regularity condition (i) is unnecessarily restrictive. An examination of the proof shows thât it is only necessary that (ii) and (iii) hold for (2) to hold. Condition (i) excludes distributions such as $f_\theta(x) = 1/\theta, 0 < x < \theta$, for which (3) fails to hold. It also excludes densities such as $f_\theta(x) = 1, \theta < x < \theta + 1$, or $f_\theta(x) = (2/\pi)\sin^2(x+\pi), \theta \leq x \leq \theta + \pi$, each of which satisfies (iii) for $h \equiv 1$, so that (3) holds but not (1) for all h with $E_\theta|h| < \infty$.

Remark 5. Sufficient conditions for regularity condition (iii) may be found in most calculus textbooks. For example, if (i) and (ii) hold, then (iii) holds provided that for all h with $E_\theta|h| < \infty$ for all $\theta \in \Theta$, both $E_\theta\{h(\mathbf{X})[\partial \log f_\theta(\mathbf{X})/\partial\theta]\}$ and $E_\theta\ |h(\mathbf{X})[\partial f_\theta(\mathbf{X})/\partial\theta]|$ are continuous functions of θ. Regularity conditions (i) to (iii) are satisfied for a one-parameter exponential family.

Remark 6. The inequality (2) holds trivially if $I(\theta) = \infty$ [and $\psi'(\theta)$ is finite] or if $\text{var}_\theta(T(\mathbf{X})) = \infty$.

Example 1. Let $X \sim b(n, p)$; $\Theta = (0, 1) \subset \mathcal{R}$. Here the Fisher information may be obtained as follows:

$$\log f_p(x) = \log\binom{n}{x} + x\log p + (n-x)\log(1-p),$$

$$\frac{\partial \log f_p(x)}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p},$$

and

$$E_p\left[\frac{\partial \log f_p(X)}{\partial p}\right]^2 = \frac{n}{p(1-p)} = I(p).$$

Let $\psi(p)$ be a function of p and $T(X)$ be an unbiased estimator of $\psi(p)$. The only condition that need be checked is differentiability under the summation sign. We have

$$\psi(p) = E_pT(X) = \sum_{x=0}^{n}\binom{n}{x}T(x)p^x(1-p)^{n-x},$$

which is a polynomial in p and hence can be differentiated with respect to p. For any unbiased estimator $T(X)$ of p, we have

$$\text{var}_p(T(X)) \geq \frac{1}{n}p(1-p) = \frac{1}{I(p)},$$

and since

$$\text{var}\left(\frac{X}{n}\right) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n},$$

it follows that the variance of the estimator X/n attains the lower bound of the FCR inequality, and hence $T(X)$ has least variance among all unbiased estimators of p. Thus $T(X)$ is the UMVUE for p.

Example 2. Let $X \sim P(\lambda)$. We leave the reader to check that the regularity conditions are satisfied and

$$\text{var}_\lambda(T(X)) \geq \lambda.$$

Since $T(X) = X$ has variance λ, X is the UMVUE of λ. Similarly, if we take a sample of size n from $P(\lambda)$, we can show that

$$I_n(\lambda) = \frac{n}{\lambda} \quad \text{and} \quad \text{var}_\lambda(T(X_1, \ldots, X_n)) \geq \frac{\lambda}{n}$$

and $\overline{X}$ is the UMVUE.

Let us next consider the problem of unbiased estimation of $\psi(\lambda) = e^{-\lambda}$, based on a sample of size 1. The estimator

$$\partial(X) = \begin{cases} 1 & \text{if } X = 0, \\ 0 & \text{if } X \geq 1, \end{cases}$$

is unbiased for $\psi(\lambda)$ since

$$E_\lambda \partial(X) = E_\lambda[\partial(X)]^2 = P_\lambda\{X = 0\} = e^{-\lambda}.$$

Also,

$$\operatorname{var}_\lambda(\partial(X)) = e^{-\lambda}(1 - e^{-\lambda}).$$

To compute the FCR lower bound, we have

$$\log f_\lambda(x) = x \log \lambda - \lambda - \log x!.$$

This has to be differentiated with respect to $e^{-\lambda}$, since we want a lower bound for an estimator of the parameter $e^{-\lambda}$. Let $\theta = e^{-\lambda}$. Then

$$\log f_\theta(x) = x \log\log \frac{1}{\theta} + \log \theta - \log x!,$$

$$\frac{\partial}{\partial \theta} \log f_\theta(x) = x \frac{1}{\theta \log \theta} + \frac{1}{\theta},$$

and

$$\begin{aligned} E_\theta\left[\frac{\partial}{\partial \theta} \log f_\theta(X)\right]^2 &= \frac{1}{\theta^2}\left\{1 + \frac{2}{\log\theta} \log \frac{1}{\theta} + \frac{1}{(\log\theta)^2}\left[\log \frac{1}{\theta} + \left(\log \frac{1}{\theta}\right)^2\right]\right\} \\ &= e^{2\lambda}\left[1 - 2 + \frac{1}{\lambda^2}(\lambda + \lambda^2)\right] \\ &= \frac{e^{2\lambda}}{\lambda} = I(e^{-\lambda}), \end{aligned}$$

so that

$$\operatorname{var}_\theta T(X) \geq \frac{\lambda}{e^{2\lambda}} = \frac{1}{I(e^{-\lambda})},$$

where $\theta = e^{-\lambda}$.

Since $e^{-\lambda}(1 - e^{-\lambda}) > \lambda e^{-2\lambda}$ for $\lambda > 0$, we see that $\operatorname{var}(\delta(X))$ is greater than the lower bound obtained from the FCR inequality. We show next that $\delta(X)$ is the only unbiased estimator of θ, and hence is the UMVUE.

If h is any unbiased estimator of θ, it must satisfy $E_\theta h(X) = \theta$. That is, for all $\lambda > 0$,

$$e^{-\lambda} = \sum_{k=0}^{\infty} h(k) e^{-\lambda} \frac{\lambda^k}{k!}.$$

Equating coefficients of powers of λ we see immediately that $h(0) = 1$ and $h(k) = 0$ for $k = 1, 2, \ldots$. It follows that $h(X) = \partial(X)$.

The same computation can be carried out when $X_1, X_2, \ldots, X_n$ is a random sample from $P(\lambda)$. We leave the reader to show that the FCR lower bound for any unbiased estimator of $\theta = e^{-\lambda}$ is $\lambda e^{-2\lambda}/n$. The estimator $\sum_{i=1}^{n} \partial(X_i)/n$ is clearly unbiased for $e^{-\lambda}$ with variance $e^{-\lambda}(1 - e^{-\lambda})/n > (\lambda e^{-2\lambda})/n$. The UMVUE of $e^{-\lambda}$ is given by $T_0 = [(n-1)/n]^{\sum_{i=1}^{n} X_i}$ with $\text{var}_\lambda(T_0) = e^{-2\lambda}(e^{\lambda/n} - 1) > (\lambda e^{-2\lambda})/n$ for all $\lambda > 0$.

Corollary. Let $X_1, X_2, \ldots, X_n$ be iid with common PDF $f_\theta(x)$. Suppose that the family $\{f_\theta : \theta \in \Theta\}$ satisfies the conditions of Theorem 1. Then equality holds in (2) if and only if for all $\theta \in \Theta$,

$$T(\mathbf{x}) - \psi(\theta) = k(\theta)\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{x}) \tag{8}$$

for some function $k(\theta)$.

Proof. Recall that we derived (2) by an application of the Cauchy–Schwatz inequality where equality holds if and only if (8) holds.

Remark 7. Integrating (8) with respect to θ, we get

$$\log f_\theta(\mathbf{x}) = Q(\theta)T(\mathbf{x}) + S(\theta) + A(\mathbf{x})$$

for some functions Q, S, and A. It follows that f_θ is a one-parameter exponential family and the statistic T is sufficient for θ.

Remark 8. A result that simplifies computations is the following. If f_θ is twice differentiable and $E_\theta\left\{\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{X})\right\}$ can be differentiated under the expectation sign, then

$$I(\theta) = E_\theta\left[\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{X})\right]^2 = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(\mathbf{X})\right]. \tag{9}$$

For the proof of (9), it is straightforward to check that

$$\frac{\partial^2}{\partial\theta^2}\log f_\theta(\mathbf{x}) = \frac{f_\theta''(\mathbf{x})}{f_\theta(\mathbf{x})} - \left[\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{x})\right]^2.$$

Taking expectations on both sides we get (9).

Example 3. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\mu, 1)$. Then

$$\log f_\mu(x) = -\frac{1}{2}\log(2\pi) - \frac{(x-\mu)^2}{2},$$

$$\frac{\partial}{\partial\mu}\log f_\mu(x) = x - \mu,$$

and

$$\frac{\partial^2}{\partial\mu^2}\log f_\mu(x) = -1.$$

Hence $I(\mu) = 1$ and $I_n(\mu) = n$.

We next consider an inequality due to Chapman, Robbins, and Kiefer (the CRK inequality) that gives a lower bound for the variance of an estimator but does not require regularity conditions of the Fréchet–Cramér–Rao type.

Theorem 2 (Chapman and Robbins [11], Kiefer [50]). Let $\Theta \subset \mathcal{R}$ and $\{f_\theta(\mathbf{x}) : \theta \in \Theta\}$ be a class of PDFs (PMFs). Let ψ be defined on Θ, and let T be an unbiased estimator of $\psi(\theta)$ with $E_\theta T^2 < \infty$ for all $\theta \in \Theta$. If $\theta \neq \varphi$, assume that f_θ and f_φ are different and assume further that there exists a $\varphi \in \Theta$ such that $\theta \neq \varphi$ and

$$S(\theta) = \{f_\theta(\mathbf{x}) > 0\} \supset S(\varphi) = \{f_\varphi(\mathbf{x}) > 0\}. \tag{10}$$

Then

$$\operatorname{var}_\theta(T(\mathbf{X})) \geq \sup_{\{\varphi: S(\varphi)\subset S(\theta),\, \varphi\neq\theta\}} \frac{[\psi(\varphi) - \psi(\theta)]^2}{\operatorname{var}_\theta[f_\varphi(\mathbf{X})/f_\theta(\mathbf{X})]} \tag{11}$$

for all $\theta \in \Omega$.

Proof. Since T is unbiased for ψ, $E_\varphi T(\mathbf{X}) = \psi(\varphi)$ for all $\varphi \in \Theta$. Hence, for $\varphi \neq \theta$,

$$\int_{S(\theta)} T(\mathbf{x})\frac{f_\varphi(\mathbf{x}) - f_\theta(\mathbf{x})}{f_\theta(\mathbf{x})} f_\theta(\mathbf{x})\,d\mathbf{x} = \psi(\varphi) - \psi(\theta), \tag{12}$$

which yields

$$\operatorname{cov}_\theta\left[T(\mathbf{X}), \frac{f_\varphi(\mathbf{X})}{f_\theta(\mathbf{X})} - 1\right] = \psi(\varphi) - \psi(\theta).$$

Using the Cauchy–Schwarz inequality, we get

$$\operatorname{cov}^2_\theta\left[T(\mathbf{X}), \frac{f_\varphi(\mathbf{X})}{f_\theta(\mathbf{X})} - 1\right] \leq \operatorname{var}_\theta(T(\mathbf{X}))\operatorname{var}_\theta\left[\frac{f_\varphi(\mathbf{X})}{f_\theta(\mathbf{X})} - 1\right]$$

$$= \text{var}_\theta(T(\mathbf{X}))\,\text{var}_\theta\left[\frac{f_\varphi(\mathbf{X})}{f_\theta(\mathbf{X})}\right].$$

Thus

$$\text{var}_\theta(T(\mathbf{X})) \geq \frac{[\psi(\varphi) - \psi(\theta)]^2}{\text{var}_\theta\{f_\varphi(\mathbf{X})/f_\theta(\mathbf{X})\}},$$

and the result follows. In the discrete case it is necessary only to replace the integral in the left side of (12) by a sum. The rest of the proof needs no change.

Remark 9. Inequality (11) holds without any regularity conditions on f_θ or $\psi(\theta)$. We will show that it covers some nonregular cases of the FCR inequality. Sometimes (11) is available in an alternative form. Let θ and $\theta + \delta(\delta \neq 0)$ be any two distinct values in Θ such that $S(\theta+\delta) \subset S(\theta)$, and take $\psi(\theta) = \theta$. Write

$$J = J(\theta, \delta) = \frac{1}{\delta^2}\left\{\left[\frac{f_{\theta+\delta}(\mathbf{X})}{f_\theta(\mathbf{X})}\right]^2 - 1\right\}.$$

Then (11) can be written as

$$\text{(13)} \qquad \text{var}_\theta(T(\mathbf{X})) \geq \frac{1}{\inf\limits_\delta E_\theta J},$$

where the infimum is taken over all $\delta \neq 0$ such that $S(\theta+\delta) \subset S(\theta)$.

Remark 10. Inequality (11) applies if the parameter space is discrete, but the Fréchet–Cramér–Rao regularity conditions do not hold in that case.

Example 4. Let X be $U[0, \theta]$. The regularity conditions of FCR inequality do not hold in this case. Let $\psi(\theta) = \theta$. If $\varphi < \theta$, then $S(\varphi) \subset S(\theta)$. Also,

$$E_\theta\left[\frac{f_\varphi(X)}{f_\theta(X)}\right]^2 = \int_0^\varphi \left(\frac{\theta}{\varphi}\right)^2 \frac{1}{\theta}\,dx = \frac{\theta}{\varphi}.$$

Thus

$$\text{var}_\theta(T(X)) \geq \sup_{\varphi:\varphi<\theta} \frac{(\varphi-\theta)^2}{(\theta/\varphi) - 1} = \sup_{\varphi:\varphi<\theta} \varphi(\theta - \varphi) = \frac{\theta^2}{4}$$

for any unbiased estimator $T(X)$ of θ. X is a complete sufficient statistic, and $2X$ is unbiased for θ so that $T(X) = 2X$ is the UMVUE. Also,

$$\text{var}_\theta(2X) = 4\,\text{var}\,X = \frac{\theta^2}{3} > \frac{\theta^2}{4}.$$

Thus the lower bound of $\theta^2/4$ of the CRK inequality is not achieved by any unbiased estimator of θ.

Example 5. Let X have PMF

$$P_N\{X=k\}=\begin{cases}\dfrac{1}{N}, & k=1,2,\dots,N,\\ 0, & \text{otherwise.}\end{cases}$$

Let $\Theta=\{N\colon N\geq M,\ M>1 \text{ given}\}$. Take $\psi(N)=N$. Although the FCR regularity conditions do not hold, (11) is applicable since for $N\neq N'\in\Theta\subset\mathcal{R}$,

$$S(N)=\{1,2,\dots,N\}\supset S(N')=\{1,2,\dots,N'\}\qquad \text{if } N'<N.$$

Also, P_N and $P_{N'}$ are different for $N\neq N'$. Thus

$$\text{var}_N(T)\geq \sup_{N'<N}\frac{(N-N')^2}{\text{var}_N\{P_{N'}/P_N\}}.$$

Now

$$\frac{P_{N'}}{P_N}(x)=\frac{P_{N'}(x)}{P_N(x)}=\begin{cases}\dfrac{N}{N'}, & x=1,2,\dots,N',\ N'<N,\\ 0, & \text{otherwise,}\end{cases}$$

$$E_N\left[\frac{P_{N'}(X)}{P_N(X)}\right]^2=\frac{1}{N}\sum_1^{N'}\left(\frac{N}{N'}\right)^2=\frac{N}{N'},$$

and

$$\text{var}_N\left[\frac{P_{N'}(X)}{P_N(X)}\right]=\frac{N}{N'}-1>0\qquad \text{for } N>N'.$$

It follows that

$$\text{var}_N(T(X))\geq \sup_{N'<N}\frac{(N-N')^2}{(N-N')/N'}=\sup_{N'<N}N'(N-N').$$

Now

$$\frac{k(N-k)}{(k-1)(N-k+1)}>1\qquad \text{if and only if } k<\frac{N+1}{2},$$

so that $N'(N-N')$ increases as long as $N'<(N+1)/2$ and decreases if $N'>(N+1)/2$. The maximum is achieved at $N'=[(N+1)/2]$ if $M\leq(N+1)/2$ and at $N'=M$ if $M>(N+1)/2$, where $[x]$ is the largest integer $\leq x$. Therefore,

$$\operatorname{var}_N(T(X)) \geq \frac{N+1}{2}\left(N - \frac{N+1}{2}\right), \qquad \text{if } M \leq \frac{N+1}{2},$$

and

$$\operatorname{var}_N(T(X)) \geq M(N-M), \qquad \text{if } M > \frac{N+1}{2}.$$

Example 6. Let $X \sim \mathcal{N}(0, \sigma^2)$. Let us compute J (see Remark 9) for $\delta \neq 0$.

$$\begin{aligned} J = \frac{1}{\delta^2}\left\{\left[\frac{f_{\sigma+\delta}(\mathbf{X})}{f_\sigma(\mathbf{X})}\right]^2 - 1\right\} &= \frac{1}{\delta^2}\left\{\frac{\sigma^{2n}}{(\sigma+\delta)^{2n}} \exp\left[-\frac{\sum X_i^2}{(\sigma+\delta)^2} + \frac{\sum X_i^2}{\sigma^2}\right] - 1\right\} \\ &= \frac{1}{\delta^2}\left\{\left(\frac{\sigma}{\sigma+\delta}\right)^{2n} \exp\left[\frac{\sum X_i^2(\delta^2 + 2\sigma\delta)}{\sigma^2(\sigma+\delta)^2}\right] - 1\right\}, \end{aligned}$$

and

$$E_\sigma J = \frac{1}{\delta^2}\left(\frac{\sigma}{\sigma+\delta}\right)^{2n} E_\sigma\left[\exp\left(c\frac{\sum X_i^2}{\sigma^2}\right)\right] - \frac{1}{\delta^2},$$

where $c = (\delta^2 + 2\sigma\delta)/(\sigma+\delta)^2$.

Since $\sum X_i^2/\sigma^2 \sim \chi^2(n)$,

$$E_\sigma J = \frac{1}{\delta^2}\left[\left(\frac{\sigma}{\sigma+\delta}\right)^{2n} \frac{1}{(1-2c)^{n/2}} - 1\right] \qquad \text{for } c < \frac{1}{2}.$$

Let $k = \delta/\sigma$; then

$$c = \frac{2k + k^2}{(1+k)^2} \quad \text{and} \quad 1 - 2c = \frac{1 - 2k - k^2}{(1+k)^2},$$

and

$$E_\sigma J = \frac{1}{k^2\sigma^2}[(1+k)^{-n}(1 - 2k - k^2)^{-n/2} - 1].$$

Here $1 + k > 0$ and $1 - 2c > 0$, so that $1 - 2k - k^2 > 0$, implying that $-\sqrt{2} < k + 1 < \sqrt{2}$ and also that $k > -1$. Thus $-1 < k < \sqrt{2} - 1$ and $k \neq 0$. Also,

$$\lim_{k\to 0} E_\sigma J = \lim_{k\to 0} \frac{(1+k)^{-n}(1-2k-k^2)^{-n/2}-1}{k^2\sigma^2}$$
$$= \frac{2n}{\sigma^2}$$

by L'Hospital's rule. We leave the reader to check that this is the FCR lower bound for $\text{var}_\sigma(T(\mathbf{X}))$. But the minimum value of $E_\sigma J$ is not achieved in the neighborhood of $k = 0$, so that the CRK inequality is sharper than the FCR inequality. Next, we show that for $n = 2$ we can do better with the CRK inequality. We have

$$E_\sigma J = \frac{1}{k^2\sigma^2}\left[\frac{1}{(1-2k-k^2)(1+k)^2} - 1\right]$$
$$= \frac{(k+2)^2}{\sigma^2(1+k)^2(1-2k-k^2)}, \qquad -1 < k < \sqrt{2}-1, \quad k \neq 0.$$

For $k = -0.1607$ we achieve the lower bound as $(E_\sigma J)^{-1} = 0.2698\sigma^2$, so that $\text{var}_\sigma(T(X)) \geq 0.2698\sigma^2 > \sigma^2/4$. Finally, we show that this bound is by no means the best available; it is possible to improve on the Chapman–Robbins–Kiefer bounds, too, in some cases. Take

$$T(X_1, X_2, \ldots, X_n) = \frac{\Gamma(n/2)}{\Gamma[(n+1)/2]}\frac{\sigma}{\sqrt{2}}\sqrt{\frac{\sum_1^n X_i^2}{\sigma^2}}$$

to be an estimate of σ. Now $E_\sigma T = \sigma$ and

$$E_\sigma T^2 = \frac{\sigma^2}{2}\left[\frac{\Gamma(n/2)}{\Gamma[(n+1)/2]}\right]^2 E\left(\frac{\sum_1^n X_i^2}{\sigma^2}\right)$$
$$= \frac{n\sigma^2}{2}\left[\frac{\Gamma(n/2)}{\Gamma[(n+1)/2]}\right]^2$$

so that

$$\text{var}_\sigma(T) = \sigma^2\left[\frac{n}{2}\left(\frac{\Gamma(n/2)}{\Gamma[(n+1)/2]}\right)^2 - 1\right].$$

For $n = 2$,

$$\text{var}_\sigma(T) = \sigma^2\left(\frac{4}{\pi} - 1\right) = 0.2732\sigma^2,$$

which is $> 0.2698\sigma^2$, the CRK bound. Note that T is the UMVUE.

