7

Point Estimation of Parameters

CHAPTER OUTLINE

LEARNING OBJECTIVES

After careful study of this chapter you should be able to do the following:

1. Explain the general concepts of estimating the parameters of a population or a probability distribution
2. Explain important properties of point estimators, including bias, variance, and mean square error
3. Know how to construct point estimators using the method of moments and the method of maximum likelihood
4. Know how to compute and explain the precision with which a parameter is estimated
5. Understand the central limit theorem
6. Explain the important role of the normal distribution as a sampling distribution

CD MATERIAL

7. Use bootstrapping to find the standard error of a point estimate
8. Know how to construct a point estimator using the Bayesian approach

Answers for most odd numbered exercises are at the end of the book. Answers to exercises whose numbers are surrounded by a box can be accessed in the e-Text by clicking on the box. Complete worked solutions to certain exercises are also available in the e-Text. These are indicated in the Answers to Selected Exercises section by a box around the exercise number. Exercises are also available for some of the text sections that appear on CD only. These exercises may be found within the e-Text immediately following the section they accompany.

7-1 INTRODUCTION

The field of statistical inference consists of those methods used to make decisions or to draw conclusions about a **population.** These methods utilize the information contained in a **sample** from the population in drawing conclusions. This chapter begins our study of the statistical methods used for inference and decision making.

Statistical inference may be divided into two major areas: **parameter estimation** and **hypothesis testing.** As an example of a parameter estimation problem, suppose that a structural engineer is analyzing the tensile strength of a component used in an automobile chassis. Since variability in tensile strength is naturally present between the individual components because of differences in raw material batches, manufacturing processes, and measurement procedures (for example), the engineer is interested in estimating the mean tensile strength of the components. In practice, the engineer will use sample data to compute a number that is in some sense a reasonable value (or guess) of the true mean. This number is called a **point estimate.** We will see that it is possible to establish the precision of the estimate.

Now consider a situation in which two different reaction temperatures can be used in a chemical process, say t_1 and t_2. The engineer conjectures that t_1 results in higher yields than does t_2. Statistical hypothesis testing is a framework for solving problems of this type. In this case, the hypothesis would be that the mean yield using temperature t_1 is greater than the mean yield using temperature t_2. Notice that there is no emphasis on estimating yields; instead, the focus is on drawing conclusions about a stated hypothesis.

Suppose that we want to obtain a point estimate of a population parameter. We know that before the data is collected, the observations are considered to be random variables, say $X_1, X_2, \ldots, X_n$. Therefore, any function of the observation, or any **statistic,** is also a random variable. For example, the sample mean $\overline{X}$ and the sample variance S^2 are statistics and they are also random variables.

Since a statistic is a random variable, it has a probability distribution. We call the probability distribution of a statistic a **sampling distribution.** The notion of a sampling distribution is very important and will be discussed and illustrated later in the chapter.

When discussing inference problems, it is convenient to have a general symbol to represent the parameter of interest. We will use the Greek symbol θ (theta) to represent the parameter. The objective of point estimation is to select a single number, based on sample data, that is the most plausible value for θ. A numerical value of a sample statistic will be used as the point estimate.

In general, if X is a random variable with probability distribution $f(x)$, characterized by the unknown parameter θ, and if $X_1, X_2, \ldots, X_n$ is a random sample of size n from X, the statistic $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$ is called a **point estimator** of θ. Note that $\hat{\Theta}$ is a random variable because it is a function of random variables. After the sample has been selected, $\hat{\Theta}$ takes on a particular numerical value $\hat{\theta}$ called the **point estimate** of θ.

Definition

> A **point estimate** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$. The statistic $\hat{\Theta}$ is called the **point estimator**.

As an example, suppose that the random variable X is normally distributed with an unknown mean μ. The sample mean is a point estimator of the unknown population mean μ. That is, $\hat{\mu} = \overline{X}$. After the sample has been selected, the numerical value $\bar{x}$ is the point estimate of μ. Thus, if $x_1 = 25$, $x_2 = 30$, $x_3 = 29$, and $x_4 = 31$, the point estimate of μ is

$$\bar{x} = \frac{25 + 30 + 29 + 31}{4} = 28.75$$

Similarly, if the population variance σ^2 is also unknown, a point estimator for σ^2 is the sample variance S^2, and the numerical value $s^2 = 6.9$ calculated from the sample data is called the point estimate of σ^2.

Estimation problems occur frequently in engineering. We often need to estimate

- The mean μ of a single population
- The variance σ^2 (or standard deviation σ) of a single population
- The proportion p of items in a population that belong to a class of interest
- The difference in means of two populations, $\mu_1 - \mu_2$
- The difference in two population proportions, $p_1 - p_2$

Reasonable point estimates of these parameters are as follows:

- For μ, the estimate is $\hat{\mu} = \bar{x}$, the sample mean.
- For σ^2, the estimate is $\hat{\sigma}^2 = s^2$, the sample variance.
- For p, the estimate is $\hat{p} = x/n$, the sample proportion, where x is the number of items in a random sample of size n that belong to the class of interest.
- For $\mu_1 - \mu_2$, the estimate is $\hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2$, the difference between the sample means of two independent random samples.
- For $p_1 - p_2$, the estimate is $\hat{p}_1 - \hat{p}_2$, the difference between two sample proportions computed from two independent random samples.

We may have several different choices for the point estimator of a parameter. For example, if we wish to estimate the mean of a population, we might consider the sample mean, the sample median, or perhaps the average of the smallest and largest observations in the sample as point estimators. In order to decide which point estimator of a particular parameter is the best one to use, we need to examine their statistical properties and develop some criteria for comparing estimators.

7-2 GENERAL CONCEPTS OF POINT ESTIMATION

7-2.1 Unbiased Estimators

An estimator should be "close" in some sense to the true value of the unknown parameter. Formally, we say that $\hat{\Theta}$ is an unbiased estimator of θ if the expected value of $\hat{\Theta}$ is equal to θ. This is equivalent to saying that the mean of the probability distribution of $\hat{\Theta}$ (or the mean of the sampling distribution of $\hat{\Theta}$) is equal to θ.

Definition

The point estimator $\hat{\Theta}$ is an **unbiased estimator** for the parameter θ if

$$E(\hat{\Theta}) = \theta \tag{7-1}$$

If the estimator is not unbiased, then the difference

$$E(\hat{\Theta}) - \theta \tag{7-2}$$

is called the **bias** of the estimator $\hat{\Theta}$.

When an estimator is unbiased, the bias is zero; that is, $E(\hat{\Theta}) - \theta = 0$.

EXAMPLE 7-1

Suppose that X is a random variable with mean μ and variance σ^2. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the population represented by X. Show that the sample mean $\overline{X}$ and sample variance S^2 are unbiased estimators of μ and σ^2, respectively.

First consider the sample mean. In Equation 5.40a in Chapter 5, we showed that $E(\overline{X}) = \mu$. Therefore, the sample mean $\overline{X}$ is an unbiased estimator of the population mean μ.

Now consider the sample variance. We have

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

The last equality follows from Equation 5-37 in Chapter 5. However, since $E(X_i^2) = \mu^2 + \sigma^2$ and $E(\overline{X}^2) = \mu^2 + \sigma^2/n$, we have

$$\begin{aligned}
E(S^2) &= \frac{1}{n-1}\left[\sum_{i=1}^{n}(\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n)\right] \\
&= \frac{1}{n-1}(n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2) \\
&= \sigma^2
\end{aligned}$$

Therefore, the sample variance S^2 is an unbiased estimator of the population variance σ^2.

Although S^2 is unbiased for σ^2, S is a biased estimator of σ. For large samples, the bias is very small. However, there are good reasons for using S as an estimator of σ in samples from normal distributions, as we will see in the next three chapters when are discuss confidence intervals and hypothesis testing.

Sometimes there are several unbiased estimators of the sample population parameter. For example, suppose we take a random sample of size $n = 10$ from a normal population and obtain the data $x_1 = 12.8$, $x_2 = 9.4$, $x_3 = 8.7$, $x_4 = 11.6$, $x_5 = 13.1$, $x_6 = 9.8$, $x_7 = 14.1$, $x_8 = 8.5$, $x_9 = 12.1$, $x_{10} = 10.3$. Now the sample mean is

$$\begin{aligned}\bar{x} &= \frac{12.8 + 9.4 + 8.7 + 11.6 + 13.1 + 9.8 + 14.1 + 8.5 + 12.1 + 10.3}{10} \\ &= 11.04\end{aligned}$$

the sample median is

$$\tilde{x} = \frac{10.3 + 11.6}{2} = 10.95$$

and a 10% trimmed mean (obtained by discarding the smallest and largest 10% of the sample before averaging) is

$$\begin{aligned}\bar{x}_{tr(10)} &= \frac{8.7 + 9.4 + 9.8 + 10.3 + 11.6 + 12.1 + 12.8 + 13.1}{8} \\ &= 10.98\end{aligned}$$

We can show that all of these are unbiased estimates of μ. Since there is not a unique unbiased estimator, we cannot rely on the property of unbiasedness alone to select our estimator. We need a method to select among unbiased estimators. We suggest a method in Section 7-2.3.

7-2.2 Proof That S is a Biased Estimator of σ (CD Only)

7-2.3 Variance of a Point Estimator

Suppose that $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are unbiased estimators of θ. This indicates that the distribution of each estimator is centered at the true value of θ. However, the variance of these distributions may be different. Figure 7-1 illustrates the situation. Since $\hat{\Theta}_1$ has a smaller variance than $\hat{\Theta}_2$, the estimator $\hat{\Theta}_1$ is more likely to produce an estimate close to the true value θ. A logical principle of estimation, when selecting among several estimators, is to choose the estimator that has minimum variance.

Definition

> If we consider all unbiased estimators of θ, the one with the smallest variance is called the **minimum variance unbiased estimator** (MVUE).

