CHAPTER 10

POINT ESTIMATION

10.1 INTRODUCTION

Traditionally, problems of statistical inference are divided into problems of estimation and tests of hypotheses, though actually they are all decision problems and, hence, could be handled by the unified approach that we presented in the preceding chapter. The main difference between the two kinds of problems is that in problems of estimation we must determine the value of a parameter (or the values of several parameters) from a possible continuum of alternatives, whereas in tests of hypotheses we must decide whether to accept or reject a specific value or a set of specific values of a parameter (or those of several parameters).

When we use the value of a statistic to estimate a population parameter, we call this point estimation and we refer to the value of the statistic as a point estimate of the parameter. For example, if we use a value of \( \bar{X} \) to estimate the mean \( \mu \) of a population, an observed sample proportion to estimate the parameter \( \pi \) of a binomial population, or a value of \( \hat{S}^2 \) to estimate a population variance, we are in each case using a point estimate of the parameter in question. These estimates are called point estimates because in each case a single number, or a single point on the real axis, is used to estimate the parameter.

Correspondingly, we refer to the statistics themselves as point estimators. For instance, \( \bar{X} \) may be used as a point estimator of \( \mu \), in which case \( \bar{X} \) is a point estimate of this parameter. Similarly, \( \hat{S}^2 \) may be used as a point estimator of \( \sigma^2 \), in which case \( \hat{S}^2 \) is a point estimate of this parameter. Here we used the word "point" to distinguish between these estimators and estimates and the interval estimators and interval estimates, which we shall present in Chapter 11.

Since estimators are random variables, one of the key problems of point estimation is to study their sampling distributions. For instance, which is to estimate the variance of a population on the basis of a random sample, we can hardly expect that the value of \( S^2 \) we get will actually equal \( \sigma^2 \), but it would be reassuring, at least, to know whether we can expect it to be close. Also, if we must decide whether to use a simple mean or a simple median to estimate the mean of a population, it would be important to know, among other things, whether \( \bar{X} \) or \( \hat{S} \) is more likely to yield a value that is actually close.

Various statistical properties of estimators can thus be used to decide which estimator is most appropriate in a given situation, which will expose us to the smallest risk, which will give us the most information at the lowest cost, and so forth. The particular properties of estimators that we shall discuss in Sections 10.2 through 10.6 are unbiasedness, minimum variance, efficiency, consistency, sufficiency, and robustness.

10.2 UNBIASED ESTIMATORS

As we saw on page 387, perfect decision functions do not exist, and in connection with problems of estimation this means that there are no perfect estimators that always give the right answer. Thus, it would seem reasonable that an estimator should do so at least on the average; that is, its expected value should equal the parameter that it is supposed to estimate. If this is the case, the estimator is said to be unbiased; otherwise, it is said to be biased. Formally,

**Definition 10.1.** A statistic \( \hat{\theta} \) is an unbiased estimator of the parameter \( \theta \) if and only if \( E(\hat{\theta}) = \theta \).

The following are some examples of unbiased and biased estimators.

**Example 10.1.** If \( X \) has the binomial distribution with the parameters \( n \) and \( \theta \), show that the sample proportion \( \frac{X}{n} \) is an unbiased estimator of \( \theta \).

**Solution.** Since \( E(X) = n\theta \), it follows that

\[
E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} \cdot n\theta = \theta
\]

and hence that \( \frac{X}{n} \) is an unbiased estimator of \( \theta \).

**Example 10.2.** Show that unless \( \theta = \frac{1}{2} \), the minimum estimator of the binomial parameter \( \theta \) on page 396 is biased.
Chapter 10  Point Estimation

Solution Since $E(Y) = n\theta$, it follows that

$$
E\left(\frac{X - \frac{1}{n} \sum_{i=1}^{n} X_i}{\frac{1}{n} \sum_{i=1}^{n} X_i}\right) = E\left(\frac{X - \frac{1}{\sqrt{n}}}{\frac{1}{\sqrt{n}}}\right) = \frac{\theta + \frac{1}{\sqrt{n}}}{\theta + \frac{1}{\sqrt{n}}} = \frac{\theta}{\theta + \frac{1}{\sqrt{n}}}
$$

and it can easily be seen that this quantity does not equal $\theta$ unless $\theta = \frac{1}{\sqrt{n}}$. ■

EXAMPLE 10.3

If $X_1, X_2, \ldots, X_n$ constitute a random sample from the population given by

$$
f(x) = \begin{cases} 
\theta^{-1} e^{-\theta x} & \text{for } x > 0 \\
0 & \text{elsewhere}
\end{cases}
$$

show that $\bar{X}$ is a biased estimator of $\delta$.

Solution Since the mean of the population is

$$
\mu = \int_{0}^{\infty} x \cdot \theta^{-1} e^{-\theta x} dx = 1 + \delta
$$

it follows from Theorem 8.6 that $E(\bar{X}) = 1 + \delta \neq \delta$ and hence that $\bar{X}$ is a biased estimator of $\delta$. ■

When $\hat{\theta}$ is a biased estimator of $\theta$, it may be of interest to know the extent of the bias, given by

$$
h(\hat{\theta}) = E(\hat{\theta}) - \theta
$$

Thus, for Example 10.2 the bias is

$$
\frac{\theta + \frac{1}{\sqrt{n}}}{\theta + \frac{1}{\sqrt{n}}} - \theta = \frac{\theta}{\theta + \frac{1}{\sqrt{n}}}
$$

and it can be seen that it tends to be small when $\theta$ is close to $\frac{1}{\sqrt{n}}$ and also when $n$ is large. Indeed, $\lim_{n \to \infty} h(\hat{\theta}) = 0$ and we say that the estimator is asymptotically unbiased.

As far as Example 10.3 is concerned, the bias is $(1 + \delta) - \delta = 1$, but here there is something we can do about it. Since $E(\bar{X}) = 1 + \delta$, it follows that $E(\bar{X} - 1) = \delta$ and hence that $\bar{X} - 1$ is an unbiased estimator of $\delta$. The following is another example where a minor modification of our estimator leads to an estimator that is unbiased.

EXAMPLE 10.4

If $X_1, X_2, \ldots, X_n$ constitute a random sample from a uniform population with $\alpha = \theta$, show that the largest sample value that is, the $n$th order statistic $X_n$ is a biased estimator of the parameter $\theta$. Also, modify the estimator of $\theta$ to make it unbiased.

Solution Substituting into the formula for $\bar{X}_n(\theta)$ on page 287, we find that the sampling distribution of $\bar{X}_n$ is given by

$$
E(\bar{X}_n) = \frac{n}{\beta} \left( \int_{0}^{\beta} \frac{1}{\beta} dx \right)^{n-1} = \frac{n}{\beta} \cdot \frac{\beta^{n-1}}{n} = \frac{\beta^{n-1}}{n}
$$

for $0 < \beta < \theta$ and $E(\bar{X}_n) = 0$ elsewhere, and hence that

$$
E(Y_n) = \frac{n}{\beta} \int_{0}^{\beta} \frac{x^n}{\beta^2} dx = \frac{n}{\beta} \cdot \frac{\beta^{n-1}}{n} = \frac{\beta^{n-1}}{n}
$$

Thus, $E(Y_n) = \beta$ and the $n$th order statistic is a biased estimator of the parameter $\beta$. However, since

$$
E\left(\frac{\beta + 1}{n} - \frac{\beta}{n}\right) = \frac{n+1}{n} - \frac{\beta}{n} = \beta
$$

it follows that $\frac{\beta + 1}{n}$ times the largest sample value is an unbiased estimator of the parameter $\beta$. ■

Having discussed unbiasedness as a desirable property of an estimator, we can now explain why we divide by $n - 1$ and not by $n$ when we defined the sample variance: it makes $S^2$ an unbiased estimator of $\sigma^2$ for random samples from infinite populations.

Theorem 10.1. If $X^2$ is the variance of a random sample from an infinite population with the finite variance $\sigma^2$, then $E(S^2) = \sigma^2$.

Proof. By Definition 8.2,

$$
E(S^2) = E\left[\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2\right] = \frac{1}{n-1} E\left[\sum_{i=1}^{n} (X_i - \mu)^2 - (\bar{X} - \mu)^2\right]
$$

$$
= \frac{1}{n-1} \left[ \sum_{i=1}^{n} E((X_i - \mu)^2) - n \cdot E((\bar{X} - \mu)^2) \right]
$$

Then, since $E((X_i - \mu)^2) = \sigma^2$ and $E((\bar{X} - \mu)^2) = \frac{\sigma^2}{n}$, it follows that

$$
E(S^2) = \frac{1}{n-1} \sum_{i=1}^{n} \sigma^2 - n \cdot \frac{\sigma^2}{n} = \sigma^2
$$
Although $\hat{\theta}$ is an unbiased estimator of the variance of an infinite population, it is not an unbiased estimator of the variance of a finite population, and in neither case is it an unbiased estimator of $\sigma$. The bias of $\hat{\theta}$ as an estimator of $\sigma$ is discussed, among others, in the book by E. S. Koenig listed among the references at the end of this chapter.

The discussion of the preceding paragraph illustrates one of the difficulties associated with the concept of unbiasedness. It may not be retained under functional transformations: that is, if $\hat{\theta}$ is an unbiased estimator of $\theta$, it does not necessarily follow that $a \hat{\theta}$ is an unbiased estimator of $a \theta$. Another difficulty associated with the concept of unbiasedness is that unbiased estimators are not necessarily unique.

For instance, in Example 10.2 we shall see that $\frac{n+1}{n} \bar{X}$ is not the only unbiased estimator of the parameter $\beta$ of Example 10.4, and in Exercise 10.8 we shall see that $\bar{X} - 1$ is not the only unbiased estimator of the parameter $\delta$ of Example 10.5.