Remark 11. In general the CRK inequality is as sharp as the FCR inequality. See Chapman and Robbins [11, pp. 584–585], for details.

We next introduce the concept of efficiency.

Definition 2. Let T_1, T_2 be two unbiased estimators for a parameter θ. Suppose that $E_\theta T_1^2 < \infty$, $E_\theta T_2^2 < \infty$. We define the *efficiency* of T_1 relative to T_2 by

$$\text{eff}_\theta(T_1 \mid T_2) = \frac{\text{var}_\theta(T_2)}{\text{var}_\theta(T_1)} \tag{14}$$

and say that T_1 is more efficient than T_2 if

$$\text{eff}_\theta(T_1 \mid T_2) > 1. \tag{15}$$

It is usual to consider the performance of an unbiased estimator by comparing its variance with the lower bound given by the FCR inequality.

Definition 3. Assume that the regularity conditions of the FCR inequality are satisfied by the family of DFs $\{F_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathcal{R}$. We say that an unbiased estimator T for parameter θ is most efficient for the family $\{F_\theta\}$ if

$$\text{var}_\theta(T) = \left\{E_\theta\left[\frac{\partial \log f_\theta(\mathbf{X})}{\partial \theta}\right]^2\right\}^{-1} = I_n(\theta). \tag{16}$$

Definition 4. Let T be the most efficient estimator for the regular family of DFs $\{F_\theta, \theta \in \Theta\}$. Then the *efficiency of any unbiased estimator* T_1 of θ is defined as

$$\text{eff}_\theta(T_1) = \text{eff}_\theta(T_1 \mid T) = \frac{\text{var}_\theta(T)}{\text{var}_\theta(T_1)} = \frac{I_n(\theta)}{\text{var}_\theta(T_1)}. \tag{17}$$

Clearly, the efficiency of the most efficient estimator is 1, and the efficiency of any unbiased estimator T_1 is < 1.

Definition 5. We say that an estimator T_1 is *asymptotically* (*most*) *efficient* if

$$\lim_{n\to\infty} \text{eff}_\theta(T_1) = 1 \tag{18}$$

and T_1 is at least asymptotically unbiased in the sense that $\lim_{n\to\infty} E_\theta T_1 = \theta$. Here n is the sample size.

Remark 12. Definition 3, although in common use, has many drawbacks. We have already seen cases in which the regularity conditions are not satisfied and yet UMVUEs exist. The definition does not cover such cases. Moreover, in many cases where the regularity conditions are satisfied and UMVUEs exist, the UMVUE is not most efficient since the variance of the best estimator (the UMVUE) does not achieve the lower bound of the FCR inequality.

Example 7. Let $X \sim b(n, p)$. Then we have seen in Example 1 that X/n is the UMVUE since its variance achieves the lower bound of the FCR inequality. It follows that X/n is most efficient.

Example 8. Let $X_1, X_2, \ldots, X_n$ be iid $P(\lambda)$ RVs and suppose that $\psi(\lambda) = P_\lambda(X = 0) = e^{-\lambda}$. From Example 2, the UMVUE of ψ is given by $T_0 = [(n-1)/n]^{\sum_{i=1}^n X_i}$ with

$$\text{var}_\lambda(T_0) = e^{-2\lambda}(e^{\lambda/n} - 1).$$

Also, $I_n(\theta) = (\lambda e^{-2\lambda})/n$. It follows that

$$\text{eff}_\lambda(T_0) = \frac{(\lambda e^{-2\lambda})/n}{e^{-2\lambda}(e^{\lambda/n} - 1)} < \frac{\lambda e^{-2\lambda})/n}{e^{-2\lambda}(\lambda/n)} = 1$$

since $e^x - 1 > x$ for $x > 0$. Thus T_0 is not most efficient. However, since $\text{eff}_\lambda(T_0) \to 1$ as $n \to \infty$, T_0 is asymptotically efficient.

In view of Remarks 6 and 7, the following result describes the relationship between most efficient unbiased estimators and UMVUEs.

Theorem 3. A necessary and sufficient condition for an unbiased estimator T of ψ to be most efficient is that T be sufficient and the relation (8) holds for some function $k(\theta)$.

Clearly, an estimator T satisfying the conditions of Theorem 3 will be the UMVUE, and two estimators coincide. We emphasize that we have assumed the regularity conditions of FCR inequality in making this statement.

Example 9. Let (X, Y) be jointly distributed with PDF

$$f_\theta(x, y) = \exp\left[-\left(\frac{x}{\theta} + \theta y\right)\right], \qquad x > 0, \quad y > 0.$$

For a sample (x, y) of size 1, we have

$$-\frac{\partial}{\partial\theta} \log f_\theta(x, y) = \frac{\partial}{\partial\theta}\left(\frac{x}{\theta} + \theta y\right) = -\frac{x}{\theta^2} + y.$$

Hence, information for this sample is

$$I(\theta) = E_\theta\left(Y - \frac{X}{\theta^2}\right)^2 = E_\theta(Y^2) + \frac{E(X^2)}{\theta^4} - \frac{2E(XY)}{\theta^2}.$$

Now

$$E_\theta(Y^2) = \frac{2}{\theta^2}, \quad E_\theta(X^2) = 2\theta^2, \quad \text{and} \quad E(XY) = 1,$$

so that

$$I(\theta) = \frac{2}{\theta^2} + \frac{2}{\theta^2} - \frac{2}{\theta^2} = \frac{2}{\theta^2}.$$

Therefore, the Fisher information in a sample of n pairs is $2n/\theta^2$.

We return to Example 8.3.23, where $X_1, X_2, \ldots, X_n$ are iid $G(1, \theta)$ and $Y_1, Y_2, \ldots, Y_n$ are iid $G(1, 1/\theta)$, and X's and Y's are independent. Then (X_1, Y_1) has common PDF $f_\theta(x, y)$ given above. We will compute Fisher's Information for θ in the family of PDFs of $S(\mathbf{X}, \mathbf{Y}) = (\sum X_i / \sum Y_i)^{1/2}$. Using the PDFs of $\sum X_i \sim G(n, \theta)$ and $\sum Y_i \sim G(n, 1/\theta)$ and the transformation technique, it is easy to see that $S(\mathbf{X}, \mathbf{Y})$ has PDF

$$g_\theta(s) = \frac{2\Gamma(2n)}{[\Gamma(n)]^2} s^{-1} \left(\frac{s}{\theta} + \frac{\theta}{s}\right)^{-2n}, \qquad s > 0.$$

Thus

$$\frac{\partial \log g_\theta(s)}{\partial\theta} = -2n\left(-\frac{s}{\theta^2} + \frac{1}{s}\right)\left(\frac{s}{\theta} + \frac{\theta}{s}\right)^{-1}.$$

It follows that

$$\begin{aligned} E_\theta\left[\frac{\partial}{\partial\theta} \log g_\theta(S)\right]^2 &= \frac{4n^2}{\theta^2} E_\theta\left[1 - 4\left(\frac{S}{\theta} + \frac{\theta}{S}\right)^{-2}\right] \\ &= \frac{4n^2}{\theta^2}\left[1 - 4\frac{n}{2(2n+1)}\right] = \frac{2n}{\theta^2}\frac{2n}{2n+1} \\ &< \frac{2n}{\theta^2}. \end{aligned}$$

That is, the information about θ in S is smaller than that in the sample.

The Fisher nformation in the conditional PDF of S given $A = a$, where $A(\mathbf{X}, \mathbf{Y}) = S_1(\mathbf{X})S_2(\mathbf{Y})$ can be shown (Problem 12) to equal

$$\frac{2a}{\theta^2}\frac{K_1(2a)}{K_0(2a)},$$

where K_0 and K_1 are Bessel functions of order 0 and 1, respectively. Averaging over all values of A, one can show that the information is $2n/\theta^2$, which is the total Fisher information in the sample of n pairs (x_j, y_j)'s.

PROBLEMS 8.5

1. Are the following families of distributions regular in the sense of Fréchet, Cramér, and Rao? If so, find the lower bound for the variance of an unbiased estimator based on a sample size n.
   (a) $f_\theta(x) = \theta^{-1}e^{-x/\theta}$ if $x > 0$, and $= 0$ otherwise; $\theta > 0$.
   (b) $f_\theta(x) = e^{-(x-\theta)}$ if $\theta < x < \infty$, and $= 0$ otherwise.
   (c) $f_\theta(x) = \theta(1-\theta)^x$, $x = 0, 1, 2, \ldots; 0 < \theta < 1$.
   (d) $f(x; \sigma^2) = (1/\sigma\sqrt{2\pi})e^{-x^2/2\sigma^2}$, $-\infty < x < \infty; \sigma^2 > 0$.

2. Find the CRK lower bound for the variance of an unbiased estimator of θ, based on a sample of size n from the PDF of Problem 1(b).

3. Find the CRK bound for the variance of an unbiased estimator of θ in sampling from $\mathcal{N}(\theta, 1)$.

4. In Problem 1 check to see whether there exists a most efficient estimator in each case.

5. Let $X_1, X_2, \ldots, X_n$ be a sample from a three-point distribution:

$$P\{X = y_1\} = \frac{1-\theta}{2}, \qquad P\{X = y_2\} = \frac{1}{2}, \quad \text{and} \quad P\{X = y_3\} = \frac{\theta}{2},$$

   where $0 < \theta < 1$. Does the FCR inequality apply in this case? If so, what is the lower bound for the variance of an unbiased estimator of θ?

6. Let $X_1, X_2, \ldots, X_n$ be iid RVs with mean μ and finite variance. What is the efficiency of the unbiased (and consistent) estimator $[2/n(n+1)]\sum_{i=1}^n iX_i$ relative to $\overline{X}$?

7. When does the equality hold in the CRK inequality?

8. Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(\mu, 1)$, and let $d(\mu) = \mu^2$.
   (a) Show that the minimum variance of any estimator of μ^2 from the FCR inequality is $4\mu^2/n$.
   (b) Show that $T(X_1, X_2, \ldots, X_n) = \overline{X}^2 - 1/n$ is the UMVUE of μ^2 with variance $(4\mu^2/n + 2/n^2)$.

9. Let $X_1, X_2, \ldots, X_n$ be iid $G(1, 1/\alpha)$ RVs.
   (a) Show that the estimator $T(X_1, X_2, \ldots, X_n) = (n-1)/n\overline{X}$ is the UMVUE for α with variance $a^2/(n-2)$.
   (b) Show that the minimum variance from FCR inequality is α^2/n.

10. In Problem 8.4.16, compute the relative efficiency of $\hat{\mu}_0$ with respect to $\hat{\mu}_1$.

11. Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent samples from $\mathcal{N}(\mu, \sigma_1^2)$ and $\mathcal{N}(\mu, \sigma_2^2)$, respectively, where $\mu, \sigma_1^2, \sigma_2^2$ are unknown. Let $\rho = \sigma_2^2/\sigma_1^2$ and $\theta = m/n$, and consider the problem of unbiased estimation of μ.

(a) If ρ is known, show that

$$\hat{\mu}_0 = \alpha\overline{X} + (1-\alpha)\overline{Y},$$

where $\alpha = \rho/(\rho+\theta)$ is the BLUE of μ. Compute $\operatorname{var}(\hat{\mu}_0)$.

(b) If ρ is unknown, the unbiased estimator

$$\bar{\mu} = \frac{\overline{X}+\theta\overline{Y}}{1+\theta}$$

is optimum in the neighborhood of $\rho = 1$. Find the variance of $\bar{\mu}$.

(c) Compute the efficiency of $\bar{\mu}$ relative to $\hat{\mu}_0$.

(d) Another unbiased estimator of μ is

$$\hat{\mu} = \frac{\rho F\overline{X}+\theta\overline{Y}}{\theta+\rho F},$$

where $F = S_2^2/\rho S_1^2$ is an $F(m-1, n-1)$ RV.

12. Show that the Fisher information on θ based on the PDF

$$\frac{1}{2K_0(2a)} s \exp\left[-a\left(\frac{s}{\theta}+\frac{\theta}{s}\right)\right]$$

for fixed a equals $(2a/\theta^2)[K_1(2a)/K_0(2a)]$, where $K_0(2a)$ and $K_1(2a)$ are Bessel functions of order 0 and 1, respectively.

8.6 SUBSTITUTION PRINCIPLE (METHOD OF MOMENTS)

One of the simplest and oldest methods of estimation is the *substitution principle*: Let $\psi(\theta)$, $\theta \in \Theta$ be a parametric function to be estimated on the basis of a random sample $X_1, X_2, \dots, X_n$ from a population DF F. Suppose that we can write $\psi(\theta) = h(F)$ for some known function h. Then the substitution principle estimator of $\psi(\theta)$ is $h(F_n^*)$, where F_n^* is the sample distribution function. Accordingly, we estimate $\mu = \mu(F)$ by $\mu(F_n^*) = \overline{X}$, $m_k = E_F X^k$ by $\sum_{j=1}^{n} X_j/n$, and so on. The *method of moments* is a special case when we need to estimate some known function of a finite number of unknown moments. Let us suppose that we are interested in estimating

$$\theta = h(m_1, m_2, \dots, m_k), \tag{1}$$

where h is some known numerical function and m_j is the jth-order moment of the population distribution that is known to exist for $1 \le j \le k$.

Definition 1. The *method of moments* consists in estimating θ by the statistic

$$(2)\qquad T(X_1,\ldots,X_n)=h\left(n^{-1}\sum_1^n X_i, n^{-1}\sum_1^n X_i^2,\ldots,n^{-1}\sum_1^n X_i^k\right).$$

To make sure that T is a statistic, we will assume that $h:\mathcal{R}_k\to\mathcal{R}$ is a Borel-measurable function.

Remark 1. It is easy to extend the method to the estimation of joint moments. Thus we use $n^{-1}\sum_1^n X_iY_i$ to estimate $E(XY)$, and so on.

Remark 2. From the WLLN, $n^{-1}\sum_{i=1}^n X_i^j \xrightarrow{P} EX^j$. Thus, if one is interested in estimating the population moments, the method of moments leads to consistent and unbiased estimators. Moreover, the method of moments estimators in this case are asymptotically normally distributed (see Section 7.5).

Again, if one estimates parameters of the type θ defined in (1) and h is a continuous function, the estimators $T(X_1, X_2,\ldots,X_n)$ defined in (2) are consistent for θ (see Problem 1). Under some mild conditions on h, the estimator T is also asymptotically normal (see Cramér [16, pp. 386–387]).

Example 1. Let $X_1, X_2,\ldots,X_n$ be iid RVs with common mean μ and variance σ^2. Then $\sigma=\sqrt{m_2-m_1^2}$, and the method of moments estimator for σ is given by

$$T(X_1,\ldots,X_n)=\sqrt{\frac{1}{n}\sum_1^n X_i^2-\frac{(\sum X_i)^2}{n^2}}.$$

Although T is consistent and asymptotically normal for σ, it is not unbiased.

In particular, if $X_1, X_2,\ldots,X_n$ are iid $P(\lambda)$ RVs, we know that $EX_1=\lambda$ and $\text{var}(X_1)=\lambda$. The method of moments leads to using either $\overline{X}$ or $\sum_1^n(X_i-\overline{X})^2/n$ as an estimator of λ. To avoid this kind of ambiguity we take the estimator involving the lowest-order sample moment.

Example 2. Let $X_1, X_2,\ldots,X_n$ be a sample from

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

Then

$$EX=\frac{a+b}{2}\quad\text{and}\quad\text{var}(X)=\frac{(b-a)^2}{12}.$$

The method of moments leads to estimating EX by $\overline{X}$ and $\operatorname{var}(X)$ by $\sum_1^n(X_i - \overline{X})^2/n$, so that the estimators for a and b, respectively, are

$$T_1(X_1, \dots, X_n) = \overline{X} - \sqrt{\frac{3\sum_1^n(X_i - \overline{X})^2}{n}}$$

and

$$T_2(X_1, \dots, X_n) = \overline{X} + \sqrt{\frac{3\sum_1^n(X_i - \overline{X})^2}{n}}.$$

Example 3. Let $X_1, X_2, \dots, X_N$ be iid $b(n, p)$ RVs, where both n and p are unknown. The method of moments estimators of p and n are given by

$$\overline{X} = EX = np$$

and

$$\frac{1}{N}\sum_1^N X_i^2 = EX^2 = np(1-p) + n^2p^2.$$

Solving for n and p, we get the estimator for p as

$$T_1(X_1, \dots, X_N) = \frac{\overline{X}}{T_2(X_1, \dots, X_N)},$$

where $T_2(X_1, \dots, X_N)$ is the estimator for n, given by

$$T_2(X_1, X_2, \dots, X_N) = \frac{(\overline{X})^2}{\overline{X} + \overline{X}^2 - \left(\sum_1^N X_i^2/N\right)}.$$

Note that $\overline{X} \xrightarrow{P} np$, $\sum_1^N X_i^2/N \xrightarrow{P} np(1-p) + n^2p^2$, so that both T_1 and T_2 are consistent estimators.

Method of moments may lead to absurd estimators. The reader is asked to compute estimators of θ in $\mathcal{N}(\theta, \theta)$ or $\mathcal{N}(\theta, \theta^2)$ by the method of moments and verify this assertion.

PROBLEMS 8.6

1. Let $X_n \xrightarrow{P} a$, and $Y_n \xrightarrow{P} b$, where a and b are constants. Let $h : \mathcal{R}_2 \to \mathcal{R}$ be a continuous function. Show that $h(X_n, Y_n) \xrightarrow{P} h(a, b)$.

2. Let $X_1, X_2, \ldots, X_n$ be a sample from $G(\alpha, \beta)$. Find the method of moments estimator for (α, β).

3. Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$. Find the method of moments estimator for (μ, σ^2).

4. Let $X_1, X_2, \ldots, X_n$ be a sample from $B(\alpha, \beta)$. Find the method of moments estimator for (α, β).

5. A random sample of size n is taken from the lognormal PDF

$$f(x; \mu, \sigma) = (\sigma\sqrt{2\pi})^{-1}x^{-1}\exp\left[-\frac{1}{2\sigma^2}(\log x - \mu)^2\right], \qquad x > 0.$$

Find the method of moments estimators for μ and σ^2.

8.7 MAXIMUM LIKELIHOOD ESTIMATORS

In this section we study a frequently used method of estimation, namely, the *method of maximum likelihood estimation*. Consider the following example.

Example 1. Let $X \sim b(n, p)$. One observation on X is available, and it is known that n is either 2 or 3 and $p = \frac{1}{2}$ or $\frac{1}{3}$. Our objective is to estimate the pair (n, p). The following table gives the probability that $X = x$ for each possible pair (n, p):

x	$(2, \frac{1}{2})$	$(2, \frac{1}{3})$	$(3, \frac{1}{2})$	$(3, \frac{1}{3})$	Maximum Probability
0	$\frac{1}{4}$	$\frac{4}{9}$	$\frac{1}{8}$	$\frac{8}{27}$	$\frac{4}{9}$
1	$\frac{1}{2}$	$\frac{4}{9}$	$\frac{3}{8}$	$\frac{12}{27}$	$\frac{1}{2}$
2	$\frac{1}{4}$	$\frac{1}{9}$	$\frac{3}{8}$	$\frac{6}{27}$	$\frac{3}{8}$
3	0	0	$\frac{1}{8}$	$\frac{1}{27}$	$\frac{1}{8}$

The last column gives the maximum probability in each row, that is, for each value that X assumes. If the value $x = 1$, say, is observed, it is more probable that it came from the distribution $b(2, \frac{1}{2})$ than from any of the other distributions, and so on. The following estimator is therefore reasonable in that it maximizes the probability of the value observed:

$$(\hat{n}, \hat{p})(x) = \begin{cases} (2, \frac{1}{3}) & \text{if } x = 0, \\ (2, \frac{1}{2}) & \text{if } x = 1, \\ (3, \frac{1}{2}) & \text{if } x = 2, \\ (3, \frac{1}{2}) & \text{if } x = 3. \end{cases}$$

The *principle of maximum likelihood* essentially assumes that the sample is representative of the population and chooses as the estimator that value of the parameter which maximizes the PDF (PMF) $f_\theta(x)$.