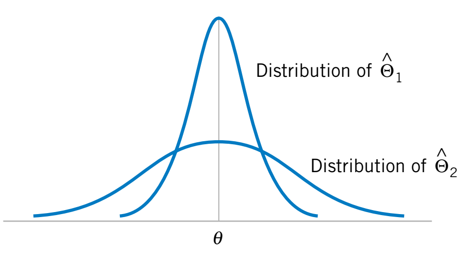


Figure 7-1 The sampling distributions of two unbiased estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$.

In a sense, the MVUE is most likely among all unbiased estimators to produce an estimate $\hat{\theta}$ that is close to the true value of θ. It has been possible to develop methodology to identify the MVUE in many practical situations. While this methodology is beyond the scope of this book, we give one very important result concerning the normal distribution.

Theorem 7-1

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and variance σ^2, the sample mean $\overline{X}$ is the MVUE for μ.

In situations in which we do not know whether an MVUE exists, we could still use a minimum variance principle to choose among competing estimators. Suppose, for example, we wish to estimate the mean of a population (not necessarily a *normal* population). We have a random sample of n observations $X_1, X_2, \ldots, X_n$ and we wish to compare two possible estimators for μ: the sample mean $\overline{X}$ and a single observation from the sample, say, X_i. Note that both $\overline{X}$ and X_i are unbiased estimators of μ; for the sample mean, we have $V(\overline{X}) = \sigma^2/n$ from Equation 5-40b and the variance of any observation is $V(X_i) = \sigma^2$. Since $V(\overline{X}) < V(X_i)$ for sample sizes $n \geq 2$, we would conclude that the sample mean is a better estimator of μ than a single observation X_i.

7-2.4 Standard Error: Reporting a Point Estimate

When the numerical value or point estimate of a parameter is reported, it is usually desirable to give some idea of the precision of estimation. The measure of precision usually employed is the standard error of the estimator that has been used.

Definition

The **standard error** of an estimator $\hat{\Theta}$ is its standard deviation, given by $\sigma_{\hat{\Theta}} = \sqrt{V(\hat{\Theta})}$. If the standard error involves unknown parameters that can be estimated, substitution of those values into $\sigma_{\hat{\Theta}}$ produces an **estimated standard error,** denoted by $\hat{\sigma}_{\hat{\Theta}}$.

Sometimes the estimated standard error is denoted by $s_{\hat{\Theta}}$ or $se(\hat{\Theta})$.

Suppose we are sampling from a normal distribution with mean μ and variance σ^2. Now the distribution of $\overline{X}$ is normal with mean μ and variance σ^2/n, so the standard error of $\overline{X}$ is

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

If we did not know σ but substituted the sample standard deviation S into the above equation, the estimated standard error of $\overline{X}$ would be

$$\hat{\sigma}_{\overline{X}} = \frac{S}{\sqrt{n}}$$

When the estimator follows a normal distribution, as in the above situation, we can be reasonably confident that the true value of the parameter lies within two standard errors of the estimate. Since many point estimators are normally distributed (or approximately so) for large n, this is a very useful result. Even in cases in which the point estimator is not normally distributed, we can state that so long as the estimator is unbiased, the estimate of the parameter will deviate from the true value by as much as four standard errors at most 6 percent of the time. Thus a very conservative statement is that the true value of the parameter differs from the point estimate by at most four standard errors. See Chebyshev's inequality in the CD only material.

EXAMPLE 7-2 An article in the *Journal of Heat Transfer* (Trans. ASME, Sec. C, 96, 1974, p. 59) described a new method of measuring the thermal conductivity of Armco iron. Using a temperature of 100°F and a power input of 550 watts, the following 10 measurements of thermal conductivity (in Btu/hr-ft-°F) were obtained:

$$41.60, 41.48, 42.34, 41.95, 41.86,$$
$$42.18, 41.72, 42.26, 41.81, 42.04$$

A point estimate of the mean thermal conductivity at 100°F and 550 watts is the sample mean or

$$\bar{x} = 41.924 \text{ Btu/hr-ft-°F}$$

The standard error of the sample mean is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$, and since σ is unknown, we may replace it by the sample standard deviation $s = 0.284$ to obtain the estimated standard error of $\bar{X}$ as

$$\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{0.284}{\sqrt{10}} = 0.0898$$

Notice that the standard error is about 0.2 percent of the sample mean, implying that we have obtained a relatively precise point estimate of thermal conductivity. If we can assume that thermal conductivity is normally distributed, 2 times the standard error is $2\hat{\sigma}_{\bar{X}} = 2(0.0898) = 0.1796$, and we are highly confident that the true mean thermal conductivity is with the interval 41.924 ± 0.1756, or between 41.744 and 42.104.

7-2.5 Bootstrap Estimate of the Standard Error (CD Only)

7-2.6 Mean Square Error of an Estimator

Sometimes it is necessary to use a biased estimator. In such cases, the mean square error of the estimator can be important. The **mean square error** of an estimator $\hat{\Theta}$ is the expected squared difference between $\hat{\Theta}$ and θ.

Definition

The **mean square error** of an estimator $\hat{\Theta}$ of the parameter θ is defined as

$$\text{MSE}(\hat{\Theta}) = E(\hat{\Theta} - \theta)^2 \tag{7-3}$$

The mean square error can be rewritten as follows:

$$\begin{aligned}\text{MSE}(\hat{\Theta}) &= E[\hat{\Theta} - E(\hat{\Theta})]^2 + [\theta - E(\hat{\Theta})]^2 \\ &= V(\hat{\Theta}) + (\text{bias})^2\end{aligned}$$

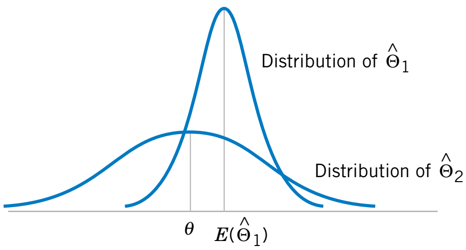


Figure 7-2 A biased estimator $\hat{\Theta}_1$ that has smaller variance than the unbiased estimator $\hat{\Theta}_2$.

That is, the mean square error of $\hat{\Theta}$ is equal to the variance of the estimator plus the squared bias. If $\hat{\Theta}$ is an unbiased estimator of θ, the mean square error of $\hat{\Theta}$ is equal to the variance of $\hat{\Theta}$.

The mean square error is an important criterion for comparing two estimators. Let $\hat{\Theta}_1$ and $\hat{\Theta}_2$ be two estimators of the parameter θ, and let MSE ($\hat{\Theta}_1$) and MSE ($\hat{\Theta}_2$) be the mean square errors of $\hat{\Theta}_1$ and $\hat{\Theta}_2$. Then the **relative efficiency** of $\hat{\Theta}_2$ to $\hat{\Theta}_1$ is defined as

$$\frac{\text{MSE}(\hat{\Theta}_1)}{\text{MSE}(\hat{\Theta}_2)} \tag{7-4}$$

If this relative efficiency is less than 1, we would conclude that $\hat{\Theta}_1$ is a more efficient estimator of θ than $\hat{\Theta}_2$, in the sense that it has a smaller mean square error.

Sometimes we find that biased estimators are preferable to unbiased estimators because they have smaller mean square error. That is, we may be able to reduce the variance of the estimator considerably by introducing a relatively small amount of bias. As long as the reduction in variance is greater than the squared bias, an improved estimator from a mean square error viewpoint will result. For example, Fig. 7-2 shows the probability distribution of a biased estimator $\hat{\Theta}_1$ that has a smaller variance than the unbiased estimator $\hat{\Theta}_2$. An estimate based on $\hat{\Theta}_1$ would more likely be close to the true value of θ than would an estimate based on $\hat{\Theta}_2$. Linear regression analysis (Chapters 11 and 12) is an area in which biased estimators are occasionally used.

An estimator $\hat{\Theta}$ that has a mean square error that is less than or equal to the mean square error of any other estimator, for all values of the parameter θ, is called an **optimal** estimator of θ. Optimal estimators rarely exist.

EXERCISES FOR SECTION 7-2

7-1. Suppose we have a random sample of size $2n$ from a population denoted by X, and $E(X) = \mu$ and $V(X) = \sigma^2$. Let

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

be two estimators of μ. Which is the better estimator of μ? Explain your choice.

7-2. Let $X_1, X_2, \ldots, X_7$ denote a random sample from a population having mean μ and variance σ^2. Consider the following estimators of μ:

$$\hat{\Theta}_1 = \frac{X_1 + X_2 + \cdots + X_7}{7}$$

$$\hat{\Theta}_2 = \frac{2X_1 - X_6 + X_4}{2}$$

(a) Is either estimator unbiased?
(b) Which estimator is best? In what sense is it best?

7-3. Suppose that $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are unbiased estimators of the parameter θ. We know that $V(\hat{\Theta}_1) = 10$ and $V(\hat{\Theta}_2) = 4$. Which estimator is best and in what sense is it best?

7-4. Calculate the relative efficiency of the two estimators in Exercise 7-2.

7-5. Calculate the relative efficiency of the two estimators in Exercise 7-3.

7-6. Suppose that $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are estimators of the parameter θ. We know that $E(\hat{\Theta}_1) = \theta$, $E(\hat{\Theta}_2) = \theta/2$, $V(\hat{\Theta}_1) = 10$, $V(\hat{\Theta}_2) = 4$. Which estimator is best? In what sense is it best?

7-7. Suppose that $\hat{\Theta}_1$, $\hat{\Theta}_2$, and $\hat{\Theta}_3$ are estimators of θ. We know that $E(\hat{\Theta}_1) = E(\hat{\Theta}_2) = \theta$, $E(\hat{\Theta}_3) \neq \theta$, $V(\hat{\Theta}_1) = 12$, $V(\hat{\Theta}_2) = 10$, and $E(\hat{\Theta}_3 - \theta)^2 = 6$. Compare these three estimators. Which do you prefer? Why?