10.3 EFFICIENCY

If we have to choose one of several unbiased estimators of a given parameter, we usually take the one whose sampling distribution has the smallest variance. We already mentioned this on page 288, where comparing the sample median with the sample mean, we said that the estimator with the smaller variance is "more reliable."

To check whether a given unbiased estimator has the smallest possible variance, that is, whether it is a minimum variance unbiased estimator (also called a best unbiased estimator), we make use of the fact that if $\hat{\theta}$ is an unbiased estimator of $\theta$, it can be shown under very general conditions (referenced on page 353) that the variance of $\hat{\theta}$ must satisfy the inequality

$$\text{var}(\hat{\theta}) \geq \frac{1}{n} \cdot E \left[ \left( \frac{\partial f(x)}{\partial \theta} \right)^2 \right]$$

where $f(x)$ is the value of the population density at $x$ and $n$ is the size of the random sample. This inequality, the Cramér-Rao inequality, leads to the following result.

**Theorem 10.2.** If $\hat{\theta}$ is an unbiased estimator of $\theta$ and

$$\text{var}(\hat{\theta}) = \frac{1}{n} \cdot E \left[ \left( \frac{\partial f(X)}{\partial \theta} \right)^2 \right]$$

then $\hat{\theta}$ is a minimum variance unbiased estimator of $\theta$.

Here, the quantity in the denominator is referred to as the information about $\theta$ that is supplied by the sample (see also Exercise 10.19). Thus, the smaller the variance is, the greater the information.

**Example 10.5**

Show that $\bar{X}$ is a minimum variance unbiased estimator of the mean $\mu$ of a normal population.

**Solution**

Since $f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for $-\infty < x < \infty$

it follows that

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

so that

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

and hence

$$E \left[ \frac{\partial f(x)}{\partial \mu} \right] = \frac{1}{\sigma^2} E \left[ \left( \frac{x - \mu}{\sigma} \right)^2 \right] = \frac{\sigma^2}{\sigma^2} = 1$$

Thus, for $\bar{X}$ unbiased and $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$ according to Theorem 10.1, it follows that $\bar{X}$ is a minimum variance unbiased estimator of $\mu$.

It would be erroneous to conclude from this example that $\bar{X}$ is a minimum variance unbiased estimator of the mean of any population. Indeed, in Exercise 10.3 the reader will be asked to verify that this is not so for random samples of size $n = 3$ from the continuous uniform population with $\alpha = -\frac{1}{2}$ and $\beta = \frac{1}{2}$.

As we have indicated, unbiased estimators of one and the same parameter are usually compared in terms of the size of their variances. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of the parameter $\theta$ of a given population, we say that $\hat{\theta}_1$ is less than the variance of $\hat{\theta}_2$; we say that $\hat{\theta}_1$ is relatively more efficient than $\hat{\theta}_2$. Also, we use the ratio

$$\frac{\text{var}(\hat{\theta}_2)}{\text{var}(\hat{\theta}_1)}$$

as a measure of the efficiency of $\hat{\theta}_2$ relative to $\hat{\theta}_1$.

**Example 10.6**

In Example 10.4 we saw that if $X_1, X_2, \ldots, X_n$ constitute a random sample from a uniform population with $\alpha = 0$, then $\frac{n+1}{n} \cdot \bar{X}$ is an unbiased estimator of $\beta$.

(a) Show that $\frac{2\bar{X}}{n}$ is also an unbiased estimator of $\beta$.

(b) Compare the efficiency of these two estimators of $\beta$. 

Solution

(a) Since the mean of the population is \( \mu = \frac{\beta}{2} \) according to Theorem 6.1, it follows from Theorem 8.1 that \( E(\bar{X}) = \frac{\beta}{2} \) and hence that \( E(2\bar{X}) = \beta \). Thus, \( 2\bar{X} \) is an unbiased estimator of \( \beta \).

(b) First we must find the variances of the two estimators. Using the sampling distribution of \( Y_n \), and the expression for \( E(Y_n) \) given in Example 10.4, we get

\[
\Sigma(Y_n^2) = \frac{n}{\beta^2} \int_0^\beta Y_n^2 \, dy_n = \frac{n}{n+2} \cdot \beta^2
\]

and

\[
\text{var}(Y_n) = \frac{n}{n+2} \cdot \beta^2 = \left( \frac{n}{n+2} \cdot \frac{\beta}{2} \right)^2
\]

If we leave the details to the reader in Exercise 10.27, it can be shown that

\[
\text{var} \left( \frac{n+1}{n} Y_n \right) = \frac{\beta^2}{n(n+2)}
\]

Since the variance of the population is \( \sigma^2 = \frac{\beta^2}{12} \) according to Theorem 6.1, it follows from Theorem 8.3 that \( \text{var}(\bar{X}) = \frac{\beta^2}{12n} \) and hence that

\[
\text{var}(2\bar{X}) = \beta \cdot \text{var}(\bar{X}) = \frac{\beta^2}{3n}
\]

Therefore, the efficiency of \( 2\bar{X} \) relative to \( \frac{n+1}{n} Y_n \) is given by

\[
\frac{\text{var} \left( \frac{n+1}{n} Y_n \right)}{\text{var}(2\bar{X})} = \frac{\frac{\beta^2}{n(n+2)}}{\frac{\beta^2}{3n}} = \frac{3}{n+2}
\]

and it can be seen that for \( n > 1 \) the estimator based on the \( n \)th order statistic is much more efficient than the other one. For \( n = 10 \), for example, the relative efficiency is only 25 percent, and for \( n = 25 \) it is only 11 percent.

Example 10.7

When the mean of a normal population is estimated on the basis of a random sample of size \( 2n + 1 \), what is the efficiency of the median relative to the mean?

Solution

From Theorem 8.1, we know that \( \bar{X} \) is unbiased and that

\[
\text{var}(\bar{X}) = \frac{\sigma^2}{2n + 1}
\]

As far as \( \bar{X} \) is concerned, it is unbiased by virtue of the symmetry of the normal distribution about its mean, and we know from the discussion following Theorem 8.17 that for large samples

\[
\text{var}(\bar{X}) = \frac{n^2 \sigma^2}{4n}
\]

Thus, for large samples, the efficiency of the median relative to the mean is

\[
\frac{\text{var}(\bar{X})}{\text{var}(\bar{X})} = \frac{\frac{\sigma^2}{2n + 1}}{\frac{\sigma^2}{4n}} = \frac{4n}{n(2n + 1)} = \frac{4}{2n + 1}
\]

and the asymptotic efficiency of the median with respect to the mean is

\[
\lim_{n \to \infty} \frac{4n}{n(2n + 1)} = \frac{2}{\pi}
\]

or about 64 percent.

The result of the preceding example may be interpreted as follows: For large samples, the mean requires only 66 percent as many observations as the median to estimate \( \mu \) with the same reliability.

It is important to note that we have limited our discussion of relative efficiency to unbiased estimators. If we included biased estimators, we could always assure ourselves of an estimator with \( n \)-th variance by letting its values equal the same constant regardless of the data that we may obtain. Therefore, if \( \hat{\theta} \) is not an unbiased estimator of a given parameter \( \theta \), we judge its merits and make efficiency comparisons on the basis of the mean square error \( E[(\hat{\theta} - \theta)^2] \) instead of the variance of \( \hat{\theta} \).

Exercises

10.1. If \( X_1, X_2, \ldots, X_n \) constitute a random sample from a population with the mean \( \mu \), what condition must be imposed on the constants \( a_1, a_2, \ldots, a_n \) so that

\[
a_1X_1 + a_2X_2 + \cdots + a_nX_n
\]

is an unbiased estimator of \( \mu \)?

10.2. If \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) are unbiased estimators of the same parameter \( \theta \), what condition must be imposed on the constants \( k_1 \) and \( k_2 \) so that

\[
k_1\hat{\theta}_1 + k_2\hat{\theta}_2
\]

is also an unbiased estimator of \( \theta \)?

10.3. Use the formula for the sampling distribution of \( \bar{X} \) on page 287 to show that for random samples of size \( n = 3 \) the median is an unbiased estimator of the parameter \( \theta \) of a uniform population with \( \alpha = \theta - 1 \) and \( \beta = \theta + 1 \).

10.4. Use the result of Example 8.4 to show that for random samples of size \( n = 3 \) the median is a biased estimator of the parameter \( \theta \) of an exponential population.
10.5. Given a random sample of size \( n \) from a population that is the known mean \( \mu \) and the finite variance \( \sigma^2 \), show that
\[
\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2
\]
is an unbiased estimator of \( \sigma^2 \).

10.6. Use the results of Theorem 6.1 to show that \( \overline{X} \) is an asymptotically unbiased estimator of \( \sigma^2 \).

10.7. Show that \( \frac{X + 1}{n + 2} \) is a biased estimator of the binomial parameter \( \theta \). Is this estimator asymptotically unbiased?

10.8. With reference to Example 10.3, find an unbiased estimator of \( \beta \) based on the smallest sample value (that is, on the first order statistic, \( X_1 \)).

10.9. With reference to Example 10.4, find an unbiased estimator of \( \beta \) based on the smallest sample value (that is, on the first order statistic, \( X_1 \)).

10.10. If \( X_1, X_2, \ldots, X_n \) constitute a random sample from a normal population with \( \mu = 0 \), show that
\[
\sum_{i=1}^{n} X_i^2 / n
\]
is an unbiased estimator of \( \sigma^2 \).

10.11. If \( X \) is a random variable having the binomial distribution with the parameters \( n \) and \( \theta \), show that \( X / n = (1 - X / n) \) is a biased estimator of the variance of \( X \).

10.12. If a random sample of size \( n \) is taken without replacement from the finite population that consists of the positive integers 1, 2, \ldots, \( k \), show that
\[
f(Y_k) = \binom{Y_k - 1}{Y_k - \frac{n}{k}}
\]
for \( Y_k = n, \ldots, k; \)

(b) \( n + 1 / k \), \( Y_k - 1 \) is an unbiased estimator of \( k \).