Definition 1. Let $(X_1, X_2, \dots, X_n)$ be a random vector with PDF (PMF) f_θ $(x_1, x_2, \dots, x_n), \theta \in \Theta$. The function

$$(1) \qquad L(\theta;\ x_1, x_2, \dots, x_n) = f_\theta(x_1, x_2, \dots, x_n),$$

considered as a function of θ, is called the *likelihood function.*

Usually, θ will be a multiple parameter. If $X_1, X_2, \dots, X_n$ are iid with PDF (PMF) $f_\theta(x)$, the likelihood function is

$$(2) \qquad L(\theta;\ x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

Let $\Theta \subseteq \mathcal{R}_k$ and $\mathbf{X} = (X_1, X_2, \dots, X_n)$.

Definition 2. The *principle of maximum likelihood estimation* consists of choosing as an estimator of $\boldsymbol{\theta}$ a $\hat{\boldsymbol{\theta}}(\mathbf{X})$ that maximizes $L(\boldsymbol{\theta};\ x_1, x_2, \dots, x_n)$, that is, to find a mapping $\hat{\boldsymbol{\theta}}$ of $\mathcal{R}_n \rightarrow \mathcal{R}_k$ that satisfies

$$(3) \qquad L(\hat{\boldsymbol{\theta}};\ x_1, x_2, \dots, x_n) = \sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} L(\boldsymbol{\theta};\ x_1, x_2, \dots, x_n).$$

(Constants are not admissible as estimators.) If a $\hat{\boldsymbol{\theta}}$ satisfying (3) exists, we call it a *maximum likelihood estimator* (MLE).

It is convenient to work with the logarithm of the likelihood function. Since log is a monotone function,

$$(4) \qquad \log L(\hat{\boldsymbol{\theta}};\ x_1, \dots, x_n) = \sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \log L(\boldsymbol{\theta};\ x_1, \dots, x_n).$$

Let Θ be an open subset of $\mathcal{R}_k$, and suppose that $f_\theta(\mathbf{x})$ is a positive, differentiable function of $\boldsymbol{\theta}$ (that is, the first-order partial derivatives exist in the components of $\boldsymbol{\theta}$). If a supremum $\hat{\boldsymbol{\theta}}$ exists, it must satisfy the *likelihood equations*

$$(5) \qquad \frac{\partial \log L(\hat{\boldsymbol{\theta}};\ x_1, \dots, x_n)}{\partial \theta_j} = 0, \qquad j = 1, 2, \dots, k, \quad \boldsymbol{\theta} = (\theta_1, \dots, \theta_k).$$

Any nontrivial root of the likelihood equations (5) is called an MLE in the *loose sense*. A parameter value that provides the absolute maximum of the likelihood function is called an MLE in the *strict* sense or, simply, an MLE.

Remark 1. If $\Theta \subseteq \mathcal{R}$, there may still be many problems. Often, the likelihood equation $\partial L/\partial \theta = 0$ has more than one root, or the likelihood function is not differentiable everywhere in Θ, or $\hat{\theta}$ may be a terminal value. Sometimes the likelihood equation may be quite complicated and difficult to solve explicitly. In that case one may have to resort to some numerical procedure to obtain the estimator. Similar remarks apply to the multiparameter case.

Example 2. Let $X_1, X_2, \ldots, X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Here $\Theta = \{(\mu, \sigma^2), -\infty < \mu < \infty, \sigma^2 > 0\}$. The likelihood function is

$$L(\mu, \sigma^2;\ x_1, \ldots, x_n) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left[-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right],$$

and

$$\log L(\mu, \sigma^2;\ \mathbf{x}) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi) - \frac{\sum_1^n (x_i - \mu)^2}{2\sigma^2}.$$

The likelihood equations are

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

and

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

Solving the first of these equations for μ, we get $\hat{\mu} = \overline{X}$ and, substituting in the second, $\hat{\sigma}^2 = \sum_{i=1}^{n}[(X_i - \overline{X})^2/n]$. We see that $(\hat{\mu}, \hat{\sigma}^2) \in \Theta$ with probability 1. We show that $(\hat{\mu}, \hat{\sigma}^2)$ maximizes the likelihood function. First note that $\overline{X}$ maximizes $L(\mu, \sigma^2;\ \mathbf{x})$ whatever σ^2 is, since $L(\mu, \sigma^2;\ \mathbf{x}) \to 0$ as $|\mu| \to \infty$, and in that case $L(\hat{\mu}, \sigma^2;\ \mathbf{x}) \to 0$ as $\sigma^2 \to 0$ or ∞ whenever $\boldsymbol{\theta} \in \Theta$, $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}^2)$.

Note that $\hat{\sigma}^2$ is not unbiased for σ^2. Indeed, $E\hat{\sigma}^2 = [(n-1)/n]\sigma^2$. But $n\hat{\sigma}^2/(n-1) = S^2$ is unbiased, as we already know. Also, $\hat{\mu}$ is unbiased, and both $\hat{\mu}$ and $\hat{\sigma}^2$ are consistent. In addition, $\hat{\mu}$ and $\hat{\sigma}^2$ are method of moments estimators for μ and σ^2, and $(\hat{\mu}, \hat{\sigma}^2)$ is jointly sufficient.

Finally, note that $\hat{\mu}$ is the MLE of μ if σ^2 is known; but if μ is known, the MLE of σ^2 is not $\hat{\sigma}^2$ but $\sum_1^n (X_i - \mu)^2/n$.

Example 3. Let $X_1, X_2, \ldots, X_n$ be a sample from PMF

$$P_N(k) = \begin{cases} \dfrac{1}{N}, & k = 1, 2, \ldots, N, \\ 0 & \text{otherwise.} \end{cases}$$

The likelihood function is

$$L(N;\ k_1, k_2, \dots, k_n) = \begin{cases} \dfrac{1}{N^n}, & 1 \le \max(k_1, \dots, k_n) \le N, \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, the MLE of N is given by

$$\hat{N}(X_1, X_2, \dots, X_n) = \max(X_1, X_2, \dots, X_n),$$

for if we take any $\hat{\alpha} < \hat{N}$ as the MLE, then $P_{\hat{\alpha}}(k_1, k_2, \dots, k_n) = 0$; and if we take any $\hat{\beta} > \hat{N}$ as the MLE, then $P_{\hat{\beta}}(k_1, k_2, \dots, k_n) = 1/(\hat{\beta})^n < 1/(\hat{N})^n = P_{\hat{N}}(k_1, k_2, \dots, k_n)$.

We see that the MLE $\hat{N}$ is consistent, sufficient, and complete, but not unbiased.

Example 4. Consider the hypergeometric PMF

$$P_N(x) = \begin{cases} \dfrac{\dbinom{M}{x}\dbinom{N-M}{n-x}}{\dbinom{N}{n}}, & \max(0, n - N + M) \le x \le \min(n, M), \\ 0, & \text{otherwise.} \end{cases}$$

To find the MLE $\hat{N} = \hat{N}(X)$ of N, consider the ratio

$$R(N) = \frac{P_N(x)}{P_{N-1}(x)} = \frac{N-n}{N} \frac{N-M}{N-M-n+x}.$$

For values of N for which $R(N) > 1$, $P_N(x)$ increases with N, and for values of N for which $R(N) < 1$, $P_N(x)$ is a decreasing function of N:

$$R(N) > 1 \qquad \text{if and only if } N < \frac{nM}{x}$$

and

$$R(N) < 1 \qquad \text{if and only if } N > \frac{nM}{x}.$$

It follows that $P_N(x)$ reaches its maximum value where $N \approx nM/x$. Thus $\hat{N}(X) = [nM/X]$, where $[x]$ denotes the largest integer $\le x$.

Example 5. Let $X_1, X_2, \dots, X_n$ be a sample from $U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. The likelihood function is

$$L(\theta;\ x_1, x_2, \ldots, x_n) = \begin{cases} 1 & \text{if } \theta - \frac{1}{2} \le \min(x_1, \ldots, x_n) \\ & \qquad \le \max(x_1, \ldots, x_n) \le \theta + \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Thus $L(\theta;\ \mathbf{x})$ attains its maximum provided that

$$\theta - \tfrac{1}{2} \le \min(x_1, \ldots, x_n) \quad \text{and} \quad \theta + \tfrac{1}{2} \ge \max(x_1, \ldots, x_n),$$

or when

$$\theta < \min(x_1, \ldots, x_n) + \tfrac{1}{2} \quad \text{and} \quad \theta \ge \max(x_1, \ldots, x_n) - \tfrac{1}{2}.$$

It follows that every statistic $T(X_1, X_2, \ldots, X_n)$ such that

$$\text{(6)} \qquad \max_{1 \le i \le n} X_i - \tfrac{1}{2} \le T(X_1, X_2, \ldots, X_n) \le \min_{1 \le i \le n} X_i + \tfrac{1}{2}$$

is an MLE of θ. Indeed, for $0 < \alpha < 1$,

$$T_\alpha(X_1, \ldots, X_n) = \max_{1 \le i \le n} X_i - \tfrac{1}{2} + \alpha(1 + \min_{1 \le i \le n} X_i - \max_{1 \le i \le n} X_i)$$

lies in interval (6), and hence for each α, $0 < \alpha < 1$, $T_\alpha(X_1, \ldots, X_n)$ is an MLE of θ. In particular, if $\alpha = \frac{1}{2}$,

$$T_{1/2}(X_1, \ldots, X_n) = \frac{\min X_i + \max X_i}{2}$$

is an MLE of θ.

Example 6. Let $X \sim b(1, p)$, $p \in [\frac{1}{4}, \frac{3}{4}]$. In this case $L(p;\ x) = p^x(1-p)^{1-x}$, $x = 0, 1$, and we cannot differentiate $L(p;\ x)$ to get the MLE of p, since that would lead to $\hat{p} = x$, a value that does not lie in $\Theta = [\frac{1}{4}, \frac{3}{4}]$. We have

$$L(p;\ x) = \begin{cases} p, & x = 1, \\ 1 - p, & x = 0, \end{cases}$$

which is maximized if we choose $\hat{p}(x) = \frac{1}{4}$ if $x = 0$, and $= \frac{3}{4}$ if $x = 1$. Thus the MLE of p is given by

$$\hat{p}(X) = \frac{2X + 1}{4}.$$

Note that $E_p\hat{p}(X) = (2p + 1)/4$, so that $\hat{p}$ is biased. Also, the mean square error for $\hat{p}$ is

$$E_p(\hat{p}(X) - p)^2 = \tfrac{1}{16} E_p(2X + 1 - 4p)^2 = \tfrac{1}{16}.$$

In the sense of the MSE, the MLE is worse than the trivial estimator $\delta(X) = \frac{1}{2}$, for $E_p(\frac{1}{2} - p)^2 = (\frac{1}{2} - p)^2 \le \frac{1}{16}$ for $p \in [\frac{1}{4}, \frac{3}{4}]$.

***Example** 7.* Let $X_1, X_2, \dots, X_n$ be iid $b(1, p)$ RVs, and suppose that $p \in (0, 1)$. If $(0, 0, \dots, 0)((1, 1, \dots, 1))$ is observed, $\overline{X} = 0(\overline{X} = 1)$ is the MLE, which is not an admissible value of p. Hence an MLE does not exist.

***Example** 8* (Oliver [76]). This example illustrates a distribution for which an MLE is necessarily an actual observation, but not necessarily any particular observation. Let $X_1, X_2, \dots, X_n$ be a sample from the PDF

$$f_\theta(x) = \begin{cases} \dfrac{2}{\alpha}\dfrac{x}{\theta}, & 0 \le x \le \theta, \\ \dfrac{2}{\alpha}\dfrac{\alpha - x}{\alpha - \theta}, & \theta \le x \le \alpha, \\ 0, & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ is a (known) constant. The likelihood function is

$$L(\theta;\, x_1, x_2, \dots, x_n) = \left(\frac{2}{\alpha}\right)^n \prod_{x_i \le \theta} \frac{x_i}{\theta} \prod_{x_i > \theta} \frac{\alpha - x_i}{\alpha - \theta},$$

where we have assumed that observations are arranged in increasing order of magnitude, $0 \le x_1 < x_2 < \cdots < x_n \le \alpha$. Clearly, L is continuous in θ (even for $\theta =$ some x_i) and differentiable for values of θ between any two x_i's. Thus, for $x_j < \theta < x_{j+1}$, we have

$$L(\theta) = \left(\frac{2}{\alpha}\right)^n \theta^{-j}(\alpha - \theta)^{-(n-j)} \prod_{i=1}^{j} x_i \prod_{i=j+1}^{n} (\alpha - x_i),$$

$$\frac{\partial \log L}{\partial \theta} = -\frac{j}{\theta} + \frac{n-j}{\alpha - \theta} \quad \text{and} \quad \frac{\partial^2 \log L}{\partial \theta^2} = \frac{j}{\theta^2} + \frac{n-j}{(\alpha - \theta)^2} > 0.$$

It follows that any stationary value that exists must be a minimum, so that there can be no maximum in any range $x_j < \theta < x_{j+1}$. Moreover, there can be no maximum in $0 \le \theta < x_1$ or $x_n < \theta \le \alpha$. This follows since for $0 \le \theta < x_1$,

$$L(\theta) = \left(\frac{2}{\alpha}\right)^n (\alpha - \theta)^{-n} \prod_{i=1}^{n} (\alpha - x_i)$$

is a strictly increasing function of θ. By symmetry, $L(\theta)$ is a strictly decreasing function of θ in $x_n < \theta \le \alpha$. We conclude that an MLE has to be one of the observations.

In particular, let $\alpha = 5$ and $n = 3$, and suppose that the observations, arranged in increasing order of magnitude, are $1, 2, 4$. In this case the MLE can be shown to be

$\hat{\theta} = 1$, which corresponds to the first-order statistic. If the sample values are 2, 3, 4, the third-order statistic is the MLE.

Example 9. Let $X_1, X_2, \ldots, X_n$ be a sample from $G(r, 1/\beta)$; $\beta > 0$ and $r > 0$ are both unknown. The likelihood function is

$$L(\beta, r;\ x_1, x_2, \ldots, x_n) = \begin{cases} \dfrac{\beta^{nr}}{[\Gamma(r)]^n} \prod_{i=1}^n x_i^{r-1} \exp\left(-\beta \sum_{i=1}^n x_i\right), & x_i \ge 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\log L(\beta, r) = nr \log \beta - n \log \Gamma(r) + (r-1) \sum_{i=1}^n \log x_i - \beta \sum_{i=1}^n x_i,$$

$$\frac{\partial \log L(\beta, r)}{\partial \beta} = \frac{nr}{\beta} - \sum_{i=1}^n x_i = 0,$$

and

$$\frac{\partial \log L(\beta, r)}{\partial r} = n \log \beta - n \frac{\Gamma'(r)}{\Gamma(r)} + \sum_{i=1}^n \log x_i = 0.$$

The first of the likelihood equations yields $\hat{\beta}(x_1, x_2, \ldots, x_n) = \hat{r}/\bar{x}$, while the second gives

$$n \log \frac{r}{\bar{x}} + \sum_{i=1}^n \log x_i - n \frac{\Gamma'(r)}{\Gamma(r)} = 0,$$

that is,

$$\log r - \frac{\Gamma'(r)}{\Gamma(r)} = \log \bar{x} - \frac{1}{n} \sum_{i=1}^n \log x_i,$$

which is to be solved for $\hat{r}$. In this case, the likelihood equation is not easily solvable and it is necessary to resort to numerical methods, using tables for $\Gamma'(r)/\Gamma(r)$.

Remark 2. We have seen that MLEs may not be unique, although frequently they are. Also, they are not necessarily unbiased even if a unique MLE exists. In terms of MSE, an MLE may be worthless. Moreover, MLEs may not even exist. We have also seen that MLEs are functions of sufficient statistics. This is a general result, which we now prove.

Theorem 1. Let T be a sufficient statistic for the family of PDFs (PMFs) $\{f_\theta : \theta \in \Theta\}$. If a unique MLE of θ exists, it is a (nonconstant) function of T. If a MLE of θ exists but is not unique, one can find a MLE that is a function of T.

Proof. Since T is sufficient, we can write

$$L(\theta) = f_\theta(\mathbf{x}) = h(\mathbf{x})g_\theta(T(\mathbf{x})),$$

for all $\mathbf{x}$, all θ, and some h and g_θ. If a unique MLE $\hat{\theta}$ exists that maximizes $L(\theta)$, it also maximizes $g_\theta(T(\mathbf{x}))$ and hence $\hat{\theta}$ is a function of T. If a MLE of θ exists but is not unique, we choose a particular MLE $\hat{\theta}$ from the set of all MLEs which is a function of T.

Example 10. Let $X_1, X_2, \ldots, X_n$ be a random sample from $U[\theta, \theta+1]$, $\theta \in R$. Then the likelihood function is given by

$$L(\theta; \mathbf{x}) = \left(\frac{1}{2}\right)^n I_{[\theta-1\le x_{(1)}\le x_{(n)}\le\theta+1]}(\mathbf{x}).$$

We note that $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is jointly sufficient for θ and any θ satisfying

$$\theta - 1 \le x_{(1)} \le x_{(n)} \le \theta + 1,$$

or, equivalently,

$$x_{(n)} - 1 \le \theta \le x_{(1)} + 1$$

maximizes the likelihood and hence is an MLE for θ. Thus, for $0 \le \alpha \le 1$,

$$\hat{\theta}_\alpha = \alpha(X_{(n)} - 1) + (1 - \alpha)(X_{(1)} + 1)$$

is an MLE of θ. If α is a constant independent of the X's, then $\hat{\theta}_\alpha$ is a function of T. If, on the other hand, α depends on the X's, then $\hat{\theta}_\alpha$ may not be a function of T alone. For example,

$$\hat{\theta}_\alpha = (\sin^2 X_1)(X_{(n)} - 1) + (\cos^2 X_1)(X_{(1)} + 1)$$

is an MLE of θ but not a function of T alone.

Theorem 2. Suppose that the regularity conditions of the FCR inequality are satisfied and θ belongs to an open interval on the real line. If an estimator $\hat{\theta}$ of θ attains the FCR lower bound for the variance, the likelihood equation has a unique solution $\hat{\theta}$ that maximizes the likelihood.

Proof. If $\hat{\theta}$ attains the FCR lower bound, we have [see (8.5.8)]

$$\frac{\partial \log f_\theta(\mathbf{X})}{\partial \theta} = [k(\theta)]^{-1}[\hat{\theta}(\mathbf{X}) - \theta]$$

with probability 1, and the likelihood equation has a unique solution $\theta = \hat{\theta}$.

Let us write $A(\theta) = [k(\theta)]^{-1}$. Then

$$\frac{\partial^2 \log f_\theta(\mathbf{X})}{\partial \theta^2} = A'(\theta)(\hat{\theta} - \theta) - A(\theta),$$

so that

$$\left.\frac{\partial^2 \log f_\theta(\mathbf{X})}{\partial \theta^2}\right|_{\theta=\hat{\theta}} = -A(\theta).$$

We need only to show that $A(\theta) > 0$.

Recall from (8.5.4) with $\psi(\theta) = \theta$ that

$$E_\theta\left\{[T(\mathbf{X}) - \theta]\,\frac{\partial \log f_\theta(\mathbf{X})}{\partial\theta}\right\} = 1,$$

and substituting $T(\mathbf{X}) - \theta = k(\theta)[\partial \log f_\theta(\mathbf{X})/\partial\theta]$, we get

$$k(\theta)E_\theta\left[\frac{\partial \log f_\theta(\mathbf{X})}{\partial\theta}\right]^2 = 1.$$

That is,

$$A(\theta) = E\left[\frac{\partial \log f_\theta(\mathbf{X})}{\partial\theta}\right]^2 > 0$$

and the proof is complete.