7-8. Let three random samples of sizes $n_1 = 20$, $n_2 = 10$, and $n_3 = 8$ be taken from a population with mean μ and variance σ^2. Let S_1^2, S_2^2, and S_3^2 be the sample variances. Show that $S^2 = (20S_1^2 + 10S_2^2 + 8S_3^2)/38$ is an unbiased estimator of σ^2.

7-9. (a) Show that $\sum_{i=1}^{n} (X_i - \overline{X})^2/n$ is a biased estimator of σ^2.
(b) Find the amount of bias in the estimator.
(c) What happens to the bias as the sample size n increases?

7-10. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with mean μ and variance σ^2.
(a) Show that $\overline{X}^2$ is a biased estimator for μ^2.
(b) Find the amount of bias in this estimator.
(c) What happens to the bias as the sample size n increases?

7-11. Data on pull-off force (pounds) for connectors used in an automobile engine application are as follows: 79.3, 75.1, 78.2, 74.1, 73.9, 75.0, 77.6, 77.3, 73.8, 74.6, 75.5, 74.0, 74.7, 75.9, 72.9, 73.8, 74.2, 78.1, 75.4, 76.3, 75.3, 76.2, 74.9, 78.0, 75.1, 76.8.
(a) Calculate a point estimate of the mean pull-off force of all connectors in the population. State which estimator you used and why.
(b) Calculate a point estimate of the pull-off force value that separates the weakest 50% of the connectors in the population from the strongest 50%.
(c) Calculate point estimates of the population variance and the population standard deviation.
(d) Calculate the standard error of the point estimate found in part (a). Provide an interpretation of the standard error.
(e) Calculate a point estimate of the proportion of all connectors in the population whose pull-off force is less than 73 pounds.

7-12. Data on oxide thickness of semiconductors are as follows: 425, 431, 416, 419, 421, 436, 418, 410, 431, 433, 423, 426, 410, 435, 436, 428, 411, 426, 409, 437, 422, 428, 413, 416.
(a) Calculate a point estimate of the mean oxide thickness for all wafers in the population.
(b) Calculate a point estimate of the standard deviation of oxide thickness for all wafers in the population.
(c) Calculate the standard error of the point estimate from part (a).
(d) Calculate a point estimate of the median oxide thickness for all wafers in the population.
(e) Calculate a point estimate of the proportion of wafers in the population that have oxide thickness greater than 430 angstrom.

7-13. $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean μ and variance σ^2. Let $X_{\min}$ and $X_{\max}$ be the smallest and largest observations in the sample.
(a) Is $(X_{\min} + X_{\max})/2$ an unbiased estimate for μ?
(b) What is the standard error of this estimate?
(c) Would this estimate be preferable to the sample mean $\overline{X}$?

7-14. Suppose that X is the number of observed "successes" in a sample of n observations where p is the probability of success on each observation.
(a) Show that $\hat{P} = X/n$ is an unbiased estimator of p.
(b) Show that the standard error of $\hat{P}$ is $\sqrt{p(1-p)/n}$. How would you estimate the standard error?

7-15. $\overline{X}_1$ and S_1^2 are the sample mean and sample variance from a population with mean μ and variance σ_2^2. Similarly, $\overline{X}_2$ and S_2^2 are the sample mean and sample variance from a second independent population with mean μ_1 and variance σ_2^2. The sample sizes are n_1 and n_2, respectively.
(a) Show that $\overline{X}_1 - \overline{X}_2$ is an unbiased estimator of $\mu_1 - \mu_2$.
(b) Find the standard error of $\overline{X}_1 - \overline{X}_2$. How could you estimate the standard error?

7-16. Continuation of Exercise 7-15. Suppose that both populations have the same variance; that is, $\sigma_1^2 = \sigma_2^2 = \sigma_2$. Show that

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

is an unbiased estimator of σ^2.

7-17. Two different plasma etchers in a semiconductor factory have the same mean etch rate μ. However, machine 1 is newer than machine 2 and consequently has smaller variability in etch rate. We know that the variance of etch rate for machine 1 is σ_1^2 and for machine 2 it is $\sigma_2^2 = a\sigma_1^2$. Suppose that we have n_1 independent observations on etch rate from machine 1 and n_2 independent observations on etch rate from machine 2.
(a) Show that $\hat{\mu} = \alpha\,\overline{X}_1 + (1 - \alpha)\,\overline{X}_2$ is an unbiased estimator of μ for any value of α between 0 and 1.
(b) Find the standard error of the point estimate of μ in part (a).
(c) What value of α would minimize the standard error of the point estimate of μ?
(d) Suppose that $a = 4$ and $n_1 = 2n_2$. What value of α would you select to minimize the standard error of the point estimate of μ. How "bad" would it be to arbitrarily choose $\alpha = 0.5$ in this case?

7-18. Of n_1 randomly selected engineering students at ASU, X_1 owned an HP calculator, and of n_2 randomly selected engineering students at Virginia Tech X_2 owned an HP calculator. Let p_1 and p_2 be the probability that randomly selected ASU and Va. Tech engineering students, respectively, own HP calculators.
(a) Show that an unbiased estimate for $p_1 - p_2$ is $(X_1/n_1) - (X_2/n_2)$.
(b) What is the standard error of the point estimate in part (a)?
(c) How would you compute an estimate of the standard error found in part (b)?
(d) Suppose that $n_1 = 200$, $X_1 = 150$, $n_2 = 250$, and $X_2 = 185$. Use the results of part (a) to compute an estimate of $p_1 - p_2$.
(e) Use the results in parts (b) through (d) to compute an estimate of the standard error of the estimate.

7-3 METHODS OF POINT ESTIMATION

The definitions of unbiasness and other properties of estimators do not provide any guidance about how good estimators can be obtained. In this section, we discuss two methods for obtaining point estimators: the method of moments and the method of maximum likelihood. Maximum likelihood estimates are generally preferable to moment estimators because they have better efficiency properties. However, moment estimators are sometimes easier to compute. Both methods can produce unbiased point estimators.

7-3.1 Method of Moments

The general idea behind the method of moments is to equate **population moments,** which are defined in terms of expected values, to the corresponding **sample moments.** The population moments will be functions of the unknown parameters. Then these equations are solved to yield estimators of the unknown parameters.

Definition

Let $X_1, X_2, \ldots, X_n$ be a random sample from the probability distribution $f(x)$, where $f(x)$ can be a discrete probability mass function or a continuous probability density function. The kth **population moment** (or **distribution moment**) is $E(X^k)$, $k = 1, 2, \ldots$. The corresponding kth **sample moment** is $(1/n)\sum_{i=1}^{n} X_i^k$, $k = 1, 2, \ldots$.

To illustrate, the first population moment is $E(X) = \mu$, and the first sample moment is $(1/n)\sum_{i=1}^{n} X_i = \overline{X}$. Thus by equating the population and sample moments, we find that $\hat{\mu} = \overline{X}$. That is, the sample mean is the **moment estimator** of the population mean. In the general case, the population moments will be functions of the unknown parameters of the distribution, say, $\theta_1, \theta_2, \ldots, \theta_m$.

Definition

Let $X_1, X_2, \ldots, X_n$ be a random sample from either a probability mass function or probability density function with m unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$. The **moment estimators** $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_m$ are found by equating the first m population moments to the first m sample moments and solving the resulting equations for the unknown parameters.

EXAMPLE 7-3

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ. Now there is only one parameter to estimate, so we must equate $E(X)$ to $\overline{X}$. For the exponential, $E(X) = 1/\lambda$. Therefore $E(X) = \overline{X}$ results in $1/\lambda = \overline{X}$, so $\hat{\lambda} = 1/\overline{X}$ is the moment estimator of λ.

As an example, suppose that the time to failure of an electronic module used in an automobile engine controller is tested at an elevated temperature to accelerate the failure mechanism.

The time to failure is exponentially distributed. Eight units are randomly selected and tested, resulting in the following failure time (in hours): $x_1 = 11.96$, $x_2 = 5.03$, $x_3 = 67.40$, $x_4 = 16.07$, $x_5 = 31.50$, $x_6 = 7.73$, $x_7 = 11.10$, and $x_8 = 22.38$. Because $\bar{x} = 21.65$, the moment estimate of λ is $\lambda = 1/\bar{x} = 1/21.65 = 0.0462$.

EXAMPLE 7-4

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with parameters μ and σ^2. For the normal distribution $E(X) = \mu$ and $E(X^2) = \mu^2 + \sigma^2$. Equating $E(X)$ to $\overline{X}$ and $E(X^2)$ to $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ gives

$$\mu = \overline{X}, \qquad \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Solving these equations gives the moment estimators

$$\hat{\mu} = \overline{X}, \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^2}{n} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$$

Notice that the moment estimator of σ^2 is not an unbiased estimator.

EXAMPLE 7-5

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a gamma distribution with parameters r and λ. For the gamma distribution $E(X) = r/\lambda$ and $E(X^2) = r(r+1)/\lambda^2$. The moment estimators are found by solving

$$r/\lambda = \overline{X}, \quad r(r+1)/\lambda^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

The resulting estimators are

$$\hat{r} = \frac{\overline{X}^2}{(1/n)\sum_{i=1}^{n} X_i^2 - \overline{X}_i^2} \qquad \hat{\lambda} = \frac{\overline{X}}{(1/n)\sum_{i=1}^{n} X_i^2 - \overline{X}^2}$$

To illustrate, consider the time to failure data introduced following Example 7-3. For this data, $\bar{x} = 21.65$ and $\sum_{i=1}^{8} x_i^2 = 6639.40$, so the moment estimates are

$$\hat{r} = \frac{(21.65)^2}{(1/8)6639.40 - (21.65)^2} = 1.30, \qquad \hat{\lambda} = \frac{21.65}{(1/8)6639.40 - (21.65)^2} = 0.0599$$

When $r = 1$, the gamma reduces to the exponential distribution. Because $\hat{r}$ slightly exceeds unity, it is quite possible that either the gamma or the exponential distribution would provide a reasonable model for the data.