See also Exercise 10.80.

10.13. Show that if \( \hat{\theta} \) is an unbiased estimator of \( \theta \) and \( \text{var}(\hat{\theta}) \neq 0 \), then \( \hat{\theta}^2 \) is not an unbiased estimator of \( \theta^2 \).

10.14. Show that the sample proportion \( \overline{X} \) is a minimum variance unbiased estimator of the binomial parameter \( \theta \). (Hint: Treat \( \overline{X} \) as the mean of a random sample of size \( n \) from a Bernoulli population with the parameter \( \theta \).)

10.15. Show that the mean of a random sample of size \( n \) is a minimum variance unbiased estimator of the parameter \( \lambda \) of a Poisson population.

10.16. If \( \theta_1 \) and \( \theta_2 \) are independent unbiased estimators of a given parameter \( \theta \) and \( \text{var}(\hat{\theta}_1) = 3 \cdot \text{var}(\hat{\theta}_2) \), find the constants \( a_1 \) and \( a_2 \) such that \( a_1 \hat{\theta}_1 + a_2 \hat{\theta}_2 \) is an unbiased estimator with minimum variance for such a linear combination.

10.17. Show that the mean of a random sample of size \( n \) from an exponential population is a minimum variance unbiased estimator of the parameter \( \theta \).

10.18. Show that for the unbiased estimator of Example 10.4, \( \frac{1}{n} \sum_{i=1}^{n} X_i \), the Cramér-Rao inequality is not satisfied.

10.19. The information about \( \theta \) in a random sample of size \( n \) is also given by
\[
-n \cdot E \left[ \frac{\partial^2 \ln f(X)}{\partial \theta^2} \right]
\]
where \( f(x) \) is the value of the probability density at \( x \), provided that the extorses of the region for which \( f(x) \neq 0 \) do not depend on \( \theta \). The derivation of this formula takes the following steps:

(a) Differentiating the expressions on both sides of
\[
\int f(x) \, dx = 1
\]
with respect to \( \theta \), show that
\[
\int \frac{\partial \ln f(x)}{\partial \theta} \, f(x) \, dx = 0
\]
by interchanging the order of integration and differentiation.

(b) Differentiating again with respect to \( \theta \), show that
\[
E \left[ \frac{\partial^2 \ln f(X)}{\partial \theta^2} \right] = \left[ E \left( \frac{\partial \ln f(X)}{\partial \theta} \right)^2 \right] = -n \cdot E \left[ \frac{\partial^2 \ln f(X)}{\partial \theta^2} \right]
\]

10.20. Rework Example 10.5 using the alternative formula for the information given in Exercise 10.19.

10.21. If \( X \) is the mean of a random sample of size \( n \) from a normal population with the mean \( \mu \) and the variance \( \sigma^2 \), \( \overline{X} \) is the mean of a random sample of size \( n \) from a normal population with the mean \( \mu \) and the variance \( \sigma_1^2 \), and the two samples are independent, show that

(a) \( \overline{X} = (1 - \omega) \overline{X}_1 + \omega \overline{X}_2 \), where \( 0 \leq \omega \leq 1 \), is an unbiased estimator of \( \mu \).

(b) The variance of this estimator is a minimum when
\[
\omega = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}
\]

10.22. With reference to Exercise 10.21, find the efficiency of the estimator of part (a) with \( \omega = \frac{1}{2} \) relative to this estimator with
\[
\omega = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}
\]

10.23. If \( \overline{X}_1 \) and \( \overline{X}_2 \) are the means of independent random samples of size \( n_1 \) and \( n_2 \) from a normal population with the mean \( \mu \) and the variance \( \sigma^2 \), show that the variance of the unbiased estimator
\[
\omega = \frac{n_1}{n_1 + n_2} \overline{X}_1 + (1 - \omega) \overline{X}_2
\]
is a minimum when \( \omega = \frac{n_1}{n_1 + n_2} \).
328 Chapter 10  Point Estimation

10.24. With reference to Exercise 10.23, and the efficiency of the estimator with \( \omega = \frac{1}{2} \) of the estimator with \( \omega = \frac{n}{n+1} \).

10.25. If \( X_1, X_2, \) and \( X_3 \) constitute a random sample of size \( n = 3 \) from a normal population with the mean \( \mu \) and the variance \( \sigma^2 \), find the efficiency of \( \frac{X_1 + 2X_2 + X_3}{3} \) relative to \( \frac{X_1 + X_2 + X_3}{3} \).

10.26. If \( X_1 \) and \( X_2 \) constitute a random sample of size \( n = 2 \) from an exponential population, find the efficiency of \( 2X_1 \), relative to \( X_1 \), where \( X_1 \) is the first order statistic and \( X_2 \) and \( X_3 \) are both unbiased estimators of the parameter \( \theta \).

10.27. Verify the result given for \( \text{var} \left( \frac{X_1 + X_2}{2} \right) \) in Example 10.6.

10.28. With reference to Example 10.3, we showed on page 320 that \( X - 1 \) is an unbiased estimator of \( \lambda \), and in Exercise 10.5 the reader was asked to find another unbiased estimator of \( \lambda \) based on the smallest sample value. Find the efficiency of the first estimator of \( \lambda \) relative to the second.

10.29. With reference to Exercise 10.12, show that \( 2X - 1 \) is also an unbiased estimator of \( \lambda \), and find the efficiency of this estimator relative to the one of part (b) of Exercise 10.12 for

(a) \( n = 2 \);
(b) \( n = 3 \).

10.30. Since the variances of the mean and the midrange are not affected if the same constant is added to each observation, we can determine these variances for random samples of size 3 from the uniform population

\[
f(x) = \begin{cases} 1 & \text{for } -\frac{1}{2} < x < \frac{1}{2} \\ 0 & \text{elsewhere} \end{cases}
\]

by referring instead to the uniform population

\[
f(u) = \begin{cases} 1 & \text{for } 0 < u < 1 \\ 0 & \text{elsewhere} \end{cases}
\]

(a) Show that \( E(X) = 0 \), \( E(X^2) = \frac{1}{3} \), and \( \text{var}(X) = \frac{1}{12} \) for this population so that for a random sample of size \( n = 3 \), \( \text{var}(X) = \frac{1}{4} \).

(b) Use the results of Exercises 8.55 and 8.56 (or derive the necessary densities and joint density) to show that for a random sample of size \( n = 3 \) from this population, the order statistics \( X_1 \) and \( X_3 \) have \( E(X_1) = \frac{1}{3} \), \( E(X_3) = \frac{2}{3} \), and \( E(Y_1) = \frac{2}{3} \), \( E(Y_3) = \frac{1}{3} \) so that \( \text{var}(Y_1) = \frac{1}{10} \), \( \text{var}(Y_3) = \frac{1}{10} \), and \( \text{cov}(Y_1, Y_3) = \frac{1}{10} \).

(c) Use the results of part (b) and Theorem 4.14 to show that \( E \left( \frac{X_1 + X_3}{2} \right) = \frac{1}{2} \).

10.31. Show that if \( \hat{\theta} \) is a biased estimator of \( \theta \), then

\[
E(\hat{\theta} - \theta)^2 = \text{var}(\hat{\theta}) + [\theta(\theta)]^2
\]

10.32. If \( \hat{\theta}_1 \), \( \hat{\theta}_2 \), and \( \hat{\theta}_3 \) are estimators of the parameter \( \theta \) of a binomial population and \( \theta = \frac{1}{2} \), for what values of \( \phi \) is

(a) the mean square error of \( \hat{\theta}_2 \) less than the variance of \( \hat{\theta}_1 \);
(b) the mean square error of \( \hat{\theta}_2 \) less than the variance of \( \hat{\theta}_1 \)?

10.4 CONSISTENCY

In the preceding section we assumed that the variance of an estimator, or its mean square error, is a good indication of its chance fluctuations. The fact that these measures may not provide good criteria for this purpose is illustrated by the following example. Suppose that we want to estimate on the basis of one observation the parameter \( \theta \) of the population given by

\[
f(x) = \frac{\omega}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} + (1-\omega) \frac{1}{\sqrt{2\pi (x-\theta)^2}}
\]

for \(-\infty < x < \infty \) and \( 0 < \omega < 1 \). Evidently, this population is a combination of a normal population with the mean \( \theta \) and the variance \( \sigma^2 \) and a Cauchy distribution (see Exercise 6.10) with \( \omega = \theta \) and \( \beta = 1 \). Now, if \( \theta \) is very close to 0, say, \( \omega = 1 - 10^{-100} \), and \( \sigma \) is very small, say, \( \sigma = 10^{-100} \), the probability that a random variable having this distribution will take on a value that is very close to \( \theta \), and hence is a very good estimate of \( \theta \), is practically 1. Yet, since the variance of the Cauchy distribution does not exist, neither will the variance of this estimator.

The example of the preceding paragraph is a bit fortetached, but it suggests that we pay more attention to the probabilities with which estimators will take on values that are close to the parameter that they are supposed to estimate. The reader may recall that we already touched upon this question in Sections 5.4 and 5.2. Having our argument on Chebyshev's theorem, we showed on page 170 that when \( n \to \infty \) the probability approaches 1 that the sample proportion \( \frac{1}{n} \) will take on a value that differs from the binomial parameter \( \phi \) by less than any arbitrary constant \( c > 0 \). Also, using Chebyshev's theorem, we showed in Theorem 8.2 that when \( n \to \infty \) the probability approaches 1 that \( X \) will take on a value that differs from the mean of the population sampled by less than any arbitrary constant \( c > 0 \). In both of these examples we were practically assured that, for large \( n \), the estimators will take on values that are very close to the respective parameters.

Formally, this concept of "closeness" is expressed by means of the following definition of consistency.