Remark 3. In Theorem 2 we assumed the differentiability of $A(\theta)$ and the existence of the second-order partial derivative $\partial^2 \log f_\theta/\partial \theta^2$. If the conditions of Theorem 2 are satisfied, the most efficient estimator is necessarily the MLE. It does not follow, however, that every MLE is most efficient. For example, in sampling from a normal population, $\hat{\sigma}^2 = \sum_1^n (X_i - \overline{X})^2/n$ is the MLE of σ^2, but it is not most efficient. Since $\sum(X_i - \overline{X})^2/\sigma^2$ is $\chi^2(n-1)$, we see that $\text{var}(\hat{\sigma}^2) = 2(n-1)\sigma^4/n^2$, which is not equal to the FCR lower bound, $2\sigma^4/n$. Note that $\hat{\sigma}^2$ is not even an unbiased estimator of σ^2.

We next consider an important property of MLEs that is not shared by other methods of estimation. Often the parameter of interest is not θ but some function $h(\theta)$. If $\hat{\theta}$ is the MLE of θ, what is the MLE of $h(\theta)$? If $\lambda = h(\theta)$ is a one-to-one function of θ, the inverse function $h^{-1}(\lambda) = \theta$ is well defined and we can write the likelihood function as a function of λ. We have

$$L^*(\lambda; \mathbf{x}) = L(h^{-1}(\lambda); \mathbf{x})$$

so that

$$\sup_{\lambda} L^*(\lambda; \mathbf{x}) = \sup_{\lambda} L(h^{-1}(\lambda); \mathbf{x}) = \sup_{\theta} L(\theta; \mathbf{x}).$$

It follows that the supremum of L^* is achieved at $\lambda = h(\hat{\theta})$. Thus $h(\hat{\theta})$ is the MLE of $h(\theta)$.

In many applications $\lambda = h(\theta)$ is not one-to-one. It is still tempting to take $\hat{\lambda} = h(\hat{\theta})$ as the MLE of λ. The following result provides a justification.

Theorem 3 (Zehna [121]). Let $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ be a family of PDFs (PMFs), and let $L(\boldsymbol{\theta})$ be the likelihood function. Suppose that $\Theta \subseteq \mathcal{R}_k$, $k \geq 1$. Let $h : \Theta \to \Lambda$ be a mapping of Θ onto Λ, where Λ is an interval in $\mathcal{R}_p (1 \leq p \leq k)$. If $\hat{\boldsymbol{\theta}}$ is an MLE of $\boldsymbol{\theta}$, then $h(\hat{\boldsymbol{\theta}})$ is an MLE of $h(\boldsymbol{\theta})$.

Proof. For each $\lambda \in \Lambda$, let us define

$$\Theta_{\lambda} = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta,\ h(\boldsymbol{\theta}) = \lambda\}$$

and

$$M(\lambda;\ \mathbf{x}) = \sup_{\boldsymbol{\theta} \in \Theta_{\lambda}} L(\boldsymbol{\theta}; \mathbf{x}).$$

Then M defined on Λ is called the likelihood function induced by h. If $\hat{\boldsymbol{\theta}}$ is any MLE of $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}}$ belongs to one and only one set, $\Theta_{\hat{\lambda}}$ say. Since $\hat{\boldsymbol{\theta}} \in \Theta_{\hat{\lambda}}$, $\hat{\lambda} = h(\hat{\boldsymbol{\theta}})$. Now

$$M(\hat{\lambda};\ \mathbf{x}) = \sup_{\boldsymbol{\theta} \in \Theta_{\lambda}} L(\boldsymbol{\theta};\ \mathbf{x}) \geq L(\hat{\boldsymbol{\theta}};\ \mathbf{x})$$

and $\hat{\lambda}$ maximizes M, since

$$M(\hat{\lambda};\ \mathbf{x}) \leq \sup_{\lambda \in \Lambda} M(\lambda;\ \mathbf{x}) = \sup_{\boldsymbol{\theta} \in \Theta_{\lambda}} L(\boldsymbol{\theta};\ \mathbf{x}) = L(\hat{\boldsymbol{\theta}}; \mathbf{x}),$$

so that $M(\hat{\lambda};\ \mathbf{x}) = \sup_{\lambda \in \Lambda} M(\lambda;\ \mathbf{x})$. It follows that $\hat{\lambda}$ is an MLE of $h(\boldsymbol{\theta})$, where $\hat{\lambda} = h(\hat{\boldsymbol{\theta}})$.

Example 11. Let $X \sim b(1, p)$, $0 \leq p \leq 1$, and let $h(p) = \text{var}(X) = p(1 - p)$. We wish to find the MLE of $h(p)$. Note that $\Lambda = [0, \frac{1}{4}]$. The function h is not one-to-one. The MLE of p based on a sample of size n is $\hat{p}(X_1, \ldots, X_n) = \overline{X}$. Hence the MLE of parameter $h(p)$ is $h(\overline{X}) = \overline{X}(1 - \overline{X})$.

Example 12. Consider a random sample from $G(1, \beta)$. It is required to find the MLE of β in the following manner. A sample of size n is taken, and it is known only that k, $0 \leq k \leq n$, of these observations are $\leq M$, where M is a fixed positive number.

Let $p = P\{X_i \le M\} = 1 - e^{-M/\beta}$, so that $-M/\beta = \log(1-p)$ and $\beta = M/\log[1/(1-p)]$. Therefore, the MLE of β is $M/\log[1/(1-\hat{p})]$, where $\hat{p}$ is the MLE of p. To compute the MLE of p we have

$$L(p;\, x_1, x_2, \dots, x_n) = p^k(1-p)^{n-k},$$

so that the MLE of p is $\hat{p} = k/n$. Thus the MLE of β is

$$\hat{\beta} = \frac{M}{\log[n/(n-k)]}.$$

Finally, we consider some important large-sample properties of MLEs. In the following we assume that $\{f_\theta,\ \theta \in \Theta\}$ is a family of PDFs (PMFs), where Θ is an open interval on $\mathcal{R}$. The conditions listed below are stated when f_θ is a PDF. Modifications for the case where f_θ is a PMF are obvious and will be left to the reader.

(i) $\partial \log f_\theta/\partial\theta$, $\partial^2 \log f_\theta/\partial\theta^2$, $\partial^3 \log f_\theta/\partial\theta^3$ exist for all $\theta \in \Theta$ and every x. Also,

$$\int_{-\infty}^{\infty} \frac{\partial f_\theta(x)}{\partial\theta}\,dx = E_\theta \frac{\partial \log f_\theta(X)}{\partial\theta} = 0 \qquad \text{for all } \theta \in \Theta.$$

(ii) $\int_{-\infty}^{\infty} \dfrac{\partial^2 f_\theta(x)}{\partial\theta^2}\,dx = 0 \qquad$ for all $\theta \in \Theta$.

(iii) $-\infty < \int_{-\infty}^{\infty} \dfrac{\partial^2 \log f_\theta(x)}{\partial\theta^2} f_\theta(x)\,dx < 0 \qquad$ for all θ.

(iv) There exists a function $H(x)$ such that for all $\theta \in \Theta$,

$$\left|\frac{\partial^3 \log f_\theta(x)}{\partial\theta^3}\right| < H(x) \quad \text{and} \quad \int_{-\infty}^{\infty} H(x) f_\theta(x)\,dx = M(\theta) < \infty.$$

(v) There exists a function $g(\theta)$ that is positive and twice differentiable for every $\theta \in \Theta$ and, a function $H(x)$ such that for all θ

$$\left|\frac{\partial^2}{\partial\theta^2}\left[g(\theta)\frac{\partial \log f_\theta}{\partial\theta}\right]\right| < H(x) \quad \text{and} \quad \int_{-\infty}^{\infty} H(x) f_\theta(x)\,dx < \infty.$$

Note that the condition (v) is equivalent to condition (iv) with the added qualification that $g(\theta) = 1$.

We state the following results without proof.

Theorem 4 (Cramér [16])

(a) Conditions (i), (iii), and (iv) imply that with probability approaching 1, as $n \to \infty$, the likelihood equation has a consistent solution.

(b) Conditions (i) through (iv) imply that a consistent solution $\hat{\theta}_n$ of the likelihood equation is asymptotically normal, that is,

$$\sigma^{-1}\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{L} Z$$

where Z is $\mathcal{N}(0, 1)$, and

$$\sigma^2 = \left\{ E_\theta \left[\frac{\partial \log f_\theta(X)}{\partial \theta} \right]^2 \right\}^{-1}.$$

On occasions one encounters examples where the conditions of Theorem 4 are not satisfied and yet a solution of the likelihood equation is consistent and asymptotically normal.

Example 13 (Kulldorf [55]). Let $X \sim \mathcal{N}(0, \theta)$, $\theta > 0$. Let $X_1, X_2, \ldots, X_n$ be n independent observations on X. The solution of the likelihood equation is $\hat{\theta}_n = \sum_{i=1}^n X_i^2/n$. Also, $EX^2 = \theta$, $\text{var}(X^2) = 2\theta^2$, and

$$E_\theta \left[\frac{\partial \log f_\theta(X)}{\partial \theta} \right]^2 = \frac{1}{2\theta^2}.$$

We note that

$$\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta$$

and

$$\sqrt{n}\,(\hat{\theta}_n - \theta) = \theta\sqrt{2}\,\frac{\sum_1^n X_i^2 - n\theta}{\sqrt{2n}\,\theta} \xrightarrow{L} \mathcal{N}(0, 2\theta^2).$$

However,

$$\frac{\partial^3 \log f_\theta}{\partial^3 \theta} = -\frac{1}{\theta^3} + \frac{3x^2}{\theta^4} \to \infty \qquad \text{as } \theta \to 0$$

and is not bounded in $0 < \theta < \infty$. Thus condition (iv) does not hold.

The following theorem covers such cases also.

Theorem 5 (Kulldorf [55])

(a) Conditions (i), (iii), and (v) imply that with probability approaching 1 as $n \to \infty$, the likelihood equation has a solution.

(b) Conditions (i), (ii), (iii), and (v) imply that a consistent solution of the likelihood equation is asymptotically normal.

For proofs of Theorems 4 and 5 we refer to Cramér [16, p. 500], and Kulldorf [55].

Remark 4. It is important to note that the results in Theorems 4 and 5 establish the consistency of some root of the likelihood equation but not necessarily that of the MLE when the likelihood equation has several roots. Huzurbazar [44] has shown that under certain conditions the likelihood equation has at most one consistent solution and that the likelihood function has a relative maximum for such a solution. Since there may be several solutions for which the likelihood function has relative maxima, Cramér's and Huzurbazar's results still do not imply that a solution of the likelihood equation that makes the likelihood function an absolute maximum is necessarily consistent.

Wald [114] has shown that under certain conditions the MLE is strongly consistent. It is important to note that Wald does not make any differentiability assumptions.

In any event, if the MLE is a unique solution of the likelihood equation, we can use Theorems 4 and 5 to conclude that it is consistent and asymptotically normal. Note that the asymptotic variance is the same as the lower bound of the FCR inequality.

Example 14. Consider $X_1, X_2, \dots, X_n$ iid $P(\lambda)$ RVs, $\lambda \in \Theta = (0, \infty)$. The likelihood equation has a unique solution, $\hat{\lambda}(x_1, \dots, x_n) = \overline{X}$, which maximizes the likelihood function. We leave the reader to check that the conditions of Theorem 4 hold and that MLE $\overline{X}$ is consistent and asymptotically normal with mean λ and variance λ/n, a result that is immediate otherwise.

We leave the reader to check that in Example 13, conditions of Theorem 5 are satisfied.

Remark 5. The invariance and the large-sample properties of MLEs permit us to find MLEs of parametric functions and their limiting distributions. The delta method introduced in Section 7.5 (Theorem 1) comes in handy in these applications. Suppose that in Example 13 we wish to estimate $\psi(\theta) = \theta^2$. By invariance of MLEs, the MLE of $\psi(\theta)$ is $\psi(\hat{\theta}_n)$ where $\hat{\theta}_n = \sum_1^n X_i^2/n$ is the MLE of θ. Applying Theorem 7.5.1, we see that $\psi(\hat{\theta}_n)$ is $AN(\theta^2, 8\theta^4/n)$.

In Example 14, suppose that we wish to estimate $\psi(\lambda) = P_\lambda(X = 0) = e^{-\lambda}$. Then $\psi(\hat{\lambda}) = e^{-\overline{X}}$ is the MLE of $\psi(\lambda)$ and, in view of Theorem 7.5.1, $\psi(\hat{\lambda}) \sim AN(e^{-\lambda}, \lambda e^{-2\lambda}/n)$.

Remark 6. The uniqueness of MLE does not guarantee its asymptotic normality. Consider, for example, a random sample from $U(0, \theta]$. Then $X_{(n)}$ is the unique MLE for θ, and in Problem 8.2.5 we asked the reader to show that $n(\theta - X_{(n)}) \xrightarrow{L} G(1, \theta)$.

PROBLEMS 8.7

1. Let $X_1, X_2, \ldots, X_n$ be iid RVs with common PMF (PDF) $f_\theta(x)$. Find an MLE for θ in each of the following cases:
   (a) $f_\theta(x) = \frac{1}{2}e^{-|x-\theta|}$, $-\infty < x < \infty$.
   (b) $f_\theta(x) = e^{-x+\theta}$, $\theta \le x < \infty$.
   (c) $f_\theta(x) = (\theta\alpha)x^{\alpha-1}e^{-\theta x^\alpha}$, $x > 0$, and α known.
   (d) $f_\theta(x) = \theta(1-x)^{\theta-1}$, $0 \le x \le 1$, $\theta > 1$.

2. Find an MLE, if it exists, in each of the following cases:
   (a) $X \sim b(n, \theta)$: both n and $\theta \in [0, 1]$ are unknown, and one observation is available.
   (b) $X_1, X_2, \ldots, X_n \sim b(1, \theta)$, $\theta \in [\frac{1}{2}, \frac{3}{4}]$.
   (c) $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\theta, \theta^2)$, $\theta \in \mathcal{R}$.
   (d) $X_1, X_2, \ldots, X_n$ is a sample from

$$P\{X = y_1\} = \frac{1-\theta}{2}, \qquad P\{X = y_2\} = \frac{1}{2}, \qquad P\{X = y_3\} = \frac{\theta}{2}(0 < \theta < 1).$$

   (e) $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\theta, \theta)$, $0 < \theta < \infty$.
   (f) $X \sim \mathcal{C}(\theta, 0)$.

3. Suppose that n observations are taken on an RV X with distribution $\mathcal{N}(\mu, 1)$, but instead of recording all the observations, one notes only whether or not the observation is less than 0. If $\{X < 0\}$ occurs $m(< n)$ times, find the MLE of μ.

4. Let $X_1, X_2, \ldots, X_n$ be a random sample from the PDF

$$f(x;\ \alpha, \beta) = \beta^{-1}e^{-\beta^{-1}(x-\alpha)}, \qquad \alpha < x < \infty, \quad -\infty < \alpha < \infty, \quad \beta > 0.$$

   (a) Find the MLE of (α, β).
   (b) Find the MLE of $P_{\alpha,\beta}\{X_1 \ge 1\}$.

5. Let $X_1, X_2, \ldots, X_n$ be a sample from exponential density $f_\theta(x) = \theta e^{-\theta x}$, $x \ge 0$, $\theta > 0$. Find the MLE of θ, and show that it is consistent and asymptotically normal.

6. For Problem 8.6.5 find the MLE for (μ, σ^2).

7. For a sample of size 1 taken from $\mathcal{N}(\mu, \sigma^2)$, show that no MLE of (μ, σ^2) exists.

8. For Problem 5.2.5 suppose that we wish to estimate N on the basis of observations $X_1, X_2, \ldots, X_M$.
   (a) Find the UMVUE of N.
   (b) Find the MLE of N.
   (c) Compare the MSEs of the UMVUE and the MLE.

9. Let $X_{ij}(1 = 1, 2, \ldots, s;\ j = 1, 2, \ldots, n)$ be independent RVs where $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$, $i = 1, 2, \ldots, s$. Find MLEs for $\mu_1, \mu_2, \ldots, \mu_s$, and σ^2. Show that the MLE for σ^2 is not consistent as $s \to \infty$ (n fixed). (Neyman and Scott [75])

10. Let (X, Y) have a bivariate normal distribution with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and ρ. Suppose that n observations are made on the pair (X, Y), and $N - n$ observations on X; that is, $N - n$ observations on Y are missing. Find the MLEs of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and ρ. [*Hint:* If $f(x, y;\ \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ is the joint PDF of (X, Y), write

$$f(x, y;\ \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) = f_1(x;\ \mu_1, \sigma_1^2) f_{Y|X}(y \mid \beta_x, \sigma_2^2(1 - \rho^2)),$$

where f_1 is the marginal (normal) PDF of X, and $f_{Y|X}$ is the conditional (normal) PDF of Y, given x with mean

$$\beta_x = \left(\mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1\right) + \rho\frac{\sigma_2}{\sigma_1}x$$

and variance $\sigma_2^2(1 - \rho^2)$. Maximize the likelihood function first with respect to μ_1 and σ_1^2 and then with respect to $\mu_2 - \rho(\sigma_2/\sigma_1)\mu_1$, $\rho\sigma_2/\sigma_1$, and $\sigma_2^2(1 - \rho^2)$.] (Anderson [1])

11. In Problem 5, let $\hat{\theta}$ denote the MLE of θ. Find the MLE of $\mu = EX_1 = 1/\theta$ and its asymptotic distribution.

12. In Problem 1(d), find the asymptotic distribution of the MLE of θ.

13. In Problem 2(a), find the MLE of $d(\theta) = \theta^2$ and its asymptotic distribution.

14. Let $X_1, X_2, \ldots, X_n$ be a random sample from some DF F on the real line. Suppose that we observe $x_1, x_2, \ldots, x_n$ which are all different. Show that the MLE of F is F_n^*, the empirical DF of the sample.

15. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\mu, 1)$. Suppose that $\Theta = \{\mu \geq 0\}$. Find the MLE of μ.

16. Let $(X_1, X_2, \ldots, X_{k-1})$ have a multinomial distribution with parameters $n, p_1, \ldots, p_{k-1}, 0 \leq p_1, p_2, \ldots, p_{k-1} \leq 1, \sum_1^{k-1} p_j \leq 1$, where n is known. Find the MLE of $(p_1, p_2, \ldots, p_{k-1})$.

17. Consider the one-parameter exponential density introduced in Section 5.5 in its natural form with the PDF

$$f_\theta(x) = \exp[\eta T(x) + D(\eta) + S(x)].$$

(a) Show that the MGF of $T(X)$ is given by

$$M(t) = \exp[D(\eta) - D(\eta + t)]$$

for t in some neighborhood of the origin. Moreover, $E_\eta T(X) = -D'(\eta)$, and $\text{var}(T(X)) = -D''(\eta)$.

(b) If the equation $E_\eta T(X) = T(x)$ has a solution, it must be the unique MLE of η.

18. In Problem 1(b), show that the unique MLE of θ is consistent. Is it asymptotically normal?

8.8 BAYES AND MINIMAX ESTIMATION

In this section we consider the problem of point estimation in a decision-theoretic setting. We consider here Bayes and minimax estimation.

Let $\{f_\theta : \theta \in \Theta\}$ be a family of PDFs (PMFs), and $X_1, X_2, \ldots, X_n$ be a sample from this distribution. Once the sample point $(x_1, x_2, \ldots, x_n)$ is observed, the statistician takes an *action* on the basis of these data. Let us denote by $\mathcal{A}$ the set of all *actions* or *decisions* open to the statistician.

Definition 1. A *decision function* δ is a statistic that takes values in $\mathcal{A}$; that is, δ is a Borel-measurable function that maps $\mathcal{R}_n$ into $\mathcal{A}$.

If $\mathbf{X} = \mathbf{x}$ is observed, the statistician takes action $\delta(\mathbf{X}) \in \mathcal{A}$.

Example 1. Let $\mathcal{A} = \{a_1, a_2\}$. Then any decision function δ partitions the space of values of $(X_1, \ldots, X_n)$, namely, $\mathcal{R}_n$, into a set C and its complement C^c, such that if $\mathbf{x} \in C$, we take action a_1, and if $\mathbf{x} \in C^c$ action a_2 is taken. This is the problem of testing hypotheses, which we discuss in Chapter 9.

Example 2. Let $\mathcal{A} = \Theta$. In this case we face the problem of estimation.