7-3.2 Method of Maximum Likelihood

One of the best methods of obtaining a point estimator of a parameter is the method of maximum likelihood. This technique was developed in the 1920s by a famous British statistician, Sir R. A. Fisher. As the name implies, the estimator will be the value of the parameter that maximizes the **likelihood function.**

Definition

Suppose that X is a random variable with probability distribution $f(x; \theta)$, where θ is a single unknown parameter. Let $x_1, x_2, \ldots, x_n$ be the observed values in a random sample of size n. Then the **likelihood function** of the sample is

$$L(\theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \cdots \cdot f(x_n; \theta) \tag{7-5}$$

Note that the likelihood function is now a function of only the unknown parameter θ. The **maximum likelihood estimator** of θ is the value of θ that maximizes the likelihood function $L(\theta)$.

In the case of a discrete random variable, the interpretation of the likelihood function is clear. The likelihood function of the sample $L(\theta)$ is just the probability

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

That is, $L(\theta)$ is just the probability of obtaining the sample values $x_1, x_2, \ldots, x_n$. Therefore, in the discrete case, the maximum likelihood estimator is an estimator that maximizes the probability of occurrence of the sample values.

EXAMPLE 7-6

Let X be a Bernoulli random variable. The probability mass function is

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

where p is the parameter to be estimated. The likelihood function of a random sample of size n is

$$\begin{aligned} L(p) &= p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} \cdots p^{x_n}(1-p)^{1-x_n} \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \end{aligned}$$

We observe that if $\hat{p}$ maximizes $L(p)$, $\hat{p}$ also maximizes $\ln L(p)$. Therefore,

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

Now

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

Equating this to zero and solving for p yields $\hat{p} = (1/n) \sum_{i=1}^{n} x_i$. Therefore, the maximum likelihood estimator of p is

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

Suppose that this estimator was applied to the following situation: n items are selected at random from a production line, and each item is judged as either defective (in which case we set $x_i = 1$) or nondefective (in which case we set $x_i = 0$). Then $\sum_{i=1}^{n} x_i$ is the number of defective units in the sample, and $\hat{p}$ is the **sample proportion defective.** The parameter p is the **population proportion defective;** and it seems intuitively quite reasonable to use $\hat{p}$ as an estimate of p.

Although the interpretation of the likelihood function given above is confined to the discrete random variable case, the method of maximum likelihood can easily be extended to a continuous distribution. We now give two examples of maximum likelihood estimation for continuous distributions.

EXAMPLE 7-7

Let X be normally distributed with unknown μ and known variance σ^2. The likelihood function of a random sample of size n, say $X_1, X_2, \ldots, X_n$, is

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

Now

$$\ln L(\mu) = -(n/2)\ln(2\pi\sigma^2) - (2\sigma^2)^{-1} \sum_{i=1}^{n} (x_i - \mu)^2$$

and

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

Equating this last result to zero and solving for μ yields

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

Thus the sample mean is the maximum likelihood estimator of μ. Notice that this is identical to the moment estimator.

EXAMPLE 7-8

Let X be exponentially distributed with parameter λ. The likelihood function of a random sample of size n, say $X_1, X_2, \ldots, X_n$, is

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

The log likelihood is

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

Now

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

and upon equating this last result to zero we obtain

$$\hat{\lambda} = n / \sum_{i=1}^{n} X_i = 1/\overline{X}$$

Thus the maximum likelihood estimator of λ is the reciprocal of the sample mean. Notice that this is the same as the moment estimator.

It is easy to illustrate graphically just how the method of maximum likelihood works. Figure 7-3(a) plots the log of the likelihood function for the exponential parameter from Example 7-8, using the $n = 8$ observations on failure time given following Example 7-3. We found that the estimate of λ was $\hat{\lambda} = 0.0462$. From Example 7-8, we know that this is a maximum likelihood estimate. Figure 7-3(a) shows clearly that the log likelihood function is maximized at a value of λ that is approximately equal to 0.0462. Notice that the log likelihood function is relatively flat in the region of the maximum. This implies that the parameter is not estimated very precisely. If the parameter were estimated precisely, the log likelihood function would be very peaked at the maximum value. The sample size here is relatively small, and this has led to the imprecision in estimation. This is illustrated in Fig. 7-3(b) where we have plotted the difference in log likelihoods for the maximum value, assuming that the sample sizes were $n = 8$, 20, and 40 but that the sample average time to failure remained constant at $\overline{x} = 21.65$. Notice how much steeper the log likelihood is for $n = 20$ in comparsion to $n = 8$, and for $n = 40$ in comparison to both smaller sample sizes.

The method of maximum likelihood can be used in situations where there are several unknown parameters, say, $\theta_1, \theta_2, \ldots, \theta_k$ to estimate. In such cases, the likelihood function is a function of the k unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$, and the maximum likelihood estimators $\{\hat{\Theta}_i\}$ would be found by equating the k partial derivatives $\partial L(\theta_1, \theta_2, \ldots, \theta_k)/\partial\theta_i$, $i = 1, 2, \ldots, k$ to zero and solving the resulting system of equations.

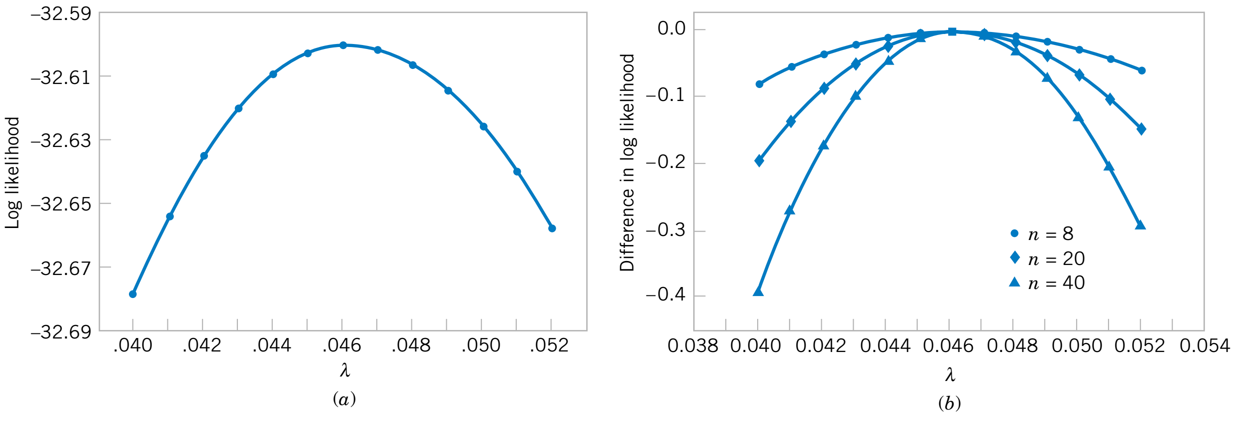


Figure 7-3 Log likelihood for the exponential distribution, using the failure time data. (a) Log likelihood with $n = 8$ (original data). (b) Log likelihood if $n = 8$, 20, and 40.

EXAMPLE 7-9 Let X be normally distributed with mean μ and variance σ^2, where both μ and σ^2 are unknown. The likelihood function for a random sample of size n is

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

and

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

Now

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

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

The solutions to the above equation yield the maximum likelihood estimators

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

Once again, the maximum likelihood estimators are equal to the moment estimators.

Properties of the Maximum Likelihood Estimator

The method of maximum likelihood is often the estimation method that mathematical statisticians prefer, because it is usually easy to use and produces estimators with good statistical properties. We summarize these properties as follows.

Properties of the Maximum Likelihood Estimator

Under very general and not restrictive conditions, when the sample size n is large and if $\hat{\Theta}$ is the maximum likelihood estimator of the parameter θ,

(1) $\hat{\Theta}$ is an approximately unbiased estimator for θ $[E(\hat{\Theta}) \simeq \theta]$,

(2) the variance of $\hat{\Theta}$ is nearly as small as the variance that could be obtained with any other estimator, and

(3) $\hat{\Theta}$ has an approximate normal distribution.

Properties 1 and 2 essentially state that the maximum likelihood estimator is approximately an MVUE. This is a very desirable result and, coupled with the fact that it is fairly easy to obtain in many situations and has an asymptotic normal distribution (the "asymptotic" means "when n is large"), explains why the maximum likelihood estimation technique is widely used. To use maximum likelihood estimation, remember that the distribution of the population must be either known or assumed.

To illustrate the "large-sample" or asymptotic nature of the above properties, consider the maximum likelihood estimator for σ^2, the variance of the normal distribution, in Example 7-9. It is easy to show that

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

The bias is

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

Because the bias is negative, $\hat{\sigma}^2$ tends to underestimate the true variance σ^2. Note that the bias approaches zero as n increases. Therefore, $\hat{\sigma}^2$ is an asymptotically unbiased estimator for σ^2.

We now give another very important and useful property of maximum likelihood estimators.

The Invariance Property

> Let $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k$ be the maximum likelihood estimators of the parameters $\theta_1, \theta_2, \ldots, \theta_k$. Then the maximum likelihood estimator of any function $h(\theta_1, \theta_2, \ldots, \theta_k)$ of these parameters is the same function $h(\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k)$ of the estimators $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k$.

EXAMPLE 7-10

In the normal distribution case, the maximum likelihood estimators of μ and σ^2 were $\hat{\mu} = \overline{X}$ and $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$. To obtain the maximum likelihood estimator of the function $h(\mu, \sigma^2) = \sqrt{\sigma^2} = \sigma$, substitute the estimators $\hat{\mu}$ and $\hat{\sigma}^2$ into the function h, which yields

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{1/2}$$

Thus, the maximum likelihood estimator of the standard deviation σ is *not* the sample standard deviation S.