**Definition 10.2.** A statistic \( \hat{\theta} \) is a **consistent estimator** of the parameter \( \theta \) if and only if for each \( \epsilon > 0 \)

\[
\lim_{n \to \infty} P(|\hat{\theta} - \theta| < \epsilon) = 1
\]

Note that consistency is an asymptotic property, that is, a limiting property of an estimator. Informally, Definition 10.2 says that when \( n \) is sufficiently large, we can be practically certain that the error made with a consistent estimator will be less
than any small preassigned positive constant. The kind of convergence expressed by the limit in Definition 10.2 is generally called convergence in probability.

Based on Chebyshev's theorem, we have thus shown in Section 5.4 that \( \bar{X} \) is a consistent estimator of the binomial parameter \( \theta \) and in Theorem 5.2 that \( \bar{X} \) is a consistent estimator of the mean of a population with a finite variance. In practice, we can often judge whether an estimator is consistent by using the following sufficient condition, which is, in fact, an immediate consequence of Chebyshev's theorem.

**Theorem 10.3**. If \( \hat{\theta} \) is an unbiased estimator of the parameter \( \theta \) and \( \text{var}(\hat{\theta}) \to 0 \) as \( n \to \infty \), then \( \hat{\theta} \) is a consistent estimator of \( \theta \).

**Example 10.8**

Show that for a random sample from a normal population, the sample variance \( S^2 \) is a consistent estimator of \( \sigma^2 \).

**Solution** Since \( S^2 \) is an unbiased estimator of \( \sigma^2 \) in accordance with Theorem 10.1, it remains to be shown that \( \text{var}(S^2) \to 0 \) as \( n \to \infty \). Referring to the result of Exercise 8.21 (or to Theorem 8.11, on which this exercise is based), we find that for a random sample from a normal population

\[
\text{var}(S^2) = \frac{2\sigma^4}{n-1}
\]

It follows that \( \text{var}(S^2) \to 0 \) as \( n \to \infty \), and we have thus shown that \( S^2 \) is a consistent estimator of the variance of a normal population.

It is of interest to note that Theorem 10.3 also holds if we substitute "asymptotically unbiased" for "unbiased." This is illustrated by the following example.

**Example 10.9**

With reference to Example 10.3, show that the smallest sample value (that is, the first order statistic) \( Y_1 \) is a consistent estimator of the parameter \( \delta \).

**Solution** Substituting into the formula for \( g_1(y) \) on page 286, we find that the sampling distribution of \( Y_1 \) is given by

\[
g_1(y) = n \cdot e^{-(\delta-y)} \left[ \int_{y}^{\infty} e^{-(\delta-x)} \, dx \right]^{n-1}
\]

for \( y > \delta \) and \( g_1(y) = 0 \) elsewhere. Based on this result, it can easily be shown that \( E(Y_1) = \delta + \frac{1}{n} \), and hence that \( Y_1 \) is an asymptotically unbiased estimator of \( \delta \).

Furthermore,

\[
P(Y_1 - \delta < \varepsilon) = P(\delta < Y_1 < \delta + \varepsilon)
\]

\[
= \int_{\delta}^{\delta + \varepsilon} n \cdot e^{-n(\delta-x)} \, dx
\]

\[
= 1 - e^{-n}
\]

Since \( \lim_{n \to \infty} (1 - e^{-n}) = 1 \), it follows from Definition 10.2 that \( Y_1 \) is a consistent estimator of \( \delta \).

As we indicated on page 30, Theorem 10.3 provides a sufficient condition for the consistency of an estimator. It is not a necessary condition because consistent estimators need not be unbiased, or even asymptotically unbiased. This is illustrated by Exercise 10.41.

### 10.5 Sufficiency

An estimator \( \hat{\theta} \) is said to be **sufficient** if it utilizes all the information in a sample relevant to the estimation of \( \theta \), that is, if all the knowledge about \( \theta \) that can be gained from the individual sample values and their order can just as well be gained from the value of \( \hat{\theta} \) alone.

Formally, we can describe this property of an estimator by referring to the conditional probability distribution or density of the sample values given \( \hat{\theta} = \theta \), which is given by

\[
f(x_1, x_2, \ldots, x_n | \theta) = \frac{f(x_1, x_2, \ldots, x_n, \hat{\theta})}{g(\hat{\theta})} = \frac{f(x_1, x_2, \ldots, x_n)}{g(\hat{\theta})}
\]

If it depends on \( \theta \), then particular values of \( X_1, X_2, \ldots, X_n \) yielding \( \hat{\theta} = \theta \) will be more probable for some values of \( \theta \) than for others, and the knowledge of those sample values will help in the estimation of \( \theta \). On the other hand, if \( \theta \) does not depend on \( \theta \), then particular values of \( X_1, X_2, \ldots, X_n \) yielding \( \hat{\theta} = \theta \) will be just as likely for any value of \( \theta \), and the knowledge of those sample values will be of no help in the estimation of \( \theta \).

**Definition 10.3**. The statistic \( \hat{\theta} \) is a **sufficient estimator** of the parameter \( \theta \) if and only if for each value of \( \theta \) the conditional probability distribution or density of the random sample \( X_1, X_2, \ldots, X_n \), given \( \hat{\theta} = \theta \), is independent of \( \theta \).

**Example 10.10**

If \( X_1, X_2, \ldots, X_n \) constitute a random sample of size \( n \) from a Bernoulli population, show that

\[
\hat{\theta} = \frac{X_1 + X_2 + \cdots + X_n}{n}
\]

is a sufficient estimator of the parameter \( \theta \).
Solution By Definition 5.2, 
\[ f(x_1; \theta) = \theta^x (1-\theta)^{n-x} \quad \text{for } x_1 = 0, 1 \]
so that
\[ f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{n-x_i} = \frac{n!}{\prod_{i=1}^{n} x_i!} \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i} = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i} = \theta^{X} (1-\theta)^{n-X} \]
for \( x_i = 0 \) or 1 and \( i = 1, 2, \ldots, n \). Also, since \( X = X_1 + X_2 + \cdots + X_n \) is a binomial random variable with the parameters \( n \) and \( \theta \), its distribution is given by
\[ b(x; n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n \]
and the transformation-of-variable technique of Section 7.3 yields
\[ g(\theta) = \binom{n}{\sum_{i=1}^{n} x_i} \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i} \quad \text{for } \theta = 0, 1, 2, \ldots, n \]
Now, substituting into the formula for \( f(x_1, x_2, \ldots, x_n; \theta) \) on page 331, we get
\[ \frac{f(x_1, x_2, \ldots, x_n; \theta)}{g(\theta)} = \frac{\theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i}}{\binom{n}{\sum_{i=1}^{n} x_i} \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i}} = \frac{1}{\binom{n}{\sum_{i=1}^{n} x_i}} = \frac{1}{\binom{n}{X}} = \frac{1}{n \choose \frac{X}{n}} \]
for \( x_i = 0 \) or 1 and \( i = 1, 2, \ldots, n \). Evidently, this does not depend on \( \theta \) and we have shown, therefore, that \( \hat{\theta} = \frac{X}{n} \) is a sufficient estimator of \( \theta \).
EXAMPLE 10.12
Show that $\bar{X}$ is a sufficient estimator of the mean $\mu$ of a normal population with the known variance $\sigma^2$.

**Solution**
Making use of the fact that

$$f(x_1, x_2, \ldots, x_n; \mu) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum (x_i - \mu)^2}$$

and that

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{X})^2 + \sum_{i=1}^{n} (\bar{X} - \mu)^2$$

we get

$$f(x_1, x_2, \ldots, x_n; \mu) = \left( \frac{1}{\sqrt{n} \sigma \sqrt{2\pi}} \right)^n e^{-\frac{1}{n} \sum (x_i - \bar{X})^2}$$

where the first factor on the right-hand side depends only on the estimate $\bar{X}$ and the population mean $\mu$, and the second factor does not involve $\mu$. According to Theorem 10.4, it follows that $\bar{X}$ is a sufficient estimator of the mean $\mu$ of a normal population with the known variance $\sigma^2$.

Based on Definition 10.3 and Theorem 10.4, respectively, we have presented two ways of checking whether a statistic $\hat{\theta}$ is a sufficient estimator of a given parameter $\theta$. As we already said, the factorization theorem usually leads to easier solutions; but if we want to show that a statistic $\hat{\theta}$ is not a sufficient estimator of a given parameter $\theta$, it is nearly always easier to proceed with Definition 10.2. This was illustrated by Example 10.11.

Let us also mention the following important property of sufficient estimators. If $\hat{\theta}$ is a sufficient estimator of $\theta$, then any single-valued function $Y = u(\hat{\theta})$, not involving $\theta$, is also a sufficient estimator of $\theta$, and therefore of $\mu(x)$, provided $x = u(\theta)$ can be solved to give the single-valued inverse $\hat{\theta} = u(x)$. This follows directly from Theorem 10.4, since we can write

$$f(x_1, x_2, \ldots, x_n; \hat{\theta}) = g[u(x), \hat{\theta}] \cdot h(x_1, x_2, \ldots, x_n)$$

where $g[u(x), \hat{\theta}]$ depends only on $x$ and $\hat{\theta}$. If we apply this result to Example 10.10, where we showed that $\hat{\theta} = \bar{X}$ is a sufficient estimator of the Bernoulli parameter $\theta$, it follows that $X = X_1 + X_2 + \cdots + X_n$ is also a sufficient estimator of the mean $\mu = n\theta$ of a binomial population.

### 10.6 ROBUSTNESS

In recent years, special attention has been paid to a statistical property called robustness. It is indicative of the extent to which estimation procedures (and, as we shall see later, other methods of inference) are adversely affected by violations of underlying assumptions. In other words, an estimator is said to be robust if its sampling distribution is not seriously affected by violations of assumptions. Such violations are often due to outliers caused by outright errors made, say, in reading instruments or recording the data or by mistakes in experimental procedures. They may also pertain to the nature of the populations sampled or their parameters. For instance, when estimating the average useful life of a certain electronic component, we may think that we are sampling an exponential population, whereas actually we are sampling a Weibull population, or when estimating the average income of a certain age group, we may use a method based on the assumption that we are sampling a normal population, whereas actually the population (income distribution) is highly skewed. Also, when estimating the difference between the average weights of two kinds of frogs, the difference between the mean IQ's of two college groups, and in general the difference $\mu_1 - \mu_2$ between the means of two populations, we may be assuming that the two populations have the same variance $\sigma^2$, whereas in reality $\sigma_1^2 \neq \sigma_2^2$.