Another element of decision theory is the specification of a *loss function*, which measures the loss incurred when we take a decision.

Definition 2. Let $\mathcal{A}$ be an arbitrary space of actions. A nonnegative function L that maps $\Theta \times \mathcal{A}$ into $\mathcal{R}$ is called a *loss function*.

The value $L(\theta, a)$ is the loss to the statistician if he takes action a when θ is the true parameter value. If we use the decision function $\delta(\mathbf{X})$ and loss function L and θ is the true parameter value, the loss is the RV $L(\theta, \delta(\mathbf{X}))$. (As always, we will assume that L is a Borel-measurable function.)

Definition 3. Let $\mathcal{D}$ be a class of decision functions that map $\mathcal{R}_n$ into $\mathcal{A}$, and let L be a loss function on $\Theta \times \mathcal{A}$. The function R defined on $\Theta \times \mathcal{D}$ by

$$R(\theta, \delta) = E_\theta L(\theta, \delta(\mathbf{X})) \tag{1}$$

is known as the *risk function* associated with δ at θ.

Example 3. Let $\mathcal{A} = \Theta \subseteq \mathcal{R}$, $L(\theta, a) = |\theta - a|^2$. Then

$$R(\theta, \delta) = E_\theta L(\theta, \delta(X)) = E_\theta\{\delta(X) - \theta\}^2,$$

which is just the MSE. If we restrict attention to estimators that are unbiased, the risk is just the variance of the estimator.

The basic problem of decision theory is the following: Given a space of actions A, and a loss function $L(\theta, a)$, find a decision function δ in $\mathcal{D}$ such that the risk $R(\theta, \delta)$ is "minimum" in some sense for all $\theta \in \Theta$. We need first to specify some criterion for comparing the decision functions δ.

Definition 4. The principle of minimax is to choose $\delta^* \in \mathcal{D}$ so that

$$\max_\theta R(\theta, \delta^*) \le \max_\theta R(\theta, \delta) \tag{2}$$

for all δ in $\mathcal{D}$. Such a rule δ^*, if it exists, is called a *minimax (decision) rule.*

If the problem is one of estimation, that is, if $\mathcal{A} = \Theta$, we call δ^* satisfying (2) a *minimax estimator* of θ.

Example 4. Let $X \sim b(1, p)$, $p \in \Theta = \{\frac{1}{4}, \frac{1}{2}\}$ and $\mathcal{A} = \{a_1, a_2\}$. Let the loss function be defined as follows.

	a_1	a_2
$p_1 = \frac{1}{4}$	1	4
$p_2 = \frac{1}{2}$	3	2

The set of decision rules includes four functions: $\delta_1, \delta_2, \delta_3, \delta_4$, defined by $\delta_1(0) = \delta_1(1) = a_1$; $\delta_2(0) = a_1$, $\delta_2(1) = a_2$; $\delta_3(0) = a_2$, $\delta_3(1) = a_1$; and $\delta_4(0) = \delta_4(1) = a_2$. The risk function takes the following values:

i	$R(p_1, \delta_i)$	$R(p_2, \delta_i)$	$\underset{p_1, p_2}{\text{Max}}\, R(p, \delta_i)$	$\underset{i}{\text{Min}}\,\underset{p_1, p_2}{\text{Max}}\, R(p, \delta_i)$
1	1	3	3	
2	$\frac{7}{4}$	$\frac{5}{2}$	$\frac{5}{2}$	$\frac{5}{2}$
3	$\frac{13}{4}$	$\frac{5}{2}$	$\frac{13}{4}$	
4	4	2	4	

Thus the minimax solution is $\delta_2(x) = a_1$ if $x = 0$ and $= a_2$ if $x = 1$.

The computation of minimax estimators is facilitated by the use of the *Bayes estimation method.* So far, we have considered θ as a fixed constant and $f_\theta(\mathbf{x})$ has represented the PDF (PMF) of the RV $\mathbf{X}$. In Bayesian estimation we treat θ as a random variable distributed according to PDF (PMF) $\pi(\theta)$ on Θ. Also, π is called the *a priori distribution*. Now $f(\mathbf{x} \mid \theta)$ represents the conditional probability density (or mass) function of RV $\mathbf{X}$, given that $\theta \in \Theta$ is held fixed. Since π is the distribution of θ, it follows that the joint density (PMF) of θ and $\mathbf{X}$ is given by

$$f(\mathbf{x}, \theta) = \pi(\theta) f(\mathbf{x} \mid \theta). \tag{3}$$

In this framework $R(\theta, \delta)$ is the conditional average loss, $E\{L(\theta, \delta(\mathbf{X})) \mid \theta\}$, given that θ is held fixed. (Note that we are using the same symbol to denote the RV θ and a value assumed by it.)

Definition 5. The *Bayes risk* of a decision function δ is defined by

$$R(\pi, \delta) = E_\pi R(\theta, \delta). \tag{4}$$

If θ is a continuous RV and $\mathbf{X}$ is of the continuous type, then

$$\begin{aligned} R(\pi, \delta) &= \int R(\theta, \delta)\pi(\theta)\, d\theta \\ &= \iint L(\theta, \delta(\mathbf{x})) f(\mathbf{x} \mid \theta)\pi(\theta)\, d\mathbf{x}\, d\theta \\ &= \iint L(\theta, \delta(\mathbf{x})) f(\mathbf{x}, \theta)\, d\mathbf{x}\, d\theta. \end{aligned} \tag{5}$$

If θ is discrete with PMF π and $\mathbf{X}$ is of the discrete type, then

$$R(\pi, \delta) = \sum_\theta \sum_{\mathbf{x}} L(\theta, \delta(\mathbf{x})) f(\mathbf{x}, \theta). \tag{6}$$

Similar expressions may be written in the other two cases.

Definition 6. A decision function δ^* is known as a *Bayes rule* (*procedure*) if it minimizes the Bayes risk, that is, if

$$R(\pi, \delta^*) = \inf_\delta R(\pi, \delta). \tag{7}$$

Definition 7. The conditional distribution of RV θ, given $\mathbf{X} = \mathbf{x}$, is called the *a posteriori probability distribution* of θ, given the sample.

Let the joint PDF (PMF) be expressed in the form

$$f(\mathbf{x}, \theta) = g(\mathbf{x}) h(\theta \mid \mathbf{x}), \tag{8}$$

where g denotes the joint marginal density (PMF) of $\mathbf{X}$. The a priori PDF (PMF) $\pi(\theta)$ gives the distribution of θ before the sample is taken, and the a posteriori PDF

(PMF) $h(\theta \mid \mathbf{x})$ gives the distribution of θ after sampling. In terms of $h(\theta \mid \mathbf{x})$ we may write

$$R(\pi, \delta) = \int g(\mathbf{x}) \left[\int L(\theta, \delta(\mathbf{x})) h(\theta \mid \mathbf{x})\, d\theta \right] d\mathbf{x} \tag{9}$$

or

$$R(\pi, \delta) = \sum_{\mathbf{x}} g(\mathbf{x}) \left[\sum_{\theta} L(\theta, \delta(\mathbf{x})) h(\theta \mid \mathbf{x}) \right], \tag{10}$$

depending on whether f and π are both continuous or both discrete. Similar expressions may be written if only one of f and π is discrete.

Theorem 1. Consider the problem of estimation of a parameter $\theta \in \Theta \subseteq \mathcal{R}$ with respect to the quadratic loss function $L(\theta, \delta) = (\theta - \delta)^2$. A Bayes solution is given by

$$\delta(\mathbf{x}) = E\{\theta \mid \mathbf{X} = \mathbf{x}\}. \tag{11}$$

[$\delta(x)$ defined by (11) is called the *Bayes estimator*].

Proof. In the continuous case, if π is the prior PDF of θ, then

$$R(\pi, \delta) = \int g(\mathbf{x}) \left\{ \int [\theta - \delta(\mathbf{x})]^2 h(\theta \mid \mathbf{x})\, d\theta \right\} d\mathbf{x},$$

where g is the marginal PDF of $\mathbf{X}$, and h is the conditional PDF of θ, given $\mathbf{x}$. The Bayes rule is a function δ that minimizes $R(\pi, \delta)$. Minimization of $R(\pi, \delta)$ is the same as minimization of

$$\int [\theta - \delta(\mathbf{x})]^2 h(\theta \mid \mathbf{x})\, d\theta,$$

which is minimum if and only if

$$\delta(\mathbf{x}) = E\{\theta \mid \mathbf{x}\}.$$

The proof for the remaining cases is similar.

Remark 1. The argument used in Theorem 1 shows that a Bayes estimator is one that minimizes $E\{L(\theta, \delta(\mathbf{X})) \mid \mathbf{X}\}$. Theorem 1 is a special case which says that if $L(\theta, \delta(\mathbf{X})) = [\theta - \delta(\mathbf{X})]^2$, the function

$$\delta(\mathbf{x}) = \int \theta\, h(\theta \mid \mathbf{x})\, d\theta$$

is the Bayes estimator for θ with respect to π, the a priori distribution on Θ.

Remark 2. Suppose that $T(\mathbf{X})$ is sufficient for the parameter θ. Then it is easily seen that the posterior distribution of θ given $\mathbf{x}$ depends on $\mathbf{x}$ only through T and it follows that the Bayes estimator of θ is a function of T.

Example 5. Let $X \sim b(n, p)$ and $L(p, \delta(x)) = [p - \delta(x)]^2$. Let $\pi(p) = 1$ for $0 < p < 1$ be the a priori PDF of p. Then

$$h(p \mid x) = \frac{\binom{n}{x}p^x(1-p)^{n-x}}{\int_0^1 \binom{n}{x}p^x(1-p)^{n-x}dp}.$$

It follows that

$$\begin{aligned} E\{p \mid x\} &= \int_0^1 p\,h\{p \mid x\}\,dp \\ &= \frac{x+1}{n+2}. \end{aligned}$$

Hence the Bayes estimator is

$$\delta^*(X) = \frac{X+1}{n+2}.$$

The Bayes risk is

$$\begin{aligned} R(\pi, \delta^*) &= \int \pi(p) \sum_{x=0}^{n} [\delta^*(x) - p]^2 f(x \mid p)\,dp \\ &= \int_0^1 E\left(\left(\frac{X+1}{n+2} - p\right)^2 \middle| p\right) dp \\ &= \frac{1}{(n+2)^2}\int_0^1 [np(1-p) + (1-2p)^2]\,dp \\ &= \frac{1}{6(n+2)}. \end{aligned}$$

Example 6. Let $X \sim \mathcal{N}(\mu, 1)$, and let the a priori PDF of μ be $\mathcal{N}(0, 1)$. Also, let $L(\mu, \delta) = [\mu - \delta(\mathbf{X})]^2$. Then

$$h(\mu \mid \mathbf{x}) = \frac{f(\mathbf{x}, \mu)}{g(\mathbf{x})} = \frac{\pi(\mu) f(\mathbf{x} \mid \mu)}{g(\mathbf{x})},$$

where

$$
\begin{aligned}
f(\mathbf{x}) &= \int f(\mathbf{x}, \mu)\, d\mu \\
&= \frac{1}{(2\pi)^{(n+1)/2}} \exp\left(-\frac{1}{2}\sum_1^n x_i^2\right) \int_{-\infty}^{\infty} \exp\left[-\frac{n+1}{2}\left(\mu^2 - 2\mu\frac{n\overline{x}}{n+1}\right)\right] d\mu \\
&= \frac{(n+1)^{-1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum x_i^2 + \frac{n^2\overline{x}^2}{2(n+1)}\right].
\end{aligned}
$$

It follows that

$$
h(\mu \mid \mathbf{x}) = \frac{1}{\sqrt{2\pi/(n+1)}} \exp\left[-\frac{n+1}{2}\left(\mu - \frac{n\overline{x}}{n+1}\right)^2\right],
$$

and the Bayes estimator is

$$
\delta^*(\mathbf{x}) = E\{\mu \mid \mathbf{x}\} = \frac{n\overline{x}}{n+1} = \frac{\sum_1^n x_i}{n+1}.
$$

The Bayes risk is

$$
\begin{aligned}
R(\pi, \delta^*) &= \int \pi(\mu) \int [\delta^*(\mathbf{x}) - \mu]^2 f(\mathbf{x} \mid \mu)\, d\mathbf{x}\, d\mu \\
&= \int_{-\infty}^{\infty} E_\mu\left(\frac{n\overline{X}}{n+1} - \mu\right)^2 \pi(\mu)\, d\mu \\
&= \int_{-\infty}^{\infty} (n+1)^{-2}(n + \mu^2)\pi(\mu)\, d\mu \\
&= \frac{1}{n+1}.
\end{aligned}
$$

The quadratic loss function used in Theorem 1 is but one example of a loss function in frequent use. Some of many other loss functions that may be used are

$$
|\theta - \delta(\mathbf{X})|, \quad \frac{|\theta - \delta(\mathbf{X})|^2}{|\theta|}, \quad |\theta - \delta(\mathbf{X})|^4, \quad \text{and} \quad \left(\frac{|\theta - \delta(\mathbf{X})|}{|\theta| + 1}\right)^{1/2}
$$

Example 7. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$ RVs. It is required to find a Bayes estimator of μ of the form $\delta(x_1, \ldots, x_n) = \delta(\overline{x})$, where $\overline{x} = \sum_1^n x_i/n$, using the loss function $L(\mu, \delta) = |\mu - \delta(\overline{x})|$. From the argument used in the proof of Theorem 1 (or by Remark 1), the Bayes estimator is one that minimizes the integral $\int |\mu - \delta(\overline{x})| h(\mu|\overline{x})\, d\mu$. This will be the case if we choose δ to be the median of the conditional distribution (see Problem 3.2.5).

Let the a priori distribution of μ be $\mathcal{N}(\theta, \tau^2)$. Since $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, we have

$$f(\overline{x}, \mu) = \frac{\sqrt{n}}{2\pi\sigma\tau} \exp\left[-\frac{(\mu-\theta)^2}{2\tau^2} - \frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right].$$

Writing

$$(\overline{x}-\mu)^2 = (\overline{x}-\theta+\theta-\mu)^2 = (\overline{x}-\theta)^2 - 2(\overline{x}-\theta)(\mu-\theta) + (\mu-\theta)^2,$$

we see that the exponent in $f(\overline{x}, \mu)$ is

$$-\frac{1}{2}\left[(\mu-\theta)^2\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right) - \frac{2n(\overline{x}-\theta)(\mu-\theta)}{\sigma^2} + \frac{n}{\sigma^2}(\overline{x}-\theta)^2\right].$$

It follows that the joint PDF of μ and $\overline{X}$ is bivariate normal with means θ, θ, variances τ^2, $\tau^2 + (\sigma^2/n)$, and correlation coefficient $\tau/\sqrt{\tau^2 + (\sigma^2/n)}$. The marginal of $\overline{X}$ is $\mathcal{N}(\theta, \tau^2 + (\sigma^2/n))$, and the conditional distribution of μ, given $\overline{X}$, is normal with mean

$$\theta + \frac{\tau}{\sqrt{\tau^2+(\sigma^2/n)}}\frac{\tau}{\sqrt{\tau^2+(\sigma^2/n)}}(\overline{x}-\theta) = \frac{\theta(\sigma^2/n)+\overline{x}\tau^2}{\tau^2+(\sigma^2/n)}$$

and variance

$$\tau^2\left[1 - \frac{\tau^2}{\tau^2+(\sigma^2/n)}\right] = \frac{\tau^2\sigma^2/n}{\tau^2+(\sigma^2/n)}$$

(see the proof of Theorem 5.4.1). The Bayes estimator is therefore the median of this conditional distribution, and since the distribution is symmetric about the mean,

$$\delta^*(\overline{x}) = \frac{\theta(\sigma^2/n)+\overline{x}\tau^2}{\tau^2+(\sigma^2/n)}$$

is the Bayes estimator of μ.

Clearly, δ^* is also the Bayes estimator under the quadratic loss function $L(\mu, \delta) = [\mu - \delta(\mathbf{X})]^2$.

Key to the derivation of Bayes estimator is the posteriori distribution, $h(\theta \mid \mathbf{x})$. The derivation of the posteriori distribution $h(\theta|\mathbf{x})$, however, is a three-step process:

1. Find the joint distribution of $\mathbf{X}$ and $\boldsymbol{\theta}$ given by $\pi(\theta)f(\mathbf{x} \mid \theta)$.
2. Find the marginal distribution with PDF (PMF) $g(\mathbf{x})$ by integrating (summing) over $\theta \in \Omega$.
3. Divide the joint PDF (PMF) by $g(\mathbf{x})$.

It is not always easy to go through these steps in practice. It may not be possible to obtain $h(\theta \mid \mathbf{x})$ in a closed form.

Example 8. Let $X \sim \mathcal{N}(\mu, 1)$ and the prior PDF of μ be given by

$$\pi(\mu) = \frac{e^{-(\mu-\theta)}}{[1+e^{-(\mu-\theta)}]^2},$$

where θ is a location parameter. Then the joint PDF of X and μ is given by

$$f(x, \mu) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} \frac{e^{-(\mu-\theta)}}{[1+e^{-(\mu-\theta)}]^2}$$

so that the marginal PDF of X is

$$g(x) = \frac{e^{\theta}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{-(x-\mu)^2/2} e^{-\mu}}{[1+e^{-(\mu-\theta)}]^2} d\mu.$$

A closed form for g is not known.

To avoid problem of integration such as that in Example 8, statisticians use *conjugate prior distributions*. Often, there is a natural parameter family of distributions such that the posterior distributions also belong to the same family. These priors make the computations much easier.

Definition 8. Let $X \sim f(x|\theta)$ and $\pi(\theta)$ be the prior distribution on Θ. Then π is said to be a *conjugate prior family* if the corresponding posterior distribution $h(\theta \mid x)$ belongs to the same family as $\pi(\theta)$.

Example 9. Consider Example 6, where $\pi(\mu)$ is $\mathcal{N}(0, 1)$ and $h(\mu \mid \mathbf{x})$ is

$$\mathcal{N}\left(\frac{n\overline{x}}{n+1}, \frac{1}{n+1}\right)$$

so that both h and π belong to the same family. Hence $\mathcal{N}(0, 1)$ is a conjugate prior for μ.

Example 10. Let $X \sim b(n, p)$, $0 < p < 1$, and $\pi(p)$ be the beta PDF with parameters (α, β). Then

$$h(p \mid x) = \frac{p^{x+\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p^{x+\alpha-1}(1-p)^{\beta-1} dp} = \frac{p^{x+\alpha-1}(1-p)^{\beta-1}}{B(x+\alpha, \beta)},$$

which is also a beta density. Thus the family of beta distributions is a conjugate family of priors for p.

Conjugate priors are popular because whenever the prior family is parametric, the posterior distributions are always computable, $h(\theta|x)$ being an updated parametric version of $\pi(\theta)$. One no longer needs to go through a computation of g, the marginal PDF (PMF) of $\mathbf{X}$. Once $h(\theta|x)$ is known, g, if needed, is easily determined from

$$g(\mathbf{x}) = \frac{\pi(\theta) f(\mathbf{x}|\theta)}{h(\theta|\mathbf{x})}.$$

Thus in Example 10, we see easily that $g(x)$ is beta $(x + \alpha, \beta)$, while in Example 6 g is given by

$$g(\mathbf{x}) = \frac{1}{(n+1)^{1/2}(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} x_i^2 + \frac{n^2\bar{x}^2}{2(n-1)} \right].$$

Conjugate priors are usually associated with a wide class of sampling distributions, namely, the exponential family of distributions.

Natural Conjugate Priors

Sampling PDF(PMF), $f(x\|\theta)$	Prior, $\pi(\theta)$	Posterior, $h(\theta\|x)$
$N(\theta, \sigma^2)$	$N(\mu, \tau^2)$	$N\left(\frac{\sigma^2\mu + x\tau^2}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)$
$G(\nu, \beta)$	$G(\alpha, \beta)$	$G(\alpha + \nu, \beta + x)$
$b(n, p)$	$B(\alpha, \beta)$	$B(\alpha + x, \beta + n - x)$
$P(\lambda)$	$G(\alpha, \beta)$	$G(\alpha + x, \beta + 1)$
$NB(r; p)$	$B(\alpha, \beta)$	$B(\alpha + r, \beta + x)$
$G(\gamma, 1/\theta)$	$G(\alpha, \beta)$	$G(\alpha + \nu, \beta + x)$

Another easy way is to use a noninformative prior $\pi(\theta)$, although one needs some integration to obtain $g(\mathbf{x})$.