Complications in Using Maximum Likelihood Estimation

While the method of maximum likelihood is an excellent technique, sometimes complications arise in its use. For example, it is not always easy to maximize the likelihood function because the equation(s) obtained from $dL(\theta)/d\theta = 0$ may be difficult to solve. Furthermore, it may not always be possible to use calculus methods directly to determine the maximum of $L(\theta)$. These points are illustrated in the following two examples.

EXAMPLE 7-11

Let X be uniformly distributed on the interval 0 to a. Since the density function is $f(x) = 1/a$ for $0 \le x \le a$ and zero otherwise, the likelihood function of a random sample of size n is

$$L(a) = \prod_{i=1}^{n}\frac{1}{a} = \frac{1}{a^n}$$

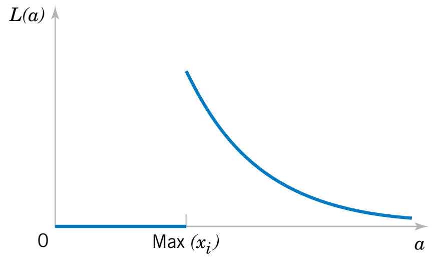


Figure 7-4 The likelihood function for the uniform distribution in Example 7-10.

if $0 \le x_1 \le a$, $0 \le x_2 \le a, \ldots, 0 \le x_n \le a$. Note that the slope of this function is not zero anywhere. That is, as long as $\max(x_i) \le a$, the likelihood is $1/a^n$, which is positive, but when $a < \max(x_i)$, the likelihood goes to zero, as illustrated in Fig. 7-4. Therefore, calculus methods cannot be used directly because the maximum value of the likelihood function occurs at a point of discontinuity. However, since $d/da(a^{-n}) = -n/a^{n+1}$ is less than zero for all values of $a > 0$, a^{-n} is a decreasing function of a. This implies that the maximum of the likelihood function $L(a)$ occurs at the lower boundary point. The figure clearly shows that we could maximize $L(a)$ by setting $\hat{a}$ equal to the smallest value that it could logically take on, which is $\max(x_i)$. Clearly, a cannot be smaller than the largest sample observation, so setting $\hat{a}$ equal to the largest sample value is reasonable.

EXAMPLE 7-12

Let $X_1, X_2, \ldots, X_n$ be a random sample from the gamma distribution. The log of the likelihood function is

$$\ln L(r, \lambda) = \ln\left(\prod_{i=1}^{n} \frac{\lambda^r X_i^{r-1} e^{-\lambda x_i}}{\Gamma(r)}\right)$$
$$= nr\ln(\lambda) + (r-1)\sum_{i=1}^{n}\ln(x_i) - n\ln[\Gamma(r)] - \lambda\sum_{i=1}^{n} x_i$$

The derivatives of the log likelihood are

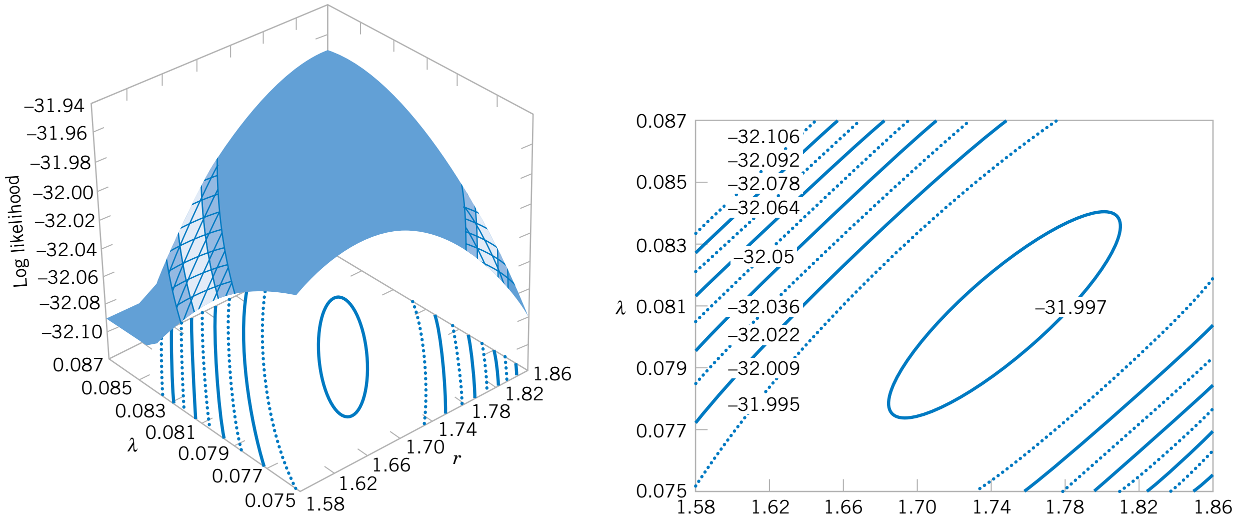
$$\frac{\partial \ln L(r, \lambda)}{\partial r} = n\ln(\lambda) + \sum_{i=1}^{n}\ln(x_i) - n\frac{\Gamma'(r)}{\Gamma(r)}$$
$$\frac{\partial \ln L(r, \lambda)}{\partial \lambda} = \frac{nr}{\lambda} - \sum_{i=1}^{n} x_i$$

When the derivatives are equated to zero, we obtain the equations that must be solved to find the maximum likelihood estimators of r and λ:

$$\hat{\lambda} = \frac{\hat{r}}{\bar{x}}$$
$$n\ln(\hat{\lambda}) + \sum_{i=1}^{n}\ln(x_i) = n\frac{\Gamma'(\hat{r})}{\Gamma(\hat{r})}$$

There is no closed form solution to these equations.

Figure 7-5 shows a graph of the log likelihood for the gamma distribution using the $n = 8$ observations on failure time introduced previously. Figure 7-5(a) shows the **log likelihood**



(a) (b)

Figure 7-5 Log likelihood for the gamma distribution using the failure time data. (a) Log likelihood surface. (b) Contour plot.

surface as a function of r and λ, and Figure 7-5(b) is a **contour plot.** These plots reveal that the log likelihood is maximized at approximately $\hat{r} = 1.75$ and $\hat{\lambda} = 0.08$. Many statistics computer programs use numerical techniques to solve for the maximum likelihood estimates when no simple solution exists.

7-3.3 Bayesian Estimation of Parameters (CD Only)

EXERCISES FOR SECTION 7-3

7-19. Consider the Poisson distribution

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

Find the maximum likelihood estimator of λ, based on a random sample of size n.

7-20. Consider the shifted exponential distribution

$$f(x) = \lambda e^{-\lambda(x-\theta)}, \qquad x \geq \theta$$

When $\theta = 0$, this density reduces to the usual exponential distribution. When $\theta > 0$, there is only positive probability to the right of θ.

(a) Find the maximum likelihood estimator of λ and θ, based on a random sample of size n.

(b) Describe a practical situation in which one would suspect that the shifted exponential distribution is a plausible model.

7-21. Let X be a geometric random variable with parameter p. Find the maximum likelihood estimator of p, based on a random sample of size n.

7-22. Let X be a random variable with the following probability distribution:

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

Find the maximum likelihood estimator of θ, based on a random sample of size n.

7-23. Consider the Weibull distribution

$$f(x) = \begin{cases} \dfrac{\beta}{\delta}\left(\dfrac{x}{\delta}\right)^{\beta-1} e^{-\left(\frac{x}{\delta}\right)^{\beta}}, & 0 < x \\ 0 \quad , & \text{otherwise} \end{cases}$$

(a) Find the likelihood function based on a random sample of size n. Find the log likelihood.

(b) Show that the log likelihood is maximized by solving the equations

$$\beta = \left[\frac{\sum_{i=1}^{n} x_i^{\beta} \ln(x_i)}{\sum_{i=1}^{n} x_i^{\beta}} - \frac{\sum_{i=1}^{n} \ln(x_i)}{n} \right]^{-1}$$

$$\delta = \left[\frac{\sum_{i=1}^{n} x_i^{\beta}}{n} \right]^{1/\beta}$$

(c) What complications are involved in solving the two equations in part (b)?

7-24. Consider the probability distribution in Exercise 7-22. Find the moment estimator of θ.

7-25. Let $X_1, X_2, \ldots, X_n$ be uniformly distributed on the interval 0 to a. Show that the moment estimator of a is $\hat{a} = 2\overline{X}$. Is this an unbiased estimator? Discuss the reasonableness of this estimator.

7-26. Let $X_1, X_2, \ldots, X_n$ be uniformly distributed on the interval 0 to a. Recall that the maximum likelihood estimator of a is $\hat{a} = \max(X_i)$.

(a) Argue intuitively why $\hat{a}$ cannot be an unbiased estimator for a.

(b) Suppose that $E(\hat{a}) = na/(n + 1)$. Is it reasonable that $\hat{a}$ consistently underestimates a? Show that the bias in the estimator approaches zero as n gets large.

(c) Propose an unbiased estimator for a.

(d) Let $Y = \max(X_i)$. Use the fact that $Y \leq y$ if and only if each $X_i \leq y$ to derive the cumulative distribution function of Y. Then show that the probability density function of Y is

$$f(y) = \begin{cases} \dfrac{ny^{n-1}}{a^n}, & 0 \leq y \leq a \\ 0\ , & \text{otherwise} \end{cases}$$

Use this result to show that the maximum likelihood estimator for a is biased.