As should be apparent, most questions of robustness are difficult to answer: indeed, much of the language used in the preceding paragraph is relatively imprecise. After all, what do we mean by "not seriously affected"? Furthermore, when we speak of violations of underlying assumptions, it should be clear that some violations are more serious than others. When it comes to questions of robustness, we are thus faced with all sorts of difficulties, mathematically and otherwise, and for the most part can be resolved only by computer simulations. The subject of robustness will be reviewed again briefly in Section 10.6.1.

### EXERCISES

10.33. Use Definition 10.2 to show that $Y_1$, the first order statistic, is a consistent estimator of the parameter $\alpha$ of a uniform population with $\beta = 0 + 1$.

10.34. With reference to Exercise 10.33, use Theorem 10.3 to show that $Y_1 = \frac{1}{n + 1}$ is a consistent estimator of the parameter $\alpha$.

10.35. With reference to the uniform population of Example 10.4, use the definition of consistency to show that $Y_n$, the nth order statistic, is a consistent estimator of the parameter $\beta$.

10.36. If $X_1, X_2, \ldots, X_n$ constitute a random sample of size $n$ from an exponential population, show that $\bar{X}$ is a consistent estimator of the parameter $\theta$.

10.37. With reference to Exercise 10.36, is $\bar{X}$ a consistent estimator of the parameter $\theta$?
10.38. Show that the estimator of Exercise 10.21 is consistent.
10.39. Substituting "asymptotically unbiased" for "unbiased" in Theorem 10.3, show that \( \frac{X + 1}{n + 2} \) is a consistent estimator of the binomial parameter \( \theta \).
10.40. Substituting "asymptotically unbiased" for "unbiased" in Theorem 10.3, use this theorem to rework Exercise 10.35.
10.41. To show that an estimator can be consistent without being unbiased or even asymptotically unbiased, consider the following estimation procedure: To estimate the mean of a population with the finite variance \( \sigma^2 \), we take a random sample of size \( n \). Then we randomly draw one of \( n \) slips of paper numbered from 1 through \( n \), and if the number we draw is 2, 3, ..., or \( n \), we use as our estimator the mean of the random sample; otherwise, we use the estimate \( x_1 \). Show that this estimation procedure is
  (a) consistent;
  (b) neither unbiased nor asymptotically unbiased.
10.42. If \( X_1, X_2, \ldots, X_n \) constitute a random sample of size \( n \) from an exponential population, show that \( X \) is a sufficient estimator of the parameter \( \theta \).
10.43. If \( X_1 \) and \( X_2 \) are independent random variables having binomial distributions with the parameters \( \theta \) and \( n_1 \) and \( \theta \) and \( n_2 \), show that \( \frac{X_1 + X_2}{n_1 + n_2} \) is an estimator of \( \theta \).
10.44. In Exercise 10.43, is \( \frac{X_1 + 2X_2}{n_1 + 2n_2} \) a sufficient estimator of \( \theta \)?
10.45. After referring to Example 10.4, is the 8th-order statistic, \( X_8 \), a sufficient estimator of the parameter \( \theta \)?
10.46. If \( X_1 \) and \( X_2 \) constitute a random sample of size \( n = 2 \) from a Poisson population, show that the mean of the sample is a sufficient estimator of the parameter \( \lambda \).
10.47. If \( X_1, X_2, \) and \( X_3 \) constitute a random sample of size \( n = 3 \) from a Bernoulli population, show that \( Y = X_1 + 2X_2 + X_3 \) is a sufficient estimator of \( \theta \).
10.48. If \( X_1, X_2, \ldots, X_n \) constitute a random sample of size \( n \) from a geometric population, show that \( Y = X_1 + X_2 + \ldots + X_n \) is a sufficient estimator of the parameter \( \theta \).
10.49. Show that the estimator of Exercise 10.5 is a sufficient estimator of the variance of a normal population with the known mean \( \mu \).

### 10.7 THE METHOD OF MOMENTS

As we have seen in this chapter, there can be many different estimators of one and the same parameter of a population. Therefore, it would seem desirable to have some general method, or methods, that yield estimators with as many desirable properties as possible. In this section and in Section 10.8 we shall present two such methods, the method of moments, which is historically one of the oldest methods, and the method of maximum likelihood. Furthermore, Bayesian estimation will be treated briefly in Section 10.9 and another method, the method of least squares, will be taken up in Chapter 14.

The method of moments consists of equating the first few moments of a population to the corresponding moments of a sample, thus getting as many equations as are needed to solve for the unknown parameters of the population.

**Definition 10.4.** The \( k \)-th sample moment of a set of observations \( x_1, x_2, \ldots, x_n \) is the mean of their \( k \)-th powers and is denoted by \( m_k \); symbolically,

\[
\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k
\]

Thus, if a population has \( r \) parameters, the method of moments consists of solving the system of equations

\[
m_k' = \mu_k', \quad k = 1, 2, \ldots, r
\]

for the \( r \) parameters.

**Example 10.13.**

Given a random sample of size \( n \) from a uniform population with \( \beta = 1 \), use the method of moments to obtain a formula for estimating the parameter \( \alpha \).

**Solution**

The equation that we shall have to solve is \( m_1' = \mu_1' \), where \( m_1' = \bar{x} \) and \( \mu_1' = \frac{\alpha + \beta}{2} = \frac{\alpha + 1}{2} \) according to Theorem 6.1. Thus,

\[
\bar{x} = \frac{\alpha + 1}{2}
\]

and we can write the estimate of \( \alpha \) as

\[
\hat{\alpha} = 2\bar{x} - 1
\]

**Example 10.14.**

Given a random sample of size \( n \) from a gamma population, use the method of moments to obtain formulas for estimating the parameters \( \alpha \) and \( \beta \).

**Solution**

The system of equations that we shall have to solve is

\[
m_1' = \mu_1' \quad \text{and} \quad m_2' = \mu_2'
\]

where \( \mu_1' = \alpha \beta \) and \( \mu_2' = \alpha (\alpha + 1) \beta^2 \) according to Theorem 6.2. Thus,

\[
m_1' = \alpha \beta \quad \text{and} \quad m_2' = \alpha (\alpha + 1) \beta^2
\]

and, solving for \( \alpha \) and \( \beta \), we get the following formulas for estimating the two parameters of the gamma distribution:

\[
\hat{\alpha} = \frac{(m_2')^2}{m_2' - (m_1')^2} \quad \text{and} \quad \hat{\beta} = \frac{m_1' - (m_1')^2}{m_2'}
\]
Chapter 10

Point estimation

For the sample values $x_1, x_2, \ldots, x_n$, we can write

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

in terms of the original observations.

In these examples we were concerned with the parameters of a specific population. It is important to note, however, that when the parameters to be estimated are the moments of the population, then the method of moments can be used without any knowledge about the nature or functional form of the population.

10.8 THE METHOD OF MAXIMUM LIKELIHOOD

In two papers published early in the last century, R. A. Fisher, the prominent statistician whom we already mentioned on page 280, proposed a general method of estimation called the method of maximum likelihood. He also demonstrated the advantages of this method by showing that it yields sufficient estimators whenever they exist and that maximum likelihood estimators are asymptotically minimum variance unbiased estimators.

To help understand the principle on which the method of maximum likelihood is based, suppose that four letters arrive in somebody's morning mail, but unfortunately one of them is misplaced before the recipient has a chance to open it. If, among the remaining three letters, two contain credit-card billings and one a bill, what might be a good estimate of $k$, the total number of credit-card billings among the four letters received? Clearly, $k$ must be two or three.

For $k = 2$, the probability of observing two credit-card billings and one bill is

$$\frac{3}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{3}$$

For $k = 3$, the probability of observing two credit-card billings and one bill is

$$\frac{3}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{3}{3}$$

Thus, the maximum likelihood estimate is $k = 3$. Therefore, if we choose as our estimate of $k$ the value that maximizes the probability of getting the observed data, we obtain $k = 3$.

The general idea applies to many cases where there are several unknown parameters. In the discrete case, if the observed sample values are $x_1, x_2, \ldots, x_n$, the probability of getting them is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1, x_2, \ldots, x_n; \theta)$$

which is just the value of the joint probability distribution of $n$ i.i.d. random variables $X_1, X_2, \ldots, X_n$ at $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. Since the sample values have been observed and are therefore fixed numbers, we regard $f(x_1, x_2, \ldots, x_n; \theta)$ as a value of a function of $\theta$, and we refer to this function as the likelihood function. An analogous definition applies when the random sample comes from a continuous population, but in that case $f(x_1, x_2, \ldots, x_n; \theta)$ is the value of the joint probability density of the random variables $X_1, X_2, \ldots, X_n$ at $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

**Definition 10.5.** If $x_1, x_2, \ldots, x_n$ are the values of a random sample from a population with the parameter $\theta$, the likelihood function of the sample is given by

$$L(\theta) = f(x_1, x_2, \ldots, x_n; \theta)$$

for values of $\theta$ within a given domain. Here $f(x_1, x_2, \ldots, x_n; \theta)$ is the value of the joint probability distribution or the joint probability density of the random variables $X_1, X_2, \ldots, X_n$ at $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

Thus, the method of maximum likelihood consists of maximizing the likelihood function with respect to $\theta$, and we refer to the value of $\theta$ that maximises the likelihood function as the maximum likelihood estimate of $\theta$.