Definition 9. A PDF $\pi(\theta)$ is said to be a *noninformative prior* if it contains no information about θ; that is, the distribution does not favor any value of θ over others.

Example 11. Some simple examples of noninformative priors are $\pi(\theta) = 1$, $\pi(\theta) = 1/\theta$, and $\pi(\theta) = \sqrt{I(\theta)}$. These may quite often lead to infinite mass and the PDF may be improper (that is, does not integrate to 1).

Calculation of $h(\theta|x)$ becomes easier bypassing the calculation of $g(\mathbf{x})$ when $f(x|\theta)$ is invariant under a group $\mathcal{G}$ of transformations following Fraser's [30] structural theory.

Let $\mathcal{G}$ be a group of Borel-measurable functions on $\mathcal{R}_n$ onto itself. The group operation is composition; that is, if g_1 and g_2 are mappings from $\mathcal{R}_n$ onto $\mathcal{R}_n$, g_2g_1 is defined by $g_2g_1(\mathbf{x}) = g_2(g_1(\mathbf{x}))$. Also, $\mathcal{G}$ is closed under composition and inverse, so that all maps in $\mathcal{G}$ are one-to-one. We define the group G of affine linear transformations $g = \{a, b\}$ by

$$gx = a + bx, \qquad a \in \mathcal{R}, \quad b > 0.$$

The inverse of $\{a, b\}$ is

$$\{a, b\}^{-1} = \left\{-\frac{a}{b}, \frac{1}{b}\right\},$$

and the composition $\{a, b\}$ and $\{c, d\} \in \mathcal{G}$ is given by

$$\begin{aligned}\{a, b\}\{c, d\}(x) &= \{a, b\}(c + dx) = a + b(c + dx)\\ &= (a + bc) + bdx = \{a + bc, bd\}(x).\end{aligned}$$

In particular,

$$\{a, b\}\{a, b\}^{-1} = \{a, b\}\left\{-\frac{a}{b}, \frac{1}{b}\right\} = \{0, 1\} = e.$$

Example 12. Let $X \sim \mathcal{N}(\mu, 1)$ and let $\mathcal{G}$ be the group of translations $\mathcal{G} = \{\{b, 1\}, \ -\infty < b < \infty\}$. Let $X_1, \dots, X_n$ be a sample from $\mathcal{N}(\mu, 1)$. Then we may write

$$X_i = \{\mu, 1\}Z_i, \qquad i = 1, \dots, n$$

where $Z_1, \dots, Z_n$ are iid $\mathcal{N}(0, 1)$.

It is clear that $\overline{Z} \sim \mathcal{N}(0, 1/n)$ with PDF

$$\sqrt{\frac{n}{2\pi}}\exp\left(-\frac{n}{2}\bar{z}^2\right)$$

and there is a one-to-one correspondence between values of $\{\bar{z}, 1\}$ and $\{\mu, 1\}$ given by

$$\{\bar{x}, 1\} = \{\mu, 1\}\{\bar{z}, 1\} = \{\mu + \bar{z}, 1\}.$$

Thus $\bar{x} = \mu + \bar{z}$ with inverse map $\bar{z} = \bar{x} - \mu$. We fix $\bar{x}$ and consider the variation in $\bar{z}$ as a function of μ. Changing the PDF element of $\bar{z}$ to μ, we get

$$\sqrt{\frac{n}{2\pi}}\exp\left[-\frac{n}{2}(\mu - \bar{x})^2\right]$$

as the posterior of μ given $\bar{x}$ with prior $\pi(\mu) = 1$.

Example 13. Let $X \sim \mathcal{N}(0, \sigma^2)$ and consider the scale group $\mathcal{G} = \{\{0, c\}, \ c > 0\}$. Let $X_1, X_2, \dots, X_n$ be iid $\mathcal{N}(0, \sigma^2)$. Write

$$X_i = \{0, \sigma\}Z_i, \qquad i = 1, 2, \dots, n$$

where Z_i are iid $\mathcal{N}(0, 1)$ RVs. Then the RV $nS_z^2 = \sum_{i=1}^n Z_i^2 \sim \chi^2(n)$ with the PDF

$$\frac{1}{2^{n/2}\Gamma(n/2)} \exp\left(-\frac{ns_z^2}{2}\right)(ns_z^2)^{n/2-1}.$$

The values of $\{0, s_z\}$ are in one-to-one correspondence with those of $\{0, \sigma\}$ through

$$\{0, s_x\} = \{0, \sigma\}\{0, s_z\},$$

where $nS_x^2 = \sum_{i=1}^n X_i^2$, so that $s_x = \sigma s_z$. Considering the variation in s_z as a function of σ for fixed s_x, we see that $ds_z = s_x(d\sigma/\sigma^2)$. Changing the PDF element of s_z to σ, we get the PDF of σ as

$$\frac{1}{2^{n/2}\Gamma(n/2)} \exp\left(-\frac{ns_x^2}{2\sigma^2}\right)\left(\frac{ns_x^2}{\sigma^2}\right)^{(n/2)-1}$$

which is the same as the posterior of σ given s_x with prior $\pi(\sigma) = 1/\sigma$.

Example 14. Let $X_1 \dots X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$ and consider the affine linear group $\mathcal{G} = \{\{a, b\}, -\infty < a < \infty, b > 0\}$. Then

$$X_i = \{\mu, \sigma\}Z_i, \qquad i = 1, \dots, n$$

where Z_i's are iid $\mathcal{N}(0, 1)$. We know that the joint distribution of $(\overline{Z}, S_z^2)$ is given by

$$\sqrt{\frac{n}{2\pi}} \exp\left(-\frac{n\bar{z}^2}{2}\right) d\bar{z} \frac{1}{\sqrt{(n-1)/2}} \left[\frac{(n-1)s_z^2}{2}\right]^{[(n-1)/2]-1}$$
$$\times \exp\left[-\frac{(n-1)s_z^2}{2}\right] d\left[\frac{(n-1)s_z^2}{2}\right].$$

Further, the values of $\{\bar{z}, s_z\}$ are in one-to-one correspondence with the values of $\{\mu, \sigma\}$ through

$$\{\bar{x}, s_x\} = \{\mu, \sigma\}\{\bar{z}, s_z\} = \{\mu + \sigma\bar{z}, \sigma s_z\}$$
$$\Rightarrow \bar{z} = \frac{\bar{x} - \mu}{\sigma} \quad \text{and} \quad s_z = \frac{s_x}{\sigma}.$$

Consider the variation of $(\bar{z}, s_z)$ as a function of (μ, σ) for fixed $(\bar{x}, s_x)$. The Jacobian of the transformation from $\{\bar{z}, s_z\}$ to $\{\mu, \sigma\}$ is given by

$$J = \begin{vmatrix} -\frac{1}{\sigma} & -\frac{\overline{x}-\mu}{\sigma^2} \\ 0 & -\frac{s_x}{\sigma^2} \end{vmatrix} = \frac{s_x}{\sigma^3}.$$

Hence, the joint PDF of (μ, σ) given $(\overline{x}, s_x)$ is given by

$$\sqrt{\frac{n}{2\pi}} \exp\left[-\frac{n(\mu-\overline{x})^2}{2\sigma^2}\right] \frac{1}{\sqrt{(n-1)/2}} \left[\frac{(n-1)s_x^2}{2\sigma^2}\right]^{[(n-1)/2]-1}$$
$$\times \exp\left[-\frac{(n-1)s_x^2}{2\sigma^2}\right] \left[\frac{(n-1)s_x^2}{2\sigma^2}\right]^{[(n-1)/2]-1} \frac{(n-1)s_x^2}{\sigma^3}.$$

This is the PDF that one obtains if $\pi(\mu) = 1$ and $\pi(\sigma) = 1/\sigma$ and μ and σ are independent RVs.

The following theorem provides a method for determining minimax estimators.

Theorem 2. Let $\{f_\theta : \theta \in \Theta\}$ be a family of PDFs (PMFs), and suppose that an estimator δ^* of θ is a Bayes estimator corresponding to an a priori distribution π on Θ. If the risk function $R(\theta, \delta^*)$ is constant on Θ, then δ^* is a minimax estimator for θ.

Proof. Since δ^* is the Bayes estimator of θ with constant risk r^* (free of θ), we have

$$\begin{aligned} r^* = R(\pi, \delta^*) &= \int_{-\infty}^{\infty} R(\theta, \delta^*)\pi(\theta)\, d\theta \\ &= \inf_{\delta \in \mathcal{D}} \int R(\theta, \delta)\pi(\theta)\, d\theta \\ &\le \sup_{\theta\in\Theta} \inf_{\delta \in \mathcal{D}} R(\theta, \delta) \le \inf_{\delta \in \mathcal{D}} \sup_{\theta\in\Theta} R(\theta, \delta). \end{aligned}$$

Similarly, since $r^* = R(\theta, \delta^*)$ for all $\theta \in \Theta$, we have

$$r^* = \sup_{\theta\in\Theta} R(\theta, \delta^*) \ge \inf_{\delta \in \mathcal{D}} \sup_{\theta\in\Theta} R(\theta, \delta).$$

Together, we then have

$$\sup_{\theta\in\Theta} R(\theta, \delta^*) = \inf_{\delta \in \mathcal{D}} \sup_{\theta\in\Theta} R(\theta, \delta)$$

which means that δ^* is minimax.

The following examples show how to obtain constant risk estimators and the suitable prior distribution.

Example 15 (Hodges and Lehmann [40]). Let $X \sim b(n, p), 0 \le p \le 1$. We seek a minimax estimator of p of the form $\alpha X + \beta$, using the squared-error loss function. We have

$$\begin{aligned} R(p, \delta) &= E_p(\alpha X + \beta - p)^2 = E_p[\alpha(X - np) + \beta + (\alpha n - 1)p]^2 \\ &= [(\alpha n - 1)^2 - \alpha^2 n]p^2 + [\alpha^2 n + 2\beta(\alpha n - 1)]p + \beta^2, \end{aligned}$$

which is a quadratic equation in p. To find α and β such that $R(p, \delta)$ is constant for all $p \in \Theta$, we set the coefficients of p^2 and p equal to 0 to get

$$(\alpha n - 1)^2 - \alpha^2 n = 0 \quad \text{and} \quad \alpha^2 n + 2\beta(\alpha n - 1) = 0.$$

It follows that

$$\alpha = \frac{1}{\sqrt{n}\,(1 + \sqrt{n})} \quad \text{or} \quad \frac{1}{\sqrt{n}\,(\sqrt{n} - 1)}$$

and

$$\beta = \frac{1}{2(1 + \sqrt{n})} \quad \text{or} \quad -\frac{1}{2(\sqrt{n} - 1)}.$$

Since $0 \le p \le 1$, we discard the second set of roots for both α and β, and then the estimator is of the form

$$\delta^*(x) = \frac{X}{\sqrt{n}\,(1 + \sqrt{n})} + \frac{1}{2(1 + \sqrt{n})}.$$

It remains to show that δ^* is Bayes against some a priori PDF π.

Consider the natural conjugate a priori PDF

$$\pi(p) = [\beta(\alpha', \beta')]^{-1} p^{\alpha' - 1}(1 - p)^{\beta' - 1}, \qquad 0 \le p \le 1, \quad \alpha', \beta' > 0.$$

The a posteriori PDF of p, given x, is expressed by

$$h(p \mid x) = \frac{p^{x + \alpha' - 1}(1 - p)^{n - x + \beta' - 1}}{B(x + \alpha', n - x + \beta')}.$$

It follows that

$$\begin{aligned} E\{p \mid x\} &= \frac{B(x + \alpha' + 1, n - x + \beta')}{B(x + \alpha', n - x + \beta')} \\ &= \frac{x + \alpha'}{n + \alpha' + \beta'}, \end{aligned}$$

which is the Bayes estimator for a squared-error loss. For this to be of the form δ^*, we must have

$$\frac{1}{\sqrt{n}(1+\sqrt{n})} = \frac{1}{n+\alpha'+\beta'} \quad \text{and} \quad \frac{1}{2(1+\sqrt{n})} = \frac{\alpha'}{n+\alpha'+\beta'},$$

giving $\alpha' = \beta' = \sqrt{n}/2$. It follows that the estimator $\delta^*(x)$ is minimax with constant risk

$$R(p, \delta^*) = \frac{1}{4(1+\sqrt{n})^2} \qquad \text{for all} \quad p \in [0, 1].$$

Note that the UMVUE (which is also the MLE) is $\delta(X) = X/n$ with risk $R(p, d) = p(1-p)/n$. Comparing the two risks (Figs. 1 and 2), we see that

$$\frac{p(1-p)}{n} \le \frac{1}{4(1+\sqrt{n})^2} \quad \text{if and only if} \quad |p - \frac{1}{2}| \ge \frac{\sqrt{1+2\sqrt{n}}}{2(1+\sqrt{n})},$$

so that

$$R(p, \delta^*) < R(p, \delta)$$

in the interval $(\frac{1}{2} - a_n, \frac{1}{2} + a_n)$, where $a_n \to 0$ as $n \to \infty$. Moreover,

$$\frac{\sup_p R(p,\delta)}{\sup_p R(p,\delta^*)} = \frac{1/4n}{1/[4(1+\sqrt{n})^2]} = \frac{n+2\sqrt{n}+1}{n} \to 1 \qquad \text{as } n \to \infty.$$

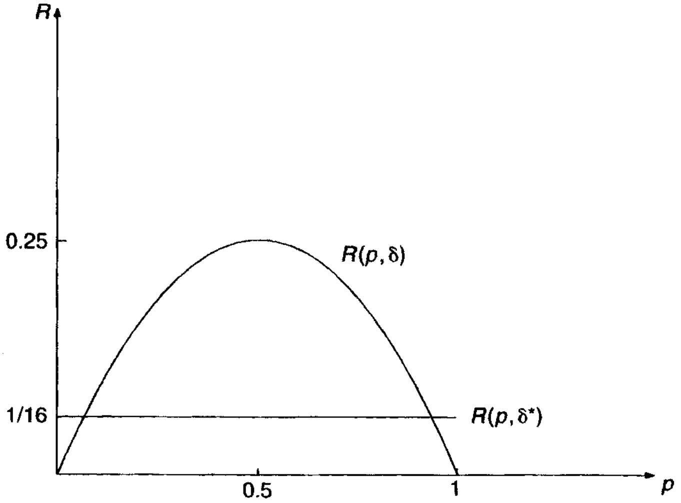

Fig. 1. Comparison of $R(p, \delta)$ and $R(p, \delta^*)$, $n = 1$.

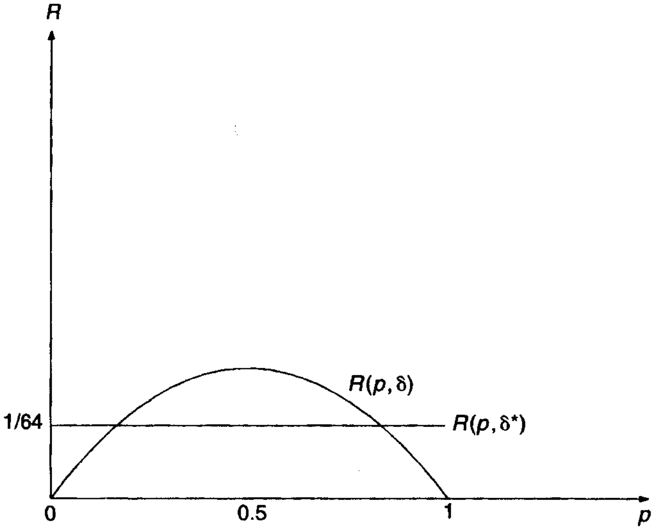

Fig. 2. Comparison of $R(p, \delta)$ and $R(p, \delta^*)$, $n = 9$.

Clearly, we would prefer the minimax estimator if n is small, and would prefer the UMVUE because of its simplicity if n is large.

Example 16 (Hodges and Lehmann [40]). A lot contains N elements, of which D are defective. A random sample of size n produces X defectives. We wish to estimate D. Clearly,

$$P_D\{X = k\} = \binom{D}{k}\binom{N-D}{n-k}\binom{N}{n}^{-1},$$

$$E_D X = n\frac{D}{N} \quad \text{and} \quad \sigma_D^2 = \frac{nD(N-n)(N-D)}{N^2(N-1)}.$$

Proceeding as in Example 15, we find a linear function of X with constant risk. Indeed, $E_D(\alpha X + \beta - D)^2 = \beta^2$ when

$$\alpha = \frac{N}{n + \sqrt{n(N-n)/(N-1)}} \quad \text{and} \quad \beta = \frac{N}{2}\left(1 - \frac{\alpha n}{N}\right).$$

We show that $\alpha X + \beta$ is the Bayes estimator corresponding to the a priori PMF

$$P\{D = d\} = c\int_0^1 \binom{N}{d} p^d (1-p)^{N-d} p^{a-1}(1-p)^{b-1}\,dp,$$

where $a, b > 0$ and $c = \Gamma(a+b)/\Gamma(a)\Gamma(b)$. First note that $\sum_{d=0}^{N} P\{D = d\} = 1$, so that

$$\sum_{d=0}^{N}\binom{N}{d}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+d)\Gamma(N+b-d)}{\Gamma(N+a+b)}=1.$$

The Bayes estimator is given by

$$\delta^*(k)=\frac{\sum_{d=k}^{N-n+k}d\binom{d}{k}\binom{N-d}{n-k}\binom{N}{d}\Gamma(a+d)\Gamma(N+b-d)}{\sum_{d=k}^{N-n+k}\binom{d}{k}\binom{N-d}{n-k}\binom{N}{d}\Gamma(a+d)\Gamma(N+b-d)}.$$

A little simplification, writing $d=(d-a)+a$ and using

$$\binom{d}{k}\binom{N-d}{n-k}\binom{N}{d}=\binom{N-n}{d-k}\binom{N}{n}\binom{n}{k},$$

yields

$$\begin{aligned}\delta^*(k)&=\frac{\sum_{i=0}^{N-n}\binom{N-n}{i}\Gamma(d+a+1)\Gamma(N+b-d)}{\sum_{0}^{N-n}\binom{N-n}{i}\Gamma(d+a)\Gamma(N+b-d)}-a\\&=k\frac{a+b+N}{a+b+n}+\frac{a(N-n)}{a+b+n}.\end{aligned}$$

Now putting

$$\alpha=\frac{a+b+N}{a+b+n}\quad\text{and}\quad\beta=\frac{a(N-n)}{a+b+n}$$

and solving for a and b, we get

$$a=\frac{\beta}{\alpha-1},\qquad b=\frac{N-\alpha n-\beta}{\alpha-1}.$$

Since $a>0$, $\beta>0$, and since $b>0$, $N>\alpha n+\beta$. Moreover, $\alpha>1$ if $N>n+1$. If $N=n+1$, the result is obtained if we give D a binomial distribution with parameter $p=\frac{1}{2}$. If $N=n$, the result is immediate.

The following theorem, which is an extension of Theorem 2, is of considerable help to prove minimaxity of various estimators.

Theorem 3. Let $\{\pi_k(\theta);\ k\geq 1\}$ be a sequence of prior distributions on Θ and let $\{\delta_k^*\}$ be the corresponding sequence of Bayes estimators with Bayes risks $R(\pi_k;\ \delta_k^*)$. If $\limsup_{k\to\infty}R(\pi_k;\ \delta_k^*)=r^*$ and there exists an estimator δ^* for which

$$\sup_{\theta\in\Theta}R(\theta,\delta^*)\leq r^*,$$

then δ^* is minimax.

Proof. Suppose that δ^* is not minimax. Then there exists an estimator $\tilde{\delta}$ such that

$$\sup_{\theta\in\Theta} R(\theta, \tilde{\delta}) \leq \sup_{\theta\in\Theta} R(\theta, \delta^*).$$

On the other hand, consider the Bayes estimators $\{\delta_k^*\}$ corresponding to the priors $\{\pi_k(\theta)\}$. We obtain

$$R(\pi_k, \delta_k^*) = \int R(\theta, \delta_k^*)\pi_k(\theta)\, d\theta \tag{12}$$

$$\leq \int R(\theta, \tilde{\delta})\pi_k(\theta)\, d\theta \tag{13}$$

$$\leq \sup_{\theta\in\Theta} R(\theta, \tilde{\delta}), \tag{14}$$

which contradicts $\sup_{\theta\in\Theta} R(\theta, \delta^*) \leq r^*$. Hence δ^* is minimax.