7-27. For the continuous distribution of the interval 0 to a, we have two unbiased estimators for a: the moment estimator $\hat{a}_1 = 2\overline{X}$ and $\hat{a}_2 = [(n + 1)/n] \max(X_i)$, where $\max(X_i)$ is the largest observation in a random sample of size n (see Exercise 7-26). It can be shown that $V(\hat{a}_1) = a^2/(3n)$ and that $V(\hat{a}_2) = a^2/[n(n + 2)]$. Show that if $n > 1$, $\hat{a}_2$ is a better estimator than $\hat{a}$. In what sense is it a better estimator of a?

7-28. Consider the probability density function

$$f(x) = \frac{1}{\theta^2} x e^{-x/\theta}, \qquad 0 \leq x < \infty, \quad 0 < \theta < \infty$$

Find the maximum likelihood estimator for θ.

7-29. The Rayleigh distribution has probability density function

$$f(x) = \frac{x}{\theta} e^{-x^2/2\theta}, \qquad x > 0, \qquad 0 < \theta < \infty$$

(a) It can be shown that $E(X^2) = 2\theta$. Use this information to construct an unbiased estimator for θ.

(b) Find the maximum likelihood estimator of θ. Compare your answer to part (a).

(c) Use the invariance property of the maximum likelihood estimator to find the maximum likelihood estimator of the median of the Raleigh distribution.

7-30. Consider the probability density function

$$f(x) = c(1 + \theta x), \quad -1 \leq x \leq 1$$

(a) Find the value of the constant c.

(b) What is the moment estimator for θ?

(c) Show that $\hat{\theta} = 3\overline{X}$ is an unbiased estimator for θ.

(d) Find the maximum likelihood estimator for θ.

7-31. Reconsider the oxide thickness data in Exercise 7-12 and suppose that it is reasonable to assume that oxide thickness is normally distributed.

(a) Use the results of Example 7-9 to compute the maximum likelihood estimates of μ and σ^2.

(b) Graph the likelihood function in the vicinity of $\hat{\mu}$ and $\hat{\sigma}^2$, the maximum likelihood estimates, and comment on its shape.

7-32. Continuation of Exercise 7-31. Suppose that for the situation of Exercise 7-12, the sample size was larger ($n = 40$) but the maximum likelihood estimates were numerically equal to the values obtained in Exercise 7-31. Graph the likelihood function for $n = 40$, compare it to the one from Exercise 7-31 (b), and comment on the effect of the larger sample size.

7-4 SAMPLING DISTRIBUTIONS

Statistical inference is concerned with making **decisions** about a population based on the information contained in a random sample from that population. For instance, we may be interested in the mean fill volume of a can of soft drink. The mean fill volume in the population is required to be 300 milliliters. An engineer takes a random sample of 25 cans and computes the sample average fill volume to be $\bar{x} = 298$ milliliters. The engineer will probably decide that the population mean is $\mu = 300$ milliliters, even though the sample mean was 298 milliliters because he or she knows that the sample mean is a reasonable estimate of μ and that a sample mean of 298 milliliters is very likely to occur, even if the true population mean is $\mu = 300$ milliliters. In fact, if the true mean is 300 milliliters, tests of 25 cans made repeatedly, perhaps every five minutes, would produce values of $\bar{x}$ that vary both above and below $\mu = 300$ milliliters.

The sample mean is a statistic; that is, it is a random variable that depends on the results obtained in each particular sample. Since a statistic is a random variable, it has a probability distribution.

Definition

The probability distribution of a statistic is called a **sampling distribution.**

For example, the probability distribution of $\overline{X}$ is called the **sampling distribution of the mean.**

The sampling distribution of a statistic depends on the distribution of the population, the size of the sample, and the method of sample selection. The next section presents perhaps the most important sampling distribution. Other sampling distributions and their applications will be illustrated extensively in the following two chapters.

7-5 SAMPLING DISTRIBUTIONS OF MEANS

Consider determining the sampling distribution of the sample mean $\overline{X}$. Suppose that a random sample of size n is taken from a normal population with mean μ and variance σ^2. Now each observation in this sample, say, $X_1, X_2, \ldots, X_n$, is a normally and independently distributed random variable with mean μ and variance σ^2. Then by the reproductive property of the normal distribution, Equation 5-41 in Chapter 5, we conclude that the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

has a normal distribution with mean

$$\mu_{\overline{X}} = \frac{\mu + \mu + \cdots + \mu}{n} = \mu$$

and variance

$$\sigma^2_{\overline{X}} = \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

If we are sampling from a population that has an unknown probability distribution, the sampling distribution of the sample mean will still be approximately normal with mean μ and variance σ^2/n, if the sample size n is large. This is one of the most useful theorems in statistics, called the **central limit theorem.** The statement is as follows:

Theorem 7-2: The Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2, and if $\overline{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{7-6}$$

as $n \rightarrow \infty$, is the standard normal distribution.

The normal approximation for $\overline{X}$ depends on the sample size n. Figure 7-6(a) shows the distribution obtained for throws of a single, six-sided true die. The probabilities are equal (1/6) for all the values obtained, 1, 2, 3, 4, 5, or 6. Figure 7-6(b) shows the distribution of the average score obtained when tossing two dice, and Fig. 7-6(c), 7-6(d), and 7-6(e) show the distributions of average scores obtained when tossing three, five, and ten dice, respectively. Notice that, while the population (one die) is relatively far from normal, the distribution of averages is approximated reasonably well by the normal distribution for sample sizes as small as five. (The dice throw distributions are discrete, however, while the normal is continuous). Although the central limit theorem will work well for small samples ($n = 4, 5$) in most cases, particularly where the population is continuous, unimodal, and symmetric, larger samples will be required in other situations, depending on the shape of the population. In many cases of practical interest, if $n \geq 30$, the normal approximation will be satisfactory regardless of the

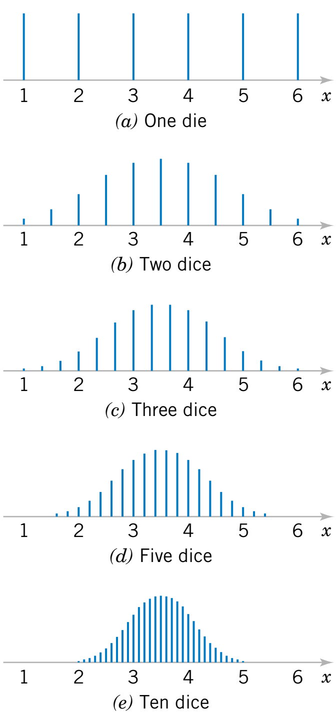


Figure 7-6 Distributions of average scores from throwing dice. [Adapted with permission from Box, Hunter, and Hunter (1978).]

shape of the population. If $n < 30$, the central limit theorem will work if the distribution of the population is not severely nonnormal.

EXAMPLE 7-13 An electronics company manufactures resistors that have a mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance is normal. Find the probability that a random sample of $n = 25$ resistors will have an average resistance less than 95 ohms.

Note that the sampling distribution of $\overline{X}$ is normal, with mean $\mu_{\overline{X}} = 100$ ohms and a standard deviation of

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

Therefore, the desired probability corresponds to the shaded area in Fig. 7-7. Standardizing the point $\overline{X} = 95$ in Fig. 7-7, we find that

$$z = \frac{95 - 100}{2} = -2.5$$

and therefore,

$$\begin{aligned} P(\overline{X} < 95) &= P(Z < -2.5) \\ &= 0.0062 \end{aligned}$$

The following example makes use of the central limit theorem.

EXAMPLE 7-14 Suppose that a random variable X has a continuous uniform distribution

$$f(x) = \begin{cases} 1/2, & 4 \le x \le 6 \\ 0, & \text{otherwise} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n = 40$.

The mean and variance of X are $\mu = 5$ and $\sigma^2 = (6 - 4)^2/12 = 1/3$. The central limit theorem indicates that the distribution of $\overline{X}$ is approximately normal with mean $\mu_{\overline{X}} = 5$ and variance $\sigma^2_{\overline{X}} = \sigma^2/n = 1/[3(40)] = 1/120$. The distributions of X and $\overline{X}$ are shown in Fig. 7-8.

Now consider the case in which we have two independent populations. Let the first population have mean μ_1 and variance σ_1^2 and the second population have mean μ_2 and variance σ_2^2. Suppose that both populations are normally distributed. Then, using the fact that linear

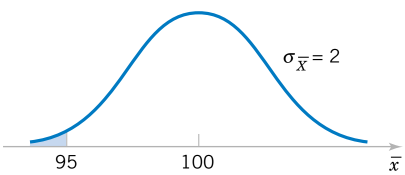


Figure 7-7 Probability for Example 7-13.

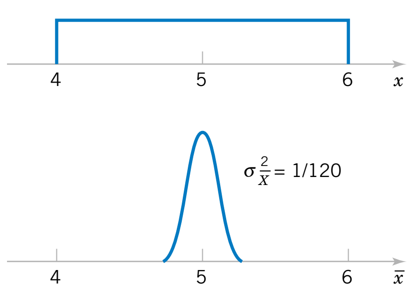


Figure 7-8 The distributions of X and $\overline{X}$ for Example 7-14.

combinations of independent normal random variables follow a normal distribution (see Equation 5-41), we can say that the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is normal with mean

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_{\overline{X}_1} - \mu_{\overline{X}_2} = \mu_1 - \mu_2 \tag{7-7}$$

and variance

$$\sigma^2_{\overline{X}_1 - \overline{X}_2} = \sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{7-8}$$

If the two populations are not normally distributed and if both sample sizes n_1 and n_2 are greater than 30, we may use the central limit theorem and assume that $\overline{X}_1$ and $\overline{X}_2$ follow approximately independent normal distributions. Therefore, the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is approximately normal with mean and variance given by Equations 7-7 and 7-8, respectively. If either n_1 or n_2 is less than 30, the sampling distribution of $\overline{X}_1 - \overline{X}_2$ will still be approximately normal with mean and variance given by Equations 7-7 and 7-8, provided that the population from which the small sample is taken is not dramatically different from the normal. We may summarize this with the following definition.