**Example 10.15.**

Given $n$ "successes" in $n$ trials, find the maximum likelihood estimate of the parameter $\theta$ of the corresponding binomial distribution.

**Solution.** To find the value of $\theta$ that maximizes

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

it will be convenient to make use of the fact that the value of $\theta$ that maximizes $L(\theta)$ will also maximize

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

Thus, we get

$$\frac{d \ln L(\theta)}{d \theta} = \frac{\ln L(\theta)}{\theta} = \frac{n - x}{1 - \theta}$$

Therefore, the maximum likelihood estimate of $\theta$ is

$$\hat{\theta} = \frac{x}{n}$$
and, equating this derivative to 0 and solving for \( \theta \), we find that the likelihood function has a maximum at \( \hat{\theta} = \frac{\bar{x}}{n} \). This is the maximum likelihood estimate of the binomial parameter \( \theta \), and we refer to \( \hat{\theta} = \frac{X}{n} \) as the corresponding maximum likelihood estimator.

**EXAMPLE 10.16**

If \( x_1, x_2, \ldots, x_n \) are the values of a random sample from an exponential population, find the maximum likelihood estimator of its parameter \( \theta \).

**Solution** Since the likelihood function is given by

\[
L(\theta) = f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = \left( \frac{1}{\theta} \right)^n e^{-\frac{1}{\theta} \sum x_i}
\]

differentiation of \( \ln L(\theta) \) with respect to \( \theta \) yields

\[
\frac{d}{d\theta} \ln L(\theta) = -\frac{h}{\theta} + \frac{1}{\theta^2} \sum x_i
\]

Equating this derivative to zero and solving for \( \theta \), we get the maximum likelihood estimate

\[
\hat{\theta} = \frac{1}{n} \sum x_i = \bar{x}
\]

Hence, the maximum likelihood estimator is \( \hat{\theta} = \bar{x} \).

Now let us consider an example in which straightforward differentiation cannot be used to find the maximum value of the likelihood function.

**EXAMPLE 10.17**

If \( x_1, x_2, \ldots, x_n \) are the values of a random sample of size \( n \) from a uniform population with \( \alpha = 0 \) (as in Example 10.4), find the maximum likelihood estimator of \( \beta \).

**Solution** The likelihood function is given by

\[
L(\beta) = \prod_{i=1}^{n} f(x_i; \beta) = \left( \frac{1}{\beta} \right)^n
\]

for \( \beta \) greater than or equal to the largest of the \( x \)'s and \( \beta \) otherwise. Since the value of this likelihood function increases as \( \beta \) decreases, we must make \( \beta \) as small as possible, and it follows that the maximum likelihood estimate of \( \beta \) is \( x_n \), the \( n \)th order statistic.

Comparing the result of this example with that of Example 10.4, we find that maximum likelihood estimators need not be unbiased. The ones of Examples 10.13 and 10.16 were unbiased.

As we have already indicated, the method of maximum likelihood can also be used for the simultaneous estimation of several parameters of a given population. In that case we must find the values of the parameters that jointly maximize the likelihood function.

**EXAMPLE 10.18**

If \( X_1, X_2, \ldots, X_n \) constitute a random sample of size \( n \) from a normal population with the mean \( \mu \) and the variance \( \sigma^2 \), find joint maximum likelihood estimates of these two parameters.

**Solution** Since the likelihood function is given by

\[
L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2\right)
\]

partial differentiation of \( \ln L(\mu, \sigma^2) \) with respect to \( \mu \) and \( \sigma^2 \) yields

\[
\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2) = -\frac{n}{2\sigma^2} \sum (x_i - \mu)
\]

and

\[
\frac{\partial}{\partial \sigma^2} \ln L(\mu, \sigma^2) = \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (x_i - \mu)^2
\]

Equating the first of these two partial derivatives to zero and solving for \( \mu \), we get

\[
\hat{\mu} = \frac{1}{n} \sum x_i = \bar{x}
\]

and equating the second of these partial derivatives to zero and solving for \( \sigma^2 \) after substituting \( \mu = \bar{x} \), we get

\[
\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2
\]

It should be observed that we did not show that \( \hat{\sigma}^2 \) is a maximum likelihood estimate of \( \sigma^2 \), only that \( \hat{\sigma}^2 \) is a maximum likelihood estimate of \( \sigma^2 \). However, it can be shown (see reference at the end of this chapter) that maximum likelihood estimators have the **invariance property** that if \( \hat{\theta} \) is a maximum likelihood estimator
of \( \sigma \) and the function given by \( g(\theta) \) is continuous, then \( g(\theta) \) is also a maximum likelihood estimator of \( g(\theta) \). It follows that

\[
\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

which differs from \( \theta \) in that we divide by \( n \) instead of \( n - 1 \), is a maximum likelihood estimate of \( \sigma \).

In Examples 10.15, 10.16, and 10.58 we maximized the logarithm of the likelihood function instead of the likelihood function itself, but this is by no means necessary. It just so happened that it was convenient in each case.

**EXERCISES**

10.50. If \( X_1, X_2, \ldots, X_n \) constitute a random sample from a population with the mean \( \mu \) and the variance \( \sigma^2 \), use the method of moments to find estimators for \( \mu \) and \( \sigma^2 \).

10.51. If a random sample of size \( n \) from an exponential population, use the method of moments to find an estimator for the parameter \( \theta \).

10.52. Given a random sample of size \( n \) from a uniform population with \( a = 0 \), find an estimator for \( \theta \) by the method of moments.

10.53. Given a random sample of size \( n \) from a Poisson population, use the method of moments to obtain an estimator for the parameter \( \lambda \).

10.54. Given a random sample of size \( n \) from a normal population with \( \beta = 1 \), use the method of moments to find a formula for estimating the parameter \( \alpha \).

10.55. If \( X_1, X_2, \ldots, X_n \) constitute a random sample of size \( n \) from a population given by

\[
f(x; \theta) = \begin{cases} 
\frac{2(\theta - x)}{\theta^2} & \text{for } 0 < x < \theta \\
0 & \text{elsewhere}
\end{cases}
\]

find an estimator for \( \theta \) by the method of moments.

10.56. If \( X_1, X_2, \ldots, X_n \) constitute a random sample of size \( n \) from a population given by

\[
f(x; \sigma^2) = \begin{cases} 
\frac{1}{\sigma \sqrt{2\pi}} e^{-x^2/2\sigma^2} & \text{for } x > 0 \\
0 & \text{elsewhere}
\end{cases}
\]

find estimators for \( \delta \) and \( \theta \) by the method of moments. This distribution is sometimes referred to as the two-parameter exponential distribution, and for \( \delta = 1 \) it is the distribution of Example 10.3.

10.57. Given a random sample of size \( n \) from a continuous uniform population, use the method of moments to find estimators for the parameters \( \alpha \) and \( \beta \).

10.58. Consider \( n \) independent random variables having identical binomial distributions with the parameters \( \theta \) and \( n \). If \( n \) of them take on the value 0, \( n \) take on the value 1, \( n \) take on the value 2, and \( n \) take on the value 3, use the method of moments to find a formula for estimating \( \theta \).

10.59. Use the method of maximum likelihood to rework Exercise 10.53.

10.60. Use the method of maximum likelihood to rework Exercise 10.54.

10.61. If \( X_1, X_2, \ldots, X_n \) constitute a random sample of size \( n \) from a gamma population with \( \alpha = 2 \), use the method of maximum likelihood to find a formula for estimating \( \beta \).

**10.9 BAYESIAN ESTIMATION**

So far we have assumed in this chapter that the parameters that we want to estimate are unknown constants; in Bayesian estimation the parameters are looked upon as random variables having prior distributions, usually reflecting the strength of one's belief about the possible values that they can assume. In Section 9.6 we already met a problem of Bayesian estimation: the parameter was that of a uniform density and its prior distribution was a gamma distribution.

Some of the concepts and language used in this section were introduced in Chapter 9, the optional chapter on decision theory.
The main problem of Bayesian estimation is that of combining prior feelings about a parameter with direct sample evidence. In Example 9.9, we accomplished this by determining \( \psi(\theta|x) \), the conditional density of \( \theta \) given \( x = x \). In contrast to the prior distribution of \( \theta \), this conditional distribution (which also reflects the direct sample evidence) is called the posterior distribution of \( \theta \). In general, if \( h(\theta) \) is the value of the prior distribution of \( \theta \) at \( \theta \) and \( w \) is the value that \( \theta \) is to convey with direct sample evidence about \( \theta \), for instance, the value of a statistic \( W = w(X_1, X_2, \ldots, X_n) \), we determine the posterior distribution of \( \theta \) by means of the formula

\[
\psi(\theta|w) = \frac{f(\theta, w)}{g(w)} = \frac{h(\theta),f(\theta|w)}{g(w)}
\]

Here \( f(w|\theta) \) is the value of the sampling distribution of \( W \) given \( \theta = \theta \) at \( w \), \( f(\theta, w) \) is the value of the joint distribution of \( \theta \) and \( W \) at \( \theta \) and \( w \), and \( g(w) \) is the value of the marginal distribution of \( W \) at \( w \). Note that the preceding formula for \( \psi(\theta|w) \) is, in fact, an extension of Bayes' theorem, Theorem 2.13, to the continuous case. Hence, the term "Bayesian estimation." Once the posterior distribution of a parameter has been obtained, it can be used to make estimates as in Example 9.9, or it can be used to make probability statements about the parameter, as will be illustrated in Example 9.20. Although the method we have described has extensive applications, we shall limit our discussion here to inferences about the parameter \( \theta \) of a binomial population and the mean of a normal population; inferences about the parameter of a Poisson population are treated in Example 9.77.

**Theorem 10.5.** If \( X \) is a binomial random variable and the prior distribution of \( \theta \) is a beta distribution with parameters \( a \) and \( b \), then the posterior distribution of \( \theta \) given \( x = x \) is a beta distribution with parameters \( x + a \) and \( n - x + b \).