Example 17. Let $X_1, \dots, X_n$ be a sample of size n from $N(\mu, 1)$. Then the MLE of μ is $\overline{X}$ with variance $1/n$. We show that $\overline{X}$ is minimax. Let $\mu \sim N(0, \tau^2)$. Then the Bayes estimator of μ is $\overline{X}[n\tau^2/(1+n\tau^2)]$. The Bayes risk of this estimator is

$$R(\pi, \delta_{\tau^2}) = \frac{1}{n}\left(\frac{n\tau^2}{1+n\tau^2}\right).$$

Now, as $\tau^2 \to \infty$, $R(\pi, \delta_{\tau^2}^*) \to 1/n$, which is the risk of $\overline{X}$. Hence $\overline{X}$ is minimax.

Definition 10. A decision rule δ is *inadmissible* if there exists a $\delta^* \in \mathcal{D}$ such that $R(\theta, \delta^*) \leq R(\theta, \delta)$, where the inequality is strict for some $\theta \in \Theta$; otherwise, δ is *admissible.*

Theorem 4. If $X_1, \dots, X_n$ is a sample from $N(\theta, 1)$, then $\overline{X}$ is an admissible estimator of θ under square-error loss $L(\theta, a) = (\theta - a)^2$.

Proof. Clearly, $\overline{X} \sim N(\theta, 1/n)$. Suppose that $\overline{X}$ is not admissible, then there exists another rule $\delta^*(\mathbf{x})$ such that $R(\theta, \delta^*) \leq R(\theta, \overline{X})$ while the inequality is strict for some $\theta = \theta_0$ (say). Now, the risk $R(\theta, \delta)$ is a continuous function of θ and hence there exists an $\varepsilon > 0$ such that $R(\theta, \delta^*) < R(\theta, \overline{X}) - \varepsilon$ for $|\theta - \theta_0| < \varepsilon$.

Now consider the prior $N(0, \tau^2)$. Then the Bayes estimator is

$$\delta(\mathbf{X}) = \overline{X}\left(1 + \frac{1}{n\tau^2}\right)^{-1} \qquad \text{with risk} \quad \frac{1}{n}\left(\frac{n\tau^2}{1+n\tau^2}\right).$$

Thus

$$R(\pi, \overline{X}) - R(\pi, \delta_{\tau^2}) = \frac{1}{n}\frac{1}{1+n\tau^2}.$$

However,

$$\tau[R(\pi,\delta^*) - R(\pi,\overline{X})] = \tau \int [R(\theta,\delta^*) - R(\theta,\overline{X})] \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{1}{2\tau^2}\theta^2\right) d\theta$$
$$\leq -\frac{\varepsilon}{\sqrt{2\pi}} \int_{\theta_0-\varepsilon}^{\theta_0+\varepsilon} \exp\left(-\frac{1}{2\tau^2}\theta^2\right) d\theta.$$

We get

$$0 \leq \tau[R(\pi,\delta^*) - R(\pi,\overline{X})] + \tau\left[R(\pi,\overline{X}) - R(\pi,\delta_{\tau^2})\right]$$
$$\leq -\frac{\varepsilon}{\sqrt{2\pi}} \int_{\theta_0+\varepsilon}^{\theta_0+\varepsilon} \exp\left(-\frac{1}{2\tau^2}\theta^2\right) d\theta + \frac{\tau}{n}\frac{1}{(1+n\tau^2)}.$$

The right-hand side goes to $-2\varepsilon^2/\sqrt{2\pi}$ as $\tau \to \infty$. This result leads to a contradiction that δ^* is admissible. Hence $\overline{X}$ is admissible under squared loss.

Thus we have proved the $\overline{X}$ is an admissible minimax estimator of the mean of a normal distribution $\mathcal{N}(\theta, 1)$.

PROBLEMS 8.8

1. It rains quite often in Bowling Green, Ohio. On a rainy day a teacher has essentially three choices: (1) to take an umbrella and face the possible prospect of carrying it around in the sunshine; (2) to leave the umbrella at home and perhaps get drenched; or (3) to just give up the lecture and stay at home. Let $\Theta = \{\theta_1, \theta_2\}$, where θ_1 corresponds to rain, and θ_2, to no rain. Let $\mathcal{A} = \{a_1, a_2, a_3\}$, where a_i corresponds to the choice i, $i = 1, 2, 3$. Suppose that the following table gives the losses for the decision problem:

	θ_1	θ_2
a_1	1	2
a_2	4	0
a_3	5	5

The teacher has to make a decision on the basis of a weather report that depends on θ as follows:

	θ_1	θ_2
W_1 (rain)	0.7	0.2
W_2 (no rain)	0.3	0.8

Find the minimax rule to help the teacher reach a decision.

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from $P(\lambda)$. For estimating λ, using the quadratic error loss function, an a priori distribution over Θ, given by the PDF

$$\begin{aligned} \pi(\lambda) &= e^{-\lambda} && \text{if } \lambda > 0, \\ &= 0 && \text{otherwise,} \end{aligned}$$

is used.

(a) Find the Bayes estimator for λ.

(b) If it is required to estimate $\varphi(\lambda) = e^{-\lambda}$ with the same loss function and same a priori PDF, find the Bayes estimator for $\varphi(\lambda)$.

3. Let $X_1, X_2, \dots, X_n$ be a sample from $b(1, \theta)$. Consider the class of decision rules δ of the form $\delta(x_1, x_2, \dots, x_n) = n^{-1} \sum_{i=1}^{n} x_i + \alpha$, where α is a constant to be determined. Find α according to the minimax principle, using the loss function $(\theta - \delta)^2$, where δ is an estimator for θ.

4. Let δ^* be a minimax estimator for $a\psi(\theta)$ with respect to the squared-error loss function. Show that $a\delta^*+b(a, b$ constants) is a minimax estimator for $a\psi(\theta)+b$.

5. Let $X \sim b(n, \theta)$, and suppose that the a priori PDF of θ is $U(0, 1)$. Find the Bayes estimator of θ, using loss function $L(\theta, \delta) = (\theta - \delta)^2/[\theta(1 - \theta)]$. Find a minimax estimator for θ.

6. In Example 5, find the Bayes estimator for p^2.

7. Let $X_1, X_2, \dots, X_n$ be a random sample from $G(1, 1/\lambda)$. to estimate λ, let the a priori PDF on λ be $\pi(\lambda) = e^{-\lambda}$, $\lambda > 0$, and let the loss function be squared error. Find the Bayes estimator of λ.

8. Let $X_1, X_2, \dots, X_n$ be iid $U(0, \theta)$ RVs. Suppose that the prior distribution of θ is a Pareto PDF $\pi(\theta) = \alpha a^\alpha/\theta^{\alpha+1}$ for $\theta \geq a$, $= 0$ for $\theta < a$. Using the quadratic loss function, find the Bayes estimator of θ.

9. Let T be the unique Bayes estimator of θ with respect to the prior density π. Then T is admissible.

10. Let $X_1, X_2, \dots, X_n$ be iid with PDF $f_\theta(x) = \exp[-(x - \theta)]$, $x > \theta$. Take $\pi(\theta) = e^{-\theta}$, $\theta > 0$. Find the Bayes estimator of θ under quadratic loss.

11. For the PDF of Problem 10, consider the estimation of θ under quadratic loss. Consider the class of estimators $a\left(X_{(1)} - 1/n\right)$ for all $a > 0$. Show that $X_{(1)} - 1/n$ is minimax in this class.

8.9 PRINCIPLE OF EQUIVARIANCE

Let $\mathcal{P} = \{P_{\boldsymbol{\theta}}\colon \boldsymbol{\theta} \in \Theta\}$ be a family of distributions of some RV $\mathbf{X}$. Let $\mathfrak{X} \subseteq \mathcal{R}_n$ be the sample space of values of $\mathbf{X}$. In Section 8.8 we saw that the statistical decision theory revolves around the following four basic elements: the parameter space Θ, the action space $\mathcal{A}$, the sample space $\mathfrak{X}$, and the loss function $L(\theta, a)$.

Let $\mathcal{G}$ be a group of transformations that map $\mathfrak{X}$ onto itself. We say that $\mathcal{P}$ is *invariant* under $\mathcal{G}$ if for each $g \in \mathcal{G}$ and every $\boldsymbol{\theta} \in \Theta$, there is a unique $\boldsymbol{\theta}' = \bar{g}\boldsymbol{\theta} \in \Theta$

such that $g(\mathbf{X}) \sim P_{\bar{g}\boldsymbol{\theta}}$ whenever $\mathbf{X} \sim P_{\boldsymbol{\theta}}$. Accordingly,

$$(1) \qquad P_{\boldsymbol{\theta}}\{g(\mathbf{X}) \in A\} = P_{\bar{g}\boldsymbol{\theta}}\{\mathbf{X} \in A\}$$

for all Borel subsets in $\mathcal{R}_n$. We note that the invariance of $\mathcal{P}$ under $\mathcal{G}$ does not change the class of distributions we begin with; it only changes the parameter or index $\boldsymbol{\theta}$ to $\bar{g}\boldsymbol{\theta}$. The group $\mathcal{G}$ induces $\bar{\mathcal{G}}$, a group of transformations $\bar{g}$ on Θ onto itself.

Example 1. Let $X \sim b(n, p), 0 \le p \le 1$. Let $\mathcal{G} = \{g, e\}$, where $g(x) = n - x$ and $e(x) = x$. Then $gg^{-1} = e$. Clearly, $g(X) \sim b(n, 1-p)$, so that $\bar{g}p = 1 - p$ and $\bar{e}p = e$. The group $\mathcal{G}$ leaves $\{b(n, p);\ 0 \le p \le 1\}$ invariant.

Example 2. Let $X_1, X_2, \dots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$ RVs. Consider the group of affine transformations $\mathcal{G} = \{\{a, b\},\ a \in \mathcal{R},\ b > 0\}$ on $\mathfrak{X}$. The joint PDF of $\{a, b\}X = (a + bX_1, \dots, a + bX_n)$ is given by

$$f(x_1, x_2, \dots, x_n) = \frac{1}{(b\sigma\sqrt{2\pi})^n} \exp\left[-\frac{1}{2b^2\sigma^2}\sum_{i=1}^{n}(x_i - a - b\mu)^2\right]$$

and we see that

$$\bar{g}(\mu, \sigma) = (a + \mu\sigma, b\sigma) = \{a, b\}\{\mu, \sigma\}.$$

Clearly, $\mathcal{G}$ leaves the family of joint PDFs of $\mathbf{X}$ invariant.

To apply invariance considerations to a decision problem we need also to ensure that the loss function is invariant.

Definition 1. A decision problem is said to be *invariant* under a group $\mathcal{G}$ if

(i) $\mathcal{P}$ is invariant under $\mathcal{G}$, and

(ii) the loss function L is invariant in the sense that for every $g \in \mathcal{G}$ and $a \in \mathcal{A}$ there is a unique $a' \in \mathcal{A}$ such that

$$L(\boldsymbol{\theta}, a) = L(\bar{g}\boldsymbol{\theta}, a') \qquad \text{for all } \boldsymbol{\theta}.$$

The $a' \in \mathcal{A}$ in Definition 1 is uniquely determined by g and may be denoted by $\tilde{g}(a)$. One can show that $\tilde{\mathcal{G}} = \{\tilde{g} : g \in \mathcal{G}\}$ is a group of transformations of $\mathcal{A}$ into itself.

Example 3. Consider the estimation of μ in sampling from $\mathcal{N}(\mu, 1)$. In Example 8.9.2 we have shown that the normal family is invariant under the location group $\mathcal{G} = \{\{b, 1\}, -\infty < b < \infty\}$. Consider the quadratic loss function

$$L(\mu, a) = (\mu - a)^2.$$

Then $\{b, 1\}a = b + a$ and $\{b, 1\}\{\mu, 1\} = \{b + \mu, 1\}$. Hence

$$L(\{b, 1\}\mu, \{b, 1\}a) = L[(b + \mu) - (b + a)]^2 = (\mu - a)^2 = L(\mu, a).$$

Thus $L(\mu, a)$ is invariant under $\mathcal{G}$ and the problem of estimation of μ is invariant under group $\mathcal{G}$.

Example 4. Consider the normal family $\mathcal{N}(0, \sigma^2)$ which is invariant under the scale group $\mathcal{G} = \{\{0, c\}, c > 0\}$. Let the loss function be

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

Now $\{0, c\}a = ca$ and $\{0, c\}\{0, \sigma^2\} = \{0, c\sigma^2\}$ and

$$L[\{0, c\}\sigma^2, \{0, c\}a] = \frac{1}{c^2\sigma^4}(c\sigma^2 - ca)^2 = \frac{1}{\sigma^4}(\sigma^2 - a)^2 = L(\sigma^2, a).$$

Thus the loss function $L(\sigma^2, a)$ is invariant under $\mathcal{G} = \{\{0, c\}, c > 0\}$ and the problem of estimation of σ^2 is invariant.

Example 5. Consider the loss function

$$L(\sigma^2, a) = \frac{a}{\sigma^2} - 1 - \log\frac{a}{\sigma^2}$$

for the estimation of σ^2 from the normal family $\mathcal{N}(0, \sigma^2)$. We show that this loss function is invariant under the scale group. Since

$$\{0, c\}\sigma^2 = \{0, c\sigma^2\} \quad \text{and} \quad \{0, c\}\{0, a\} = \{0, ca\},$$

we have

$$\begin{aligned} L[\{0, c\}\sigma^2, \{0, c\}a] &= \frac{ca}{c\sigma^2} - 1 - \log\frac{ca}{c\sigma^2} \\ &= L(\sigma^2, a). \end{aligned}$$

Let us now return to the problem of estimation of a parametric function $\psi : \Theta \to \mathcal{R}$. For convenience let us take $\Theta \subseteq \mathcal{R}$ and $\psi(\theta) = \theta$. Then $\mathcal{A} = \Theta$ and $\tilde{\mathcal{G}} = \bar{\mathcal{G}}$.

Suppose that θ is the mean of PDF f_θ, $\mathcal{G} = \{\{b, 1\},\ b \in \mathcal{R}\}$, and $\{f_\theta\}$ is invariant under $\mathcal{G}$. Consider the estimator $\partial(\mathbf{X}) = \overline{X}$. What we want in an estimator ∂^* of θ is that it changes in the same prescribed way as the data are changed. In our case, since $\mathbf{X}$ changes to $\{b, 1\}\mathbf{X} = \mathbf{X} + b$, we would like $\overline{X}$ to transform to $\{b, 1\}\overline{X} = \overline{X} + b$.

Definition 2. An estimator $\delta(\mathbf{X})$ of θ is said to be *equivariant*, under $\mathcal{G}$, if

$$\delta(g\mathbf{X}) = \bar{g}\delta(\mathbf{X}) \qquad \text{for all } g \in \mathcal{G}, \tag{2}$$

where we have written $g\mathbf{X}$ for $g(\mathbf{X})$ for convenience.

Indeed, g on S induces $\bar{g}$ on Θ. Thus if $\mathbf{X} \sim f_\theta$, then $g\mathbf{X} \sim f_{\bar{g}\theta}$, so if $\delta(\mathbf{X})$ estimates θ then $\delta(g\mathbf{X})$ should estimate $\bar{g}\theta$. The *principle of equivariance* requires that we restrict attention to equivariant estimators and select the "best" estimator in this class in a sense to be described later in this section.

Example 6. In Example 3, consider the estimators $\partial_1(\mathbf{X}) = \overline{X}$, $\partial_2(\mathbf{X}) = (X_{(1)} + X_{(n)})/2$, and $\partial_3(\mathbf{X}) = \alpha\overline{X}$, α a fixed real number. Then $\mathcal{G} = \{(b, 1), \ -\infty < b < \infty\}$ induces $\bar{\mathcal{G}} = \mathcal{G}$ on Θ and both ∂_1, ∂_2 are equivariant under $\mathcal{G}$. The estimator δ_3 is not equivariant unless $\alpha = 1$. In Example 1, $\partial(X) = X/n$ is an equivariant estimator of p.

In Example 6, consider the statistic $\partial(\mathbf{X}) = S^2$. Note that under the translation group $\{b, 1\}\mathbf{X} = \mathbf{X} + b$ and $\partial(\{b, 1\}\mathbf{X}) = \partial(\mathbf{X})$. That is, for every $g \in \mathcal{G}$, $\partial(g\mathbf{X}) = \partial(\mathbf{X})$. A statistic ∂ is said to be *invariant* under a group of transformations $\mathcal{G}$ if $\partial(g\mathbf{X}) = \partial(\mathbf{X})$ for all $g \in \mathcal{G}$. When $\mathcal{G}$ is the translation group, an invariant statistic (function) under $\mathcal{G}$ is called *location invariant*. Similarly, if $\mathcal{G}$ is the scale group, we call ∂ *scale invariant*, and if $\mathcal{G}$ is the location-scale group, we call ∂ *location-scale invariant*. In Example 6, $\partial_4(\mathbf{X}) = S^2$ is location invariant but not equivariant, and $\partial_2(\mathbf{X})$ and $\partial_3(\mathbf{X})$ are not location invariant.

A very important property of equivariant estimators is that their risk function is constant on orbits of θ.

Theorem 1. Suppose that ∂ is an equivariant estimator of θ in a problem that is invariant under $\mathcal{G}$. Then the risk function of ∂ satisfies

$$R(\bar{g}\theta, \partial) = R(\theta, \partial) \tag{3}$$

for all $\theta \in \Theta$ and $g \in \mathcal{G}$. If, in particular, $\bar{\mathcal{G}}$ is transitive over Θ, then $R(\theta, \partial)$ is independent of θ.

Proof. We have for $\theta \in \Theta$ and $g \in \mathcal{G}$,

$$\begin{aligned} R(\theta, \partial(\mathbf{X})) &= E_\theta L(\theta, \partial(\mathbf{X})) \\ &= E_\theta L(\bar{g}\theta, \bar{g}\partial(\mathbf{X})) && \text{(invariance of } L) \\ &= E_\theta L(\bar{g}\theta, \partial(g(\mathbf{X})) && \text{(equivariance of } \delta) \\ &= E_{\bar{g}\theta} L(\bar{g}\theta, \partial(\mathbf{X})) && \text{(invariance of } \{P_\theta\}) \\ &= R(\bar{g}\theta, \partial(\mathbf{X})). \end{aligned}$$

In the special case when $\bar{\mathcal{G}}$ is transitive over Θ, then for any $\theta_1, \theta_2 \in \Theta$ there exists a $\bar{g} \in \bar{\mathcal{G}}$ such that $\theta_2 = \bar{g}\theta_1$. It follows that

$$R(\theta_2, \partial) = R(\bar{g}\theta_1, \partial) = R(\theta_1, \partial)$$

so that R is independent of θ.

Remark 1. When the risk function of every equivariant estimator is constant, an estimator (in the class equivariant estimators) that is obtained by minimizing the constant is called the *minimum risk equivariant* (MRE) *estimator.*

Example 7. Let $X_1, X_2, \ldots, X_n$ iid RVs with common PDF

$$f(x, \theta) = \exp[-(x - \theta)], \qquad x \ge \theta \quad \text{and} \quad = 0 \quad \text{if } x < 0.$$

Consider the location group $\mathcal{G} = \{\{b, 1\}, -\infty < b < \infty\}$, which induces $\bar{\mathcal{G}}$ on Θ where $\bar{\mathcal{G}} = \mathcal{G}$. Clearly, $\bar{\mathcal{G}}$ is transitive. Let $L(\theta, \partial) = (\theta - \partial)^2$. Then the problem of estimation of θ is invariant, and according to Theorem 1, the risk of every equivariant estimator is free of θ. The estimator $\delta_0(\mathbf{X}) = X_{(1)} - 1/n$ is equivariant under $\mathcal{G}$ since

$$\delta_0(\{b, 1\}\mathbf{X}) = \min_{1 \le i \le n} (X_i + b) - \frac{1}{n} = b + X_{(1)} - \frac{1}{n} = b + \delta_0(\mathbf{X}).$$

We leave the reader to check that

$$R(\theta, \partial_0) = E_\theta \left(X_{(1)} - \frac{1}{n} - \theta \right)^2 = \frac{1}{n^2},$$

and it will be seen later that ∂_0 is the MRE estimator of θ.