Definition

If we have two independent populations with means μ_1 and μ_2 and variances σ_2^2 and σ_2^2 and if $\overline{X}_1$ and $\overline{X}_2$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations, then the sampling distribution of

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \tag{7-9}$$

is approximately standard normal, if the conditions of the central limit theorem apply. If the two populations are normal, the sampling distribution of Z is exactly standard normal.

EXAMPLE 7-15

The effective life of a component used in a jet-turbine aircraft engine is a random variable with mean 5000 hours and standard deviation 40 hours. The distribution of effective life is fairly close to a normal distribution. The engine manufacturer introduces an improvement into the manufacturing process for this component that increases the mean life to 5050 hours and decreases the standard deviation to 30 hours. Suppose that a random sample of $n_1 = 16$ components is selected from the "old" process and a random sample of $n_2 = 25$ components is selected from the "improved" process. What is the probability that the difference in the two sample means $\overline{X}_2 - \overline{X}_1$ is at least 25 hours? Assume that the old and improved processes can be regarded as independent populations.

To solve this problem, we first note that the distribution of $\overline{X}_1$ is normal with mean $\mu_1 = 5000$ hours and standard deviation $\sigma_1/\sqrt{n_1} = 40/\sqrt{16} = 10$ hours, and the distribution of $\overline{X}_2$ is normal with mean $\mu_2 = 5050$ hours and standard deviation $\sigma_2/\sqrt{n_2} = 30/\sqrt{25} = 6$ hours. Now the distribution of $\overline{X}_2 - \overline{X}_1$ is normal with mean $\mu_2 - \mu_1 = 5050 - 5000 = 50$ hours and variance $\sigma_2^2/n_2 + \sigma_1^2/n_1 = (6)^2 + (10)^2 = 136$ hours^2. This sampling distribution is shown in Fig. 7-9. The probability that $\overline{X}_2 - \overline{X}_1 \geq 25$ is the shaded portion of the normal distribution in this figure.

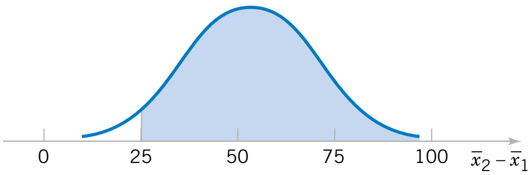


Figure 7-9 The sampling distribution of $\bar{X}_2 - \bar{X}_1$ in Example 7-15.

Corresponding to the value $\bar{x}_2 - \bar{x}_1 = 25$ in Fig. 7-9, we find that

$$z = \frac{25 - 50}{\sqrt{136}} = -2.14$$

and we find that

$$\begin{aligned} P(\bar{X}_2 - \bar{X}_1 \geq 25) &= P(Z \geq -2.14) \\ &= 0.9838 \end{aligned}$$

EXERCISES FOR SECTION 7-5

7-33. PVC pipe is manufactured with a mean diameter of 1.01 inch and a standard deviation of 0.003 inch. Find the probability that a random sample of $n = 9$ sections of pipe will have a sample mean diameter greater than 1.009 inch and less than 1.012 inch.

7-34. Suppose that samples of size $n = 25$ are selected at random from a normal population with mean 100 and standard deviation 10. What is the probability that the sample mean falls in the interval from $\mu_{\bar{X}} - 1.8\sigma_{\bar{X}}$ to $\mu_{\bar{X}} + 1.0\sigma_{\bar{X}}$?

7-35. A synthetic fiber used in manufacturing carpet has tensile strength that is normally distributed with mean 75.5 psi and standard deviation 3.5 psi. Find the probability that a random sample of $n = 6$ fiber specimens will have sample mean tensile strength that exceeds 75.75 psi.

7-36. Consider the synthetic fiber in the previous exercise. How is the standard deviation of the sample mean changed when the sample size is increased from $n = 6$ to $n = 49$?

7-37. The compressive strength of concrete is normally distributed with $\mu = 2500$ psi and $\sigma = 50$ psi. Find the probability that a random sample of $n = 5$ specimens will have a sample mean diameter that falls in the interval from 2499 psi to 2510 psi.

7-38. Consider the concrete specimens in the previous example. What is the standard error of the sample mean?

7-39. A normal population has mean 100 and variance 25. How large must the random sample be if we want the standard error of the sample average to be 1.5?

7-40. Suppose that the random variable X has the continuous uniform distribution

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

Suppose that a random sample of $n = 12$ observations is selected from this distribution. What is the probability distribution of $\bar{X} - 6$? Find the mean and variance of this quantity.

7-41. Suppose that X has a discrete uniform distribution

$$f(x) = \begin{cases} 1/3, & x = 1, 2, 3 \\ 0, & \text{otherwise} \end{cases}$$

A random sample of $n = 36$ is selected from this population. Find the probability that the sample mean is greater than 2.1 but less than 2.5, assuming that the sample mean would be measured to the nearest tenth.

7-42. The amount of time that a customer spends waiting at an airport check-in counter is a random variable with mean 8.2 minutes and standard deviation 1.5 minutes. Suppose that a random sample of $n = 49$ customers is observed. Find the probability that the average time waiting in line for these customers is

(a) Less than 10 minutes
(b) Between 5 and 10 minutes
(c) Less than 6 minutes

7-43. A random sample of size $n_1 = 16$ is selected from a normal population with a mean of 75 and a standard deviation of 8. A second random sample of size $n_2 = 9$ is taken from another normal population with mean 70 and standard deviation 12. Let $\bar{X}_1$ and $\bar{X}_2$ be the two sample means. Find

(a) The probability that $\bar{X}_1 - \bar{X}_2$ exceeds 4
(b) The probability that $3.5 \leq \bar{X}_1 - \bar{X}_2 \leq 5.5$

7-44. A consumer electronics company is comparing the brightness of two different types of picture tubes for use in its television sets. Tube type A has mean brightness of 100 and standard deviation of 16, while tube type B has unknown mean brightness, but the standard deviation is assumed to be identical to that for type A. A random sample of $n = 25$ tubes of each type is selected, and $\overline{X}_B - \overline{X}_A$ is computed. If μ_B equals or exceeds μ_A, the manufacturer would like to adopt type B for use. The observed difference is $\bar{x}_B - \bar{x}_A = 3.5$. What decision would you make, and why?

7-45. The elasticity of a polymer is affected by the concentration of a reactant. When low concentration is used, the true mean elasticity is 55, and when high concentration is used the mean elasticity is 60. The standard deviation of elasticity is 4, regardless of concentration. If two random samples of size 16 are taken, find the probability that $\overline{X}_{\text{high}} - \overline{X}_{\text{low}} \geq 2$.

Supplemental Exercises

7-46. Suppose that a random variable is normally distributed with mean μ and variance σ^2, and we draw a random sample of five observations from this distribution. What is the joint probability density function of the sample?

7-47. Transistors have a life that is exponentially distributed with parameter λ. A random sample of n transistors is taken. What is the joint probability density function of the sample?

7-48. Suppose that X is uniformly distributed on the interval from 0 to 1. Consider a random sample of size 4 from X. What is the joint probability density function of the sample?

7-49. A procurement specialist has purchased 25 resistors from vendor 1 and 30 resistors from vendor 2. Let $X_{1,1}, X_{1,2}, \ldots, X_{1,25}$ represent the vendor 1 observed resistances, which are assumed to be normally and independently distributed with mean 100 ohms and standard deviation 1.5 ohms. Similarly, let $X_{2,1}, X_{2,2}, \ldots, X_{2,30}$ represent the vendor 2 observed resistances, which are assumed to be normally and independently distributed with mean 105 ohms and standard deviation of 2.0 ohms. What is the sampling distribution of $\overline{X}_1 - \overline{X}_2$?

7-50. Consider the resistor problem in Exercise 7-49. What is the standard error of $\overline{X}_1 - \overline{X}_2$?

7-51. A random sample of 36 observations has been drawn from a normal distribution with mean 50 and standard deviation 12. Find the probability that the sample mean is in the interval $47 \leq \overline{X} \leq 53$.

7-52. Is the assumption of normality important in Exercise 7-51? Why?

7-53. A random sample of $n = 9$ structural elements is tested for compressive strength. We know that the true mean compressive strength $\mu = 5500$ psi and the standard deviation is $\sigma = 100$ psi. Find the probability that the sample mean compressive strength exceeds 4985 psi.

7-54. A normal population has a known mean 50 and known variance $\sigma^2 = 2$. A random sample of $n = 16$ is selected from this population, and the sample mean is $\bar{x} = 52$. How unusual is this result?

7-55. A random sample of size $n = 16$ is taken from a normal population with $\mu = 40$ and $\sigma^2 = 5$. Find the probability that the sample mean is less than or equal to 37.

7-56. A manufacturer of semiconductor devices takes a random sample of 100 chips and tests them, classifying each chip as defective or nondefective. Let $X_i = 0$ if the chip is nondefective and $X_i = 1$ if the chip is defective. The sample fraction defective is

$$\hat{P} = \frac{X_1 + X_2 + \cdots + X_{100}}{100}$$

What is the sampling distribution of the random variable $\hat{P}$?

7-57. Let X be a random variable with mean μ and variance σ^2. Given two independent random samples of sizes n_1 and n_2, with sample means $\overline{X}_1$ and $\overline{X}_2$, show that

$$\overline{X} = a\overline{X}_1 + (1 - a)\overline{X}_2, \quad 0 < a < 1$$

is an unbiased estimator for μ. If $\overline{X}_1$ and $\overline{X}_2$ are independent, find the value of a that minimizes the standard error of $\overline{X}$.

7-58. A random variable x has probability density function

$$f(x) = \frac{1}{2\theta^3} x^2 e^{-x/\theta}, \qquad 0 < x < \infty, \quad 0 < \theta < \infty$$

Find the maximum likelihood estimator for θ.