**Proof.** For \( 0 < \theta < 1 \) we have

\[
f(x|\theta) = \binom{n}{x} \theta^x(1-\theta)^{n-x} \text{ for } x = 0, 1, 2, \ldots, n
\]

\[
h(\theta) = \frac{\Gamma(a + b)}{\Gamma(a) \cdot \Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \text{ for } 0 < \theta < 1
\]

and hence

\[
f(\theta|x) = \frac{\Gamma(a + b)}{\Gamma(a) \cdot \Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \cdot \left( \frac{n}{x} \right) \theta^x(1-\theta)^{n-x} = \frac{n}{x} \cdot \frac{\Gamma(a + b)}{\Gamma(a) \cdot \Gamma(b)} \theta^{a+x-1}(1-\theta)^{n-x+b-1}
\]

for \( 0 < \theta < 1 \) and \( x = 0, 1, 2, \ldots, n \), and \( f(\theta, x) = 0 \) elsewhere. To obtain the marginal density of \( X \), let us make use of the fact that the integral of the beta density from \( 0 \) to 1 equals 1; that is,

\[
\int_0^1 \theta^{a+x-1}(1-\theta)^{n-x+b-1} d\theta = \frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a + b)}
\]

Thus, we get

\[
g(x) = \frac{n}{x} \cdot \frac{\Gamma(a + b)}{\Gamma(a) \cdot \Gamma(b)} \frac{\Gamma(a + x) \cdot \Gamma(n - x + b)}{\Gamma(n + a + b)}
\]

for \( x = 0, 1, \ldots, n \), and hence

\[
\psi(\theta|x) = \frac{\Gamma(a + n + b)}{\Gamma(a + x) \cdot \Gamma(n - x + b + 1)} \theta^{a+x-1}(1-\theta)^{n-x+b-1}
\]

for \( 0 < \theta < 1 \), and \( \psi(\theta|x) = 0 \) elsewhere. As can be seen by inspection, this is a beta density with the parameters \( x + a \) and \( n - x + b \). \( \square \)

To make use of this theorem, let us refer to the fact that (under very general conditions) the mean of the posterior distribution minimizes the Bayes risk when the loss function is quadratic. That is, when the loss function is given by

\[
L(c) = c^2
\]

where \( c \) is a positive constant. Note that this is the loss function that we used in Example 9.9. Since the posterior distribution of \( \theta \) is a beta distribution with parameters \( x + a \) and \( n - x + b \), it follows from Theorem 6.5 that

\[
E(\theta|x) = \frac{x + a}{a + b + n}
\]

is a value of an estimator of \( \theta \) that minimizes the Bayes risk when the loss function is quadratic and the prior distribution of \( \theta \) is of the given form.

**Example 10.19.**

Find the mean of the posterior distribution as an estimate of the "true" probability of a success if 42 successes are obtained in 120 binomial trials and the prior distribution of \( \theta \) is a beta distribution with \( a = \beta = 40 \).

**Solution.** Substituting \( x = 42, n = 120, a = 40, \) and \( \beta = 40 \) into the formula for \( E(\theta|x) \), we get

\[
E(\theta|42) = \frac{42 + 40}{40 + 40 + 120} = 0.41
\]

Note that without knowledge of the prior distribution of \( \theta \), the minimum variance unbiased estimate of \( \theta \) (see Exercise 10.14) would be the sample proportion

\[
\hat{\theta} = \frac{x}{n} = \frac{42}{120} = 0.35
\]
Theorem 10.6. Let \( \bar{X} \) be the mean of a random sample of size \( n \) from a normal population with the known variance \( \sigma^2 \) and the prior distribution of \( \mu \) (capital Greek \( \mu \)) is a normal distribution with the mean \( \mu_0 \) and the variance \( \sigma_0^2 \), then the posterior distribution of \( \mu \) given \( \bar{X} = \bar{x} \) is a normal distribution with the mean \( \mu_1 \) and the variance \( \sigma_1^2 \), where:

\[
\mu_1 = \frac{n\bar{x}^2 + \mu \sigma^2}{\sigma^2 + \sigma_0^2} \quad \text{and} \quad \sigma_1^2 = \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}.
\]

**Proof.** For \( \mathcal{M} = \mu \) we have

\[
f(\bar{x}|\mu) = \frac{\sqrt{n}}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}} \left( \frac{\bar{x} - \mu}{\sigma} \right)^2 \quad \text{for} \quad -\infty < \bar{x} < \infty.
\]

According to Theorem 8.4, and

\[
h(\mu) = \frac{1}{\sigma_0 \sqrt{2\pi}} e^{-\frac{1}{2}} \left( \frac{\mu - \mu_0}{\sigma_0} \right)^2 \quad \text{for} \quad -\infty < \mu < \infty
\]

so that

\[
\psi(\mu|\bar{x}) = \frac{h(\mu) \cdot f(\bar{x}|\mu)}{f(\bar{x})} = \frac{\sqrt{n}}{\sigma \sigma_0 \sqrt{2\pi}} e^{-\frac{1}{2}} \left( \frac{\bar{x} - \mu}{\sigma} \right)^2 \cdot \left( \frac{\mu - \mu_0}{\sigma_0} \right)^2 \quad \text{for} \quad -\infty < \mu < \infty
\]

Now, if we collect powers of \( \mu \) in the exponent of \( e \), we get

\[
-\frac{1}{2} \left( \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 + \left( \frac{n\bar{x}^2}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \mu - \frac{1}{2} \left( \frac{n\bar{x}^2}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)
\]

and if we let

\[
\frac{1}{\sigma_1^2} = \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \quad \text{and} \quad \mu_1 = \frac{n\bar{x}^2 + \mu \sigma^2}{n\sigma^2 + \sigma_0^2}
\]

factor out \(-\frac{1}{2\sigma_1^2}\), and complete the square, the exponent of \( e \) in the expression for \( \psi(\mu|\bar{x}) \) becomes

\[
-\frac{1}{2\sigma_1^2} \left[ \mu - \mu_1 \right]^2 + R
\]

where \( R \) involves \( n, \bar{x}, \mu_0, \sigma, \) and \( \sigma_0 \) but not \( \mu \). Thus, the posterior distribution of \( \mu \) becomes

\[
\psi(\mu|\bar{x}) = \frac{\sqrt{n}}{\sigma \sigma_0 \sqrt{2\pi}} e^{-\frac{1}{2}} \left( \frac{\mu - \mu_1}{\sigma_1} \right)^2 \quad \text{for} \quad -\infty < \mu < \infty
\]

which is easily identified as a normal distribution with the mean \( \mu_1 \) and the variance \( \sigma_1^2 \). Hence, it can be written as

\[
\psi(\mu|\bar{x}) = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{1}{2}} \left( \frac{\mu - \mu_1}{\sigma_1} \right)^2 \quad \text{for} \quad -\infty < \mu < \infty
\]

where \( \mu_1 \) and \( \sigma_1 \) are defined on page 346. Note that we did not have to determine \( f(\bar{x}) \) as it was absorbed in the constant in the final result.

**Example 10.20**

A distributor of soft-drink vending machines feels that in a supermarket one of his machines will sell on the average \( \mu_0 = 4.738 \) drinks per week. Of course, the mean will vary somewhat from market to market, and the distributor feels that this variation is measured by the standard deviation \( \sigma_0 = 13.4 \). As far as a machine placed in a particular market is concerned, the number of drinks sold will vary from week to week, and this variation is measured by the standard deviation \( \sigma_1 = 42.5 \). If one of the distributor's machines puts into a new supermarket averaged \( \bar{x} = 692 \) during the first 10 weeks, what is the probability (the distributor's personal probability) that for this market the value of \( \mu \) is actually between 700 and 720?

**Solution**

Assuming that the population sampled is approximately normal and that 9 is reasonable to treat the prior distribution of \( \mu \) as a normal distribution with the mean \( \mu_0 \) and the standard deviation \( \sigma_0 = 13.4 \), we find that substitution into the two formulas of Theorem 10.6 yields

\[
\mu_1 = \frac{10.629(4.738^2 + 738(42.5)^2)}{10(4.738^2 + 42.5^2)} - 715
\]

and

\[
\frac{1}{\sigma_1^2} = \frac{10}{42.5^2} + \frac{1}{(13.4)^2} = 0.0011
\]

so that \( \sigma_1^2 = 90.8 \) and \( \sigma_1 = 9.5 \). Now, the answer to our question is given by the area of the shaded region of Figure 10.1, that is, the area under the standard normal curve between

\[
z = \frac{700 - 715}{9.5} \approx -1.58 \quad \text{and} \quad z = \frac{720 - 715}{9.5} = 0.53
\]

Thus, the probability that the value of \( \mu \) is between 700 and 720 is 0.0440 + 0.2099 = 0.0448, or approximately 0.645.

**Exercises**

10.4. Making use of the results of Exercise 6.29, show that the mean of the posterior distribution of \( \theta \) given on page 345 can be written as

\[
E(\theta|\mu) = \frac{\mu + (\mu - \mu_0) \theta}{\theta + 1}
\]
that is, as a weighted mean of $\hat{\theta}$ and $\theta_0$, where $\theta_0$ and $\sigma_0^2$ are the mean and the variance of the prior beta distribution of $\theta$ and

$$w = \frac{n \theta_0 + n \hat{\theta}}{n + n \theta_0} - 1$$

10.76. Show that the mean of the posterior distribution of $\mu$ given $w$ in Theorem 10.6 can be written as

$$\mu_1 = w \cdot \mu + (1 - w) \cdot \mu_0,$$

that is, as a weighted mean of $\mu$ and $\mu_0$, where

$$w = \frac{\sigma^2}{\sigma^2 + \sigma_0^2}$$

10.77. If $X$ has a Poisson distribution and the prior distribution of its parameter $\lambda$ (capital Greek lambda) is a gamma distribution with the parameters $\alpha$ and $\beta$, show that

(a) the posterior distribution of $\lambda$ given $X = x$ is a gamma distribution with the parameters $\alpha + x$ and $\beta + 1$

(b) the mean of the posterior distribution of $\lambda$ is

$$\mu_1 = \frac{\alpha x + \beta + 1}{\alpha + x}.$$

10.10 The Theory in Practice

The sample mean, $\bar{x}$, is most frequently used to estimate the mean of a distribution from a random sample taken from that distribution. It has been shown to be the minimum-variance unbiased estimator as well as a sufficient estimator for the mean of a normal distribution. It is at least asymptotically unbiased as an estimator for the mean of most frequently encountered distributions.