Example 8. In this example we consider sampling from a normal PDF. Let us first consider estimation of μ when $\sigma = 1$. Let $\mathcal{G} = \{\{b, 1\}, \ -\infty < b < \infty\}$. Then $\partial(\mathbf{X}) = \overline{X}$ is equivariant under $\mathcal{G}$ and it has the smallest risk $1/n$. Note that $\{\bar{x}, 1\}^{-1} = \{-\bar{x}, 1\}$ may be used to designate $\mathbf{x}$ on its orbits

$$\{\bar{x}, 1\}^{-1}\mathbf{x} = (x_1 - \bar{x}, \ldots, x_n - \bar{x}) = A(\mathbf{x}).$$

Clearly, $A(\mathbf{x})$ is invariant under $\mathcal{G}$ and $A(\mathbf{X})$ is ancillary to μ. By Basu's theorem $A(\mathbf{X})$ and $\bar{X}$ are independent.

Next, consider estimation of σ^2 with $\mu = 0$ and $\mathcal{G} = \{\{0, c\}, c > 0\}$. Then $S_x^2 = \sum_1^n X_i^2$ is an equivariant estimator of σ^2. Note that $\{0, s_x\}^{-1}$ may be used to designate $\mathbf{x}$ on its orbits

$$\{0, s_x\}^{-1}\mathbf{x} = \left(\frac{x_1}{s_x}, \ldots, \frac{x_n}{s_x} \right) = A(\mathbf{x}).$$

Again, $A(\mathbf{x})$ is invariant under $\mathcal{G}$ and $A(\mathbf{X})$ is ancillary to σ^2. Moreover, S_x^2 and $A(\mathbf{X})$ are independent.

Finally, we consider estimation of (μ, σ^2) when $\mathcal{G} = \{\{b, c\}, \ -\infty < b < \infty, \ c > 0\}$. Then $(\overline{X}, S_x^2)$, where $S_x^2 = \sum_1^n (X_i - \overline{X})^2$ is an equivariant estimator of (μ, σ^2). Also, $\{\overline{x}, s_x\}^{-1}$ may be used to designate $\mathbf{x}$ on its orbits

$$\{\overline{x}, s_x\}^{-1}\mathbf{x} = \left(\frac{x_1 - \overline{x}}{s_x}, \dots, \frac{x_n - \overline{x}}{s_x}\right) = A(\mathbf{x}).$$

Note that the statistic $A(\mathbf{X})$ defined in each of the three cases considered in Example 8 is constant on its orbits. A statistic A is said to be *maximal invariant* if

(i) A is invariant, and
(ii) A is maximal, that is, $A(\mathbf{x}_1) = A(\mathbf{x}_2) \Rightarrow \mathbf{x}_1 = g(\mathbf{x}_2)$ for some $g \in \mathcal{G}$.

We now derive an explicit expression for MRE estimator for a location parameter. Let $X_1, X_2, \dots, X_n$ be iid with common PDF $f_\theta(x) = f(x - \theta)$, $-\infty < \theta < \infty$. Then $\{f_\theta : \theta \in \Theta\}$ is invariant under $\mathcal{G} = \{\{b, 1\}, \ -\infty < b < \infty\}$, and an estimator of θ is equivariant if

$$\partial(\{b, 1\}\mathbf{X}) = \partial(\mathbf{X}) + b$$

for all real b.

Lemma 1. An estimator ∂ is equivariant for θ if and only if

$$\partial(\mathbf{X}) = X_1 + q(X_2 - X_1, \dots, X_n - X_1), \tag{4}$$

for some function q.

Proof. If (4) holds, then

$$\begin{aligned}\partial(\{b, 1\}\mathbf{x}) &= b + x_1 + q(x_2 - x_1, \dots, x_n - x_1)\\ &= b + \partial(\mathbf{x}).\end{aligned}$$

Conversely,

$$\begin{aligned}\partial(\mathbf{x}) &= \partial(x_1 + x_1 - x_1, x_1 + x_2 - x_1, \dots, x_1 + x_n - x_1)\\ &= x_1 + \partial(0, x_2 - x_1, \dots, x_n - x_1),\end{aligned}$$

which is (4) with $q(x_2 - x_1, \dots, x_n - x_1) = \partial(0, x_2 - x_1, \dots, x_n - x_1)$.

From Theorem 1 the risk function of an equivariant estimator ∂ is constant with risk

$$R(\theta, \partial) = R(0, \partial) = E_0[\partial(\mathbf{X})]^2 \qquad \text{for all } \theta,$$

where the expectation is with respect to the PDF $f_0(\mathbf{x}) = f(\mathbf{x})$. Consequently, among all equivariant estimators ∂ for θ, the MRE estimator is ∂_0, satisfying

$$R(0, \partial_0) = \min_{\partial} R(0, \partial).$$

Thus we only need to choose the function q in (4).

Let $L(\theta, \partial)$ be the loss function. Invariance considerations require that

$$L(\theta, \partial) = L(\bar{g}\theta, \bar{g}\partial) = L(\theta + b, \partial + b)$$

for all real b so that $L(\theta, \partial)$ must be some function w of $\partial - \theta$.

Let $Y_i = X_i - X_1$, $i = 2, \dots, n$, and $\mathbf{Y} = (Y_2, \dots, Y_n)$ and $g(\mathbf{y})$ be the joint PDF of $\mathbf{Y}$ under $\theta = 0$. Let $h(x_1|\mathbf{y})$ be the conditional density, under $\theta = 0$, of X_1 given $\mathbf{Y} = y$. Then

$$\text{(5)} \qquad \begin{aligned} R(0, \partial) &= E_0[w(X_1 - q(\mathbf{Y}))] \\ &= \int \left[\int w(x_1 - q(\mathbf{y}))h(x_1|\mathbf{y})\,dx\right] g(\mathbf{y})\,d\mathbf{y}. \end{aligned}$$

Then $R(0, \partial)$ will be minimized by choosing, for each fixed $\mathbf{y}$, $q(\mathbf{y})$ to be that value of c that minimizes

$$\text{(6)} \qquad \int w(u - c)h(u|\mathbf{y})\,du.$$

Necessarily, q depends on $\mathbf{y}$. In the special case $w(d - \theta) = (d - \theta)^2$, the integral in (6) is minimum when c is chosen to be the mean of the conditional distribution. Thus the unique MRE estimator of θ is given by

$$\text{(7)} \qquad \partial_0(\mathbf{x}) = x_1 - E_\theta\{X_1|\mathbf{Y} = \mathbf{y}\}.$$

This is the *Pitman estimator*. Let us simplify it a little more by computing $E_0\{x_1 - X_1|\mathbf{Y} = \mathbf{y}\}$.

First we need to compute $h(u|\mathbf{y})$. When $\theta = 0$, the joint PDF of $X_1, Y_2, \dots, Y_n$ is easily seen to be

$$f(x_1)f(x_1 + y_2) \cdots f(x_1 + y_n),$$

so the joint PDF of $(Y_2, \dots, Y_n)$ is given by

$$\int_{-\infty}^{\infty} f(u)f(u + y_2) \cdots f(u + y_n)\,du.$$

It follows that

$$\text{(8)} \qquad h(u|\mathbf{y}) = \frac{f(u)f(u + y_2) \cdots f(u + y_n)}{\int_{-\infty}^{\infty} f(u)f(u + y_2) \cdots f(u + y_n)\,du}.$$

Now let $Z = x_1 - X_1$. Then the conditional PDF of Z given $\mathbf{y}$ is $h(x_1 - z \mid \mathbf{y})$. It follows from (8) that

$$\partial_0(\mathbf{x}) = E_0\{Z|\mathbf{y}\} = \int_{-\infty}^{\infty} zh(x_1 - z)\,dz = \frac{\int_{-\infty}^{\infty} z \prod_{j=1}^{n} f(x_j - z)\,dz}{\int_{-\infty}^{\infty} \prod_{j=1}^{n} f(x_j - z)\,dz}. \tag{9}$$

Remark 2. Since the joint PDF of $X_1, X_2, \ldots, X_n$ is $\prod_{j=1}^{n} f_\theta(x_j) = \prod_{j=1}^{n} f(x_j - \theta)$, the joint PDF of θ and $\mathbf{X}$ when θ has prior $\pi(\theta)$ is $\pi(\theta) \prod_{j=1}^{n} f(x_j - \theta)$. The joint marginal of $\mathbf{X}$ is $\int_{-\infty}^{\infty} \pi(\theta) \prod_{j=1}^{n} f(x_j - \theta)\,d\theta$. It follows that the conditional PDF of θ given $\mathbf{X} = \mathbf{x}$ is given by

$$\frac{\pi(\theta) \prod_{j=1}^{n} f(x_j - \theta)}{\int_{-\infty}^{\infty} \pi(\theta) \prod_{j=1}^{n} f(x_j - \theta)\,d\theta}.$$

Taking $\pi(\theta) = 1$, the improper uniform prior on Θ, we see from (9) that $\partial_0(\mathbf{x})$ is the Bayes estimator of θ under squared-error loss and prior $\pi(\theta) = 1$. Since the risk of ∂_0 is constant, it follows that ∂_0 is also a minimax estimator of θ.

Remark 3. Suppose that S is sufficient for θ. Then $\prod_{j=1}^{n} f_\theta(x_j) = g_\theta(s)h(\mathbf{x})$, so that the Pitman estimator of θ can be rewritten as

$$\partial_0(\mathbf{x}) = \frac{\int_{-\infty}^{\infty} \theta \prod_{j=1}^{n} f_\theta(x_j)\,d\theta}{\int_{-\infty}^{\infty} \prod_{j=1}^{n} f_\theta(x_j)\,d\theta} = \frac{\int_{-\infty}^{\infty} \theta g_\theta(s)h(\mathbf{x})\,d\theta}{\int_{-\infty}^{\infty} g_\theta(s)h(\mathbf{x})\,d\theta} = \frac{\int_{-\infty}^{\infty} \theta g_\theta(s)\,d\theta}{\int_{-\infty}^{\infty} g_\theta(s)\,d\theta},$$

which is a function of s alone.

Examples 7 and 8 (*continued*). A direct computation using (9) shows that $X_{(1)} - 1/n$ is the Pitman MRE estimator of θ in Example 7, and $\overline{X}$ is the MRE estimator of μ in Example 8 (when $\sigma = 1$). The results can be obtained by using sufficiency reduction. In Example 7, $X_{(1)}$ is the minimal sufficient statistic for θ. Every (translation) equivariant function based on $X_{(1)}$ must be of the form $\partial_c(\mathbf{X}) = X_{(1)} + c$, where c is a real number. Then

$$R(\theta, \partial_c) = E_\theta\{X_{(1)} + c - \theta\}^2 = E_\theta\left\{X_{(1)} - \frac{1}{n} - \theta + \left(c + \frac{1}{n}\right)\right\}^2$$

$$= R(\theta, \partial_0) + \left(c + \frac{1}{n}\right)^2 = \left(\frac{1}{n}\right)^2 + \left(c + \frac{1}{n}\right)^2,$$

which is minimized for $c = -1/n$. In Example 8, $\overline{X}$ is the minimal sufficient statistic, so every equivariant function of $\overline{X}$ must be of the form $\partial_c(\mathbf{X}) = \overline{X} + c$, where c is a real constant. Then

$$R(\mu, \partial_c) = E_\mu(\overline{X} + c - \mu)^2 = \frac{1}{n} + c^2,$$

which is minimized for $c = 0$.

Example 9. Let $X_1, X_2, \dots, X_n$ be iid $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Then $(X_{(1)}, X_{(n)})$ is jointly sufficient for θ. Clearly,

$$f(x_1 - \theta, \dots, x_n - \theta) = \begin{cases} 1, & x_{(1)} < \theta < x_{(n)}, \\ 0, & \text{otherwise,} \end{cases}$$

so that the Pitman estimator of θ is given by

$$\partial_0(\mathbf{x}) = \frac{\int_{x_{(1)}}^{x_{(n)}} \theta \, d\theta}{\int_{x_{(1)}}^{x_{(n)}} d\theta} = \frac{x_{(n)} + x_{(1)}}{2}.$$

We now consider, briefly, the Pitman estimator of a scale parameter. Let $\mathbf{X}$ have a joint PDF

$$f_\sigma(\mathbf{x}) = \frac{1}{\sigma^n} f\left(\frac{x_1}{\sigma}, \dots, \frac{x_n}{\sigma}\right)$$

where f is known and $\sigma > 0$ is a scale parameter. The family $\{f_\sigma : \sigma > 0\}$ remains invariant under $\mathcal{G} = \{\{0, c\}, c > 0\}$, which induces $\bar{\mathcal{G}} = \mathcal{G}$ on Θ. Then for estimation of σ^k loss function $L(\sigma, a)$ is invariant under these transformations if and only if $L(\sigma, a) = w(a/\sigma^k)$. An estimator ∂ of σ^k is equivariant under $\mathcal{G}$ if

$$\partial(\{0, c\}\mathbf{X}) = c^k \partial(\mathbf{X}) \qquad \text{or all } c > 0.$$

Some simple examples of scale-equivariant estimators of σ are the mean deviation $\sum_1^n |X_i - \overline{X}|/n$ and the standard deviation $\sqrt{\sum_1^n (X_i - \overline{X})^2/(n-1)}$. We note that the group $\bar{\mathcal{G}}$ over Θ is transitive, so according to Theorem 1, the risk of any equivariant estimator of σ^k is free of σ and an MRE estimator minimizes this risk over the class of all equivariant estimators of σ^k. Using the loss function $L(\sigma, a) = w(a/\sigma^k) = (a - \sigma^k)^2/\sigma^{2k}$, it can be shown that the MRE estimator of σ^k, also known as the *Pitman estimator* of σ^k, is given by

$$\partial_0(\mathbf{x}) = \frac{\int_0^\infty v^{n+k-1} f(vx_1, \dots, vx_n)\,dv}{\int_0^\infty v^{n+2k-1} f(vx_1, \dots, vx_n)\,dv}.$$

Just as in the location case, one can show that ∂_0 is a function of the minimal sufficient statistic and ∂_0 is the Bayes estimator of σ^k with improper prior $\pi(\sigma) = 1/\sigma^{2k+1}$. Consequently, ∂_0 is minimax.

Example 8. (*continued*). In Example 8, the Pitman estimator of σ^k is easily shown to be

$$\partial_0(\mathbf{X}) = \frac{\Gamma[(n+k)/2]}{\Gamma[(n+2k)/2]} \left(\sum_1^n X_i^2\right)^{k/2}.$$

Thus the MRE estimator of σ is given by $\{\Gamma[(n+1)/2]\sqrt{\sum_1^n X_i^2}/\,\Gamma[(n+2)/2]\}$ and that of σ^2 by $\sum_1^n X_i^2/(n+2)$.

Example 10. Let $X_1, X_2, \dots, X_n$ be iid $U(0, \theta)$. The Pitman estimator of θ is given by

$$\partial_0(\mathbf{X}) = \frac{\int_{X_{(n)}}^\infty v^n\,dv}{\int_{X_{(n)}}^\infty v^{n+1}\,dv} = \frac{n+2}{n+1} X_{(n)}.$$

Finally, we consider, briefly, estimation of the mean vector of a multivariate normal distribution. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)'$ be a column vector and $\mathbf{I}_p$ be the $p \times p$ identity matrix. Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be a sample from a p-variate normal distribution with mean vector $\boldsymbol{\theta}$ and variance–covariance matrix $\mathbf{I}_p$. Let $L(\boldsymbol{\theta}, \mathbf{a}) = (\boldsymbol{\theta} - \mathbf{a})'(\boldsymbol{\theta} - \mathbf{a}) = \sum_{i=1}^p (\theta_i - a_i)^2$. In the univariate ($p = 1$) case we have seen that the sample mean $\bar{\mathbf{X}}$ is a minimax and admissible estimator of θ. It is therefore natural to consider $\bar{\mathbf{X}} = (\overline{X}_1, \overline{X}_2, \dots, \overline{X}_p)'$ as an estimator of $\boldsymbol{\theta}$ also in the p-variate case and suspect that it has the same properties as in $p = 1$ case. Certainly, $\bar{\mathbf{X}}$ is a minimax estimator, but is it admissible, too? Stein [108] showed that $\bar{\mathbf{X}}$ is admissible for $p = 2$. But for $p \geq 3$, James and Stein [45] showed that the estimator

$$\text{(10)} \qquad \boldsymbol{\theta}^s(\mathbf{X}) = \left(1 - \frac{p-2}{n\bar{\mathbf{X}}'\bar{\mathbf{X}}}\right)\bar{\mathbf{X}}$$

improves on $\bar{\mathbf{X}}$ for all $\boldsymbol{\theta}$.

This is a surprising result but is typical in a variety of multiparameter estimation problems. What is optimal in independent estimation problems is not necessarily optimal if the problems are considered simultaneously. It should be noted, however, that $\boldsymbol{\theta}^s$ does not share the other optimality properties of $\bar{\mathbf{X}}$. It is not MLE, is biased, and is not equivariant. It only dominates $\bar{\mathbf{X}}$ under quadratic loss.

The estimator $\boldsymbol{\theta}^s$ takes $\bar{\mathbf{X}}$ and *shrinks* it toward the origin (provided $\bar{\mathbf{X}}'\bar{\mathbf{X}} > p-2$).

PROBLEMS 8.9

In all problems assume that $X_1, X_2, \ldots, X_n$ is a random sample from the distribution under consideration.

1. Show that the following statistics are equivariant under translation group:
   (a) Median (X_i).
   (b) $(X_{(1)} + X_{(n)})/2$.
   (c) $X_{[np]+1}$, the quantile of order p, $0 < p < 1$.
   (d) $\left(X_{(r)} + X_{(r+1)} + \cdots + X_{(n-r)}\right)/(n-2r)$.
   (e) $\overline{X} + \overline{Y}$, where $\overline{Y}$ is the mean of a sample of size m, $m \neq n$.

2. Show that the following statistics are invariant under location or scale or location-scale group:
   (a) $\overline{X} - \text{median}(X_i)$.
   (b) $X_{(n+1-k)} - X_{(k)}$.
   (c) $\sum_{i=1}^n |X_i - \overline{X}|/n$.
   (d) $\dfrac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2\right]^{1/2}}$, where $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from a bivariate distribution.

3. Let the common distribution be $G(\alpha, \sigma)$, where α (> 0) is known and $\sigma > 0$ is unknown. Find the MRE estimator for σ under loss $L(\sigma, a) = (1 - a/\sigma)^2$.

4. Let the common PDF be the folded normal distribution

$$\sqrt{\frac{2}{\pi}} \exp\left[-\tfrac{1}{2}(x-\mu)^2\right] I_{[\mu,\infty)}(x).$$

   Verify that the best equivariant estimator of μ under quadratic loss is given by

$$\hat{\mu} = \overline{X} - \frac{\exp[-(n/2)(X_{(1)} - \overline{X})^2]}{\sqrt{2n\pi}\left[\int_0^{\sqrt{n}(X_{(1)}-\overline{X})} (1/\sqrt{2\pi}) \exp(-z^2/2)\, dz\right]}.$$

5. Let $X \sim U(\theta, 2\theta)$.
   (a) Show that $(X_{(1)}, X_{(n)})$ is jointly sufficient statistic for θ.
   (b) Verify whether or not $(X_{(n)} - X_{(1)})$ is an unbiased estimator of θ. Find an ancillary statistic.
   (c) Determine the best invariant estimator of θ under the loss function $L(\theta, a) = (1 - a/\theta)^2$.

6. Let

$$f_\theta(x) = \tfrac{1}{2}\exp\{-|x-\theta|\}.$$

Find the Pitman estimator of θ.

7. Let $f_\theta(x) = \exp[-(x-\theta)] \cdot \{[1+\exp-(x-\theta)]\}^{-2}$, for $x \in \mathcal{R}, \theta \in \mathcal{R}$. Find the Pitman estimator of θ.

8. Show that an estimator ∂ is (location) equivariant if and only if

$$\partial(\mathbf{x}) = \partial_0(\mathbf{x}) + \phi(\mathbf{x}),$$

where ∂_0 is any equivariant estimator and ϕ is an invariant function.

9. Let X_1, X_2 be iid with PDF

$$f_\sigma(x) = \frac{2}{\sigma}\left(1-\frac{x}{\sigma}\right), \qquad 0 < x < \sigma \quad \text{and} \quad = 0 \text{ otherwise.}$$

Find, explicitly, the Pitman estimator of σ^r.

10. Let $X_1, X_2, \ldots, X_n$ be iid with PDF

$$f_\theta(x) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right), \quad x > 0 \quad \text{and} \quad = 0, \text{ otherwise.}$$

Find the Pitman estimator of θ^k.