7-59. Let $f(x) = \theta x^{\theta - 1}$, $0 < \theta < \infty$, and $0 < x < 1$. Show that $\hat{\Theta} = -n/(\ln \prod_{i=1}^{n} X_i)$ is the maximum likelihood estimator for θ.

7-60. Let $f(x) = (1/\theta)x^{(1-\theta)/\theta}$, $0 < x < 1$, and $0 < \theta < \infty$. Show that $\hat{\Theta} = -(1/n)\sum_{i=1}^{n} \ln(X_i)$ is the maximum likelihood estimator for θ and that $\hat{\Theta}$ is an unbiased estimator for θ.

MIND-EXPANDING EXERCISES

7-61. A lot consists of N transistors, and of these M ($M \le N$) are defective. We randomly select two transistors without replacement from this lot and determine whether they are defective or nondefective. The random variable

$$X_i = \begin{cases} 1, & \text{if the } i\text{th transistor is nondefective} \\ 0, & \text{if the } i\text{th transistor is defective} \end{cases} \quad i = 1, 2$$

Determine the joint probability function for X_1 and X_2. What are the marginal probability functions for X_1 and X_2? Are X_1 and X_2 independent random variables?

7-62. When the sample standard deviation is based on a random sample of size n from a normal population, it can be shown that S is a biased estimator for σ. Specifically,

$$E(S) = \sigma\sqrt{2/(n-1)}\ \Gamma(n/2)/\Gamma[(n-1)/2]$$

(a) Use this result to obtain an unbiased estimator for σ of the form $c_n S$, when the constant c_n depends on the sample size n.

(b) Find the value of c_n for $n = 10$ and $n = 25$. Generally, how well does S perform as an estimator of σ for large n with respect to bias?

7-63. A collection of n randomly selected parts is measured twice by an operator using a gauge. Let X_i and Y_i denote the measured values for the ith part. Assume that these two random variables are independent and normally distributed and that both have true mean μ_i and variance σ^2.

(a) Show that the maximum likelihood estimator of σ^2 is $\hat{\sigma}^2 = (1/4n)\sum_{i=1}^{n}(X_i - Y_i)^2$.

(b) Show that $\hat{\sigma}^2$ is a biased estimator for σ^2. What happens to the bias as n becomes large?

(c) Find an unbiased estimator for σ^2.

7-64. Consistent Estimator. Another way to measure the closeness of an estimator $\hat{\Theta}$ to the parameter θ is in terms of consistency. If $\hat{\Theta}_n$ is an estimator of θ based on a random sample of n observations, $\hat{\Theta}_n$ is consistent for θ if

$$\lim_{n\to\infty} P(|\hat{\Theta}_n - \theta| < \epsilon) = 1$$

Thus, consistency is a large-sample property, describing the limiting behavior of $\hat{\Theta}_n$ as n tends to infinity. It is usually difficult to prove consistency using the above definition, although it can be done from other approaches. To illustrate, show that $\overline{X}$ is a consistent estimator of μ (when $\sigma^2 < \infty$) by using Chebyshev's inequality. See Section 5-10 (CD Only).

7-65. Order Statistics. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from X, a random variable having distribution function $F(x)$. Rank the elements in order of increasing numerical magnitude, resulting in $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, where $X_{(1)}$ is the smallest sample element ($X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$) and $X_{(n)}$ is the largest sample element ($X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$). $X_{(i)}$ is called the ith order statistic. Often the distribution of some of the order statistics is of interest, particularly the minimum and maximum sample values. $X_{(1)}$ and $X_{(n)}$, respectively. Prove that the cumulative distribution functions of these two order statistics, denoted respectively by $F_{X_{(1)}}(t)$ and $F_{X_{(n)}}(t)$ are

$$F_{X_{(1)}}(t) = 1 - [1 - F(t)]^n$$
$$F_{X_{(n)}}(t) = [F(t)]^n$$

Prove that if X is continuous with probability density function $f(x)$, the probability distributions of $X_{(1)}$ and $X_{(n)}$ are

$$f_{X_{(1)}}(t) = n[1 - F(t)]^{n-1} f(t)$$
$$f_{X_{(n)}}(t) = n[F(t)]^{n-1} f(t)$$

7-66. Continuation of Exercise 7-65. Let $X_1, X_2, \ldots, X_n$ be a random sample of a Bernoulli random variable with parameter p. Show that

$$P(X_{(n)} = 1) = 1 - (1-p)^n$$
$$P(X_{(1)} = 0) = 1 - p^n$$

Use the results of Exercise 7-65.

7-67. Continuation of Exercise 7-65. Let $X_1, X_2, \ldots, X_n$ be a random sample of a normal random variable with mean μ and variance σ^2. Using the results of Exercise 7-65, derive the probability density functions of $X_{(1)}$ and $X_{(n)}$.

MIND-EXPANDING EXERCISES

7-68. Continuation of Exercise 7-65. Let $X_1, X_2, \ldots, X_n$ be a random sample of an exponential random variable of parameter λ. Derive the cumulative distribution functions and probability density functions for $X_{(1)}$ and $X_{(n)}$. Use the result of Exercise 7-65.

7-69. Let $X_1, X_2, \ldots, X_n$ be a random sample of a continuous random variable with cumulative distribution function $F(x)$. Find

$$E[F(X_{(n)})]$$

and

$$E[F(X_{(1)})]$$

7-70. Let X be a random variable with mean μ and variance σ^2, and let $X_1, X_2, \ldots, X_n$ be a random sample of size n from X. Show that the statistic $V = k\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2$ is an unbiased estimator for σ^2 for an appropriate choice for the constant k. Find this value for k.

7-71. When the population has a normal distribution, the estimator

$$\hat{\sigma} = \text{median}\,(|X_1 - \overline{X}|,\ |X_2 - \overline{X}|,\ \cdots, |X_n - \overline{X}|)/0.6745$$

is sometimes used to estimate the population standard deviation. This estimator is more robust to outliers than the usual sample standard deviation and usually does not differ much from S when there are no unusual observations.

(a) Calculate $\hat{\sigma}$ and S for the data 10, 12, 9, 14, 18, 15, and 16.

(b) Replace the first observation in the sample (10) with 50 and recalculate both S and $\hat{\sigma}$.

7-72. **Censored Data.** A common problem in industry is life testing of components and systems. In this problem, we will assume that lifetime has an exponential distribution with parameter λ, so $\hat{\mu} = 1/\hat{\lambda} = \overline{X}$ is an unbiased estimate of μ. When n components are tested until failure and the data $X_1, X_2, \ldots, X_n$ represent actual lifetimes, we have a complete sample, and $\overline{X}$ is indeed an unbiased estimator of μ. However, in many situations, the components are only left under test until $r < n$ failures have occurred. Let Y_1 be the time of the first failure, Y_2 be the time of the second failure, ..., and Y_r be the time of the last failure. This type of test results in **censored data.** There are $n - r$ units still running when the test is terminated. The total accumulated test time at termination is

$$T_r = \sum_{i=1}^{r} Y_i + (n - r)Y_r$$

(a) Show that $\hat{\mu} = T_r/r$ is an unbiased estimator for μ. [*Hint:* You will need to use the memoryless property of the exponential distribution and the results of Exercise 7-68 for the distribution of the minimum of a sample from an exponential distribution with parameter λ.]

(b) It can be shown that $V(T_r/r) = 1/(\lambda^2 r)$. How does this compare to $V(\overline{X})$ in the uncensored experiment?

IMPORTANT TERMS AND CONCEPTS

In the E-book, click on any term or concept below to go to that subject.

Bias in parameter estimation
Central limit theorem
Estimator versus estimate
Likelihood function
Maximum likelihood estimator
Mean square error of an estimator
Minimum variance unbiased estimator
Moment estimator
Normal distribution as the sampling distribution of a sample mean
Normal distribution as the sampling distribution of the difference in two sample means
Parameter estimation
Point estimator
Population or distribution moments
Sample moments
Sampling distribution
Standard error and estimated standard error of an estimator
Statistic
Statistical inference
Unbiased estimator

CD MATERIAL

Bayes estimator
Bootstrap
Posterior distribution
Prior distribution

7-2.2 Proof That S is a Biased Estimator of σ (CD Only)

We proved that the sample variance is an unbiased estimator of the population variance, that is, $E(S^2) = \sigma^2$, and that this result does not depend on the form of the distribution. However, the sample standard deviation is not an unbiased estimator of the population standard deviation. This is easy to demonstrate for the case where the random variable X follows a normal distribution.

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a normal population with mean μ and variance σ^2. Now it can be shown that the distribution of the random variable

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

is chi-square with $n - 1$ degrees of freedom, denoted χ^2_{n-1} (the chi-squared distribution was introduced in our discussion of the gamma distribution in Chapter 4, and the above result will be presented formally in Chapter 8). Therefore the distribution of S^2 is $\sigma^2/(n-1)$ times a χ^2_{n-1} random variable. So when sampling from a normal distribution, the expected value of S^2 is

$$E(S^2) = E\left(\frac{\sigma^2}{n-1}\chi^2_{n-1}\right) = \frac{\sigma^2}{n-1}E(\chi^2_{n-1}) = \frac{\sigma^2}{n-1}(n-1) = \sigma^2$$

because the mean of a chi-squared random variable with $n - 1$ degrees of freedom is $n - 1$. Now it follows that the distribution of

$$\frac{\sqrt{(n-1)}S}{\sigma}$$

is a chi distribution with $n - 1$ degrees of freedom, denoted χ_{n-1}. The expected value of S can be written as

$$E(S) = E\left(\frac{\sigma}{\sqrt{n-1}}\chi_{n-1}\right) = \frac{\sigma}{\sqrt{n-1}}E(\chi_{n-1})$$

The mean of the chi distribution with $n - 1$ degrees of freedom is

$$E(\chi_{n-1}) = \