In spite of these desirable properties of the sample mean as an estimator for a population mean, we know that the sample mean will never equal the population mean. Let us examine the error we make when using $\bar{x}$ to estimate $\mu$, $E = |\bar{x} - \mu|$. If the sample size, $n$, is large, the quantity

$$\frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}}$$

is a value of a random variable having approximately the standard normal distribution. Thus, we can state with probability $1 - \alpha$ that

$$\frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

or

$$E \leq z_{\alpha/2} \cdot \sigma/\sqrt{n}.$$
Another consideration related to the accuracy of a sample estimate deals with the concept of sampling bias. Sampling bias occurs when a sample is chosen that does not accurately represent the population from which it is taken. For example, a national poll based on automobile registrations in each of the 50 states probably is biased, because it omits people who do not own cars. Such people may well have different opinions than those who do. A sample of product stored on shelves in a warehouse is likely to be biased if all units of the product in the sample were selected from the bottom shelf. Ambient conditions, such as temperature and humidity may well have affected the top-shelf units differently than those on the bottom shelf.

The mean square error defined in page 225 can be viewed as the expected squared error loss encountered when we estimate the parameter $\theta$ with the point estimator $\hat{\theta}$. We can write

$$\text{MSE}(\hat{\theta}) = (\hat{\theta} - \theta)^2$$

$$= E[(\hat{\theta} - \theta) + (\hat{\theta} - \theta)]^2$$

$$= E[(\hat{\theta} - \theta)]^2 + E[(\hat{\theta} - \theta)] + 2E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)]$$

The first term of the cross product, $E[\hat{\theta} - \theta] = E\hat{\theta} - \theta = 0$, and we are left with

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = [\text{Bias}^2]$$

The first term is readily seen to be the variance of $\hat{\theta}$ and the second term is the square of the bias, the difference between the expected value of the estimate of the parameter $\theta$ and its true value. Thus, we can write

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}^2$$

While it is possible to estimate $E\hat{\theta}$, variance of $\hat{\theta}$ in most applications, the sampling bias usually is unknown. Great care should be taken to avoid, or at least minimize sampling bias, for it can be much greater than the sampling variance $\sigma^2$. This can be done by carefully calibrating all instruments to be used in measuring the sample units, by eliminating human sub-activity as much as possible, and by assuring that the method of sampling is appropriately randomized over the entire population for which sampling estimates are to be made. These and other related issues are more thoroughly discussed in the book by Hogg and Tanis, referenced at the end of this chapter.

**APPLIED EXERCISES**

10.76. Random samples of size $n$ are taken from normal populations with the mean $\mu$ and the variances $\sigma_1^2 = 4$ and $\sigma_2^2 = 9$. If $\bar{X}_1 = 26.0$ and $\bar{X}_2 = 32.5$, estimate $\mu$ using the estimator of part (b) of Exercise 10.21.

10.79. Independent random samples of size $n_1$ and $n_2$ are taken from normal populations with different variances, $\sigma_1^2 = 8$ and $\sigma_2^2 = 16$. If $\bar{X}_1 = 25$, $n_1 = 30$, $\bar{X}_2 = 27.6$, and $n_2 = 38.1$, estimate $\mu$ using the estimator of Exercise 10.23.

4.80. A country's military intelligence knows that an enemy built certain new tanks numbered serially from 1 to 10. In three of these tanks are captured and their serial numbers are 230, 238, and 153, use the estimator of part (b) of Exercise 10.12 to estimate $\mu$.

10.81. On 12 days selected at random, a city's consumption of electricity was 6.4, 4.5, 10.8, 7.2, 6.8, 4.9, 3.9, 13.6, 4.8, 7.0, 8.8, and 5.4 million kilowatt-hours. Assuming that these data may be looked upon as a random sample from a Pareto population, use the estimators obtained in Exercise 10.14 to estimate the parameters $\alpha$ and $\beta$.

10.82. The size of an animal's population is sometimes estimated by the capture-recapture method. In this method, $n_1$ of the animals are captured in the area $X$ of them are found to be tagged. Later, $n_2$ of the animals are captured, the total number of animals of the given kind in the area under consideration. If $n_2 = 35$ and only one of them is found to be tagged, estimate $N$.

10.83. Certain radial sites had useful lives of 35.2, 41.0, 44.7, 66.9, and 41.50 miles. Assuming that these data can be looked upon as a random sample from an exponential population, the estimator obtained in Exercise 10.51 to estimate the parameter $\theta$.

10.84. Among six measurements of the boiling point of a silicon compound, the size of the error was 0.07, 0.03 0.14, 0.04, 0.08, and 0.06$^\circ$C. Assuming that these data can be looked upon as a random sample from the population of Exercising 10.55, use the estimator obtained there by the method of moments to estimate the parameter $\theta$.

10.85. Not counting the ones that failed immediately, certain light bulbs had useful lives of 415, 433, 459, 451, 466, 419, 479, 403, 562, 422, 475, and 439 hours. Assuming that these data can be looked upon as a random sample from a two-parameter exponential population, use the estimators obtained in Exercise 10.56 to estimate the parameters $\alpha$ and $\beta$.

10.86. Work Exercise 10.85 using the estimators obtained in Exercise 10.66 by the method of maximum likelihood.

10.87. Data collected over a number of years show that when a broker called a random sample of eight of her clients, she got a busy signal 6.5, 10.6, 8.1, 4.1, 9.3, 11.5, and 7.4, and 3.7 percent of the time. Assuming that these figures can be looked upon as a random sample from a continuous uniform population, use the estimators obtained in Exercise 10.57 to estimate the parameters $\alpha$ and $\beta$.

10.88. Work Exercise 10.87 using the estimators obtained in Exercise 10.67.

10.89. Every time Mr. Jones goes to the race track he bets on three races. In a random sample of 20 of his visits to the race track, he lost all his bets 11 times, won once seven times, and won twice on two occasions. If $\alpha$ is the probability that he will win any one of his bets, estimate $\alpha$ by using the maximum likelihood estimator obtained in Exercise 10.68.

10.90. In a random sample of the teachers in a large school district, their annual salaries were $23,900, 31,200, 26,400, 24,800, 35,700, 24,500, 29,200, 36,200, 22,700,$ and $25,100. Assuming that these data can be looked upon as a random sample from a Pareto population, use the estimator obtained in Exercise 10.65 to estimate the parameter $\alpha$.

10.91. On 20 very cold days, a farmer got her tractor started on the first, third, fifth, first, second, first, third, seventh, second, fourth, fourth, eighth, first, third, sixth, and second try. Assuming that these data can be looked
upon as a random sample from a geometric population, estimate its parameter
by either of the methods of Exercise 10.63.

10.92. The I.Q.'s of 100 teenagers belonging to one ethnic group are 98, 114, 105, 101, 123,
117, 106, 92, 110, and 108, whereas those of six teenagers belonging to another
ethnic group are 122, 165, 99, 126, 114, and 58. Assuming that these data can
be looked upon as independent random samples from normal populations with
the means $\mu_1$ and $\mu_2$ and the common variance $\sigma^2$, estimate these parameters
by means of the maximum likelihood estimators obtained in Exercise 10.71.

10.93. The output of a certain integrated-circuit production line is checked daily by
inspecting a sample of 100 units. Over a long period of time, the process has
maintained a yield of 60 percent, this is, a proportion defective of 20 percent,
and the variation of the proportion defective from day to day is measured
by a standard deviation of 0.04. If on a certain day the sample contains 38
defectives, find the mean of the posterior distribution of $\theta$ as an estimate
of this day’s proportion defective. Assume that the prior distribution of $\theta$ is a beta
distribution.

10.94. Records of a university (collected every many years) show that on the average
74 percent of all incoming freshmen have I.Q.'s of at least 115. Of course, the
percentage varies somewhat from year to year, and this variation is measured
by a standard deviation of 3 percent. If a sample of 30 freshmen entering the
university in 2003 showed that only 18 of them have I.Q.'s of at least 115, estimate
the true proportion of students with I.Q.'s of at least 115 in that freshman class using
(a) only the prior information;
(b) only the direct information;
(c) the result of Exercise 10.74 to combine the prior information with the direct
information.

10.95. With reference to Example 10.20, find $P(72 < M < 72.5 | \mu = 65.2)$.

10.96. A history professor is making up a final examination that is to be given to a very
large group of students. His feelings about the average grade that they should
get is expressed subjectively by a normal distribution with the mean $\mu_0 = 65.2$
and the standard deviation $\sigma_0 = 15$.
(a) What prior probability does the professor assign to the actual average grade
being somewhere on the interval from 63.0 to 68.0?
(b) What posterior probability would he assign to the event if the examination
is tried on a random sample of 40 students whose grades have a mean of
72.9 and a standard deviation of 7.4? Use $\gamma = 7.4$ as an estimate of $\sigma$.

10.97. An office manager feels that for a certain kind of business the daily number
of incoming telephone calls is a random variable having a Poisson distribution
whose parameter has a prior gamma distribution with $\alpha = 50$ and $\beta = 2$. Being
told that one such business had 113 incoming calls on a given day, what would
be the estimate of that particular business’s average daily number of incoming
calls if she considers
(a) only the prior information;
(b) only the direct information;
(c) both kinds of information and the theory of Exercise 10.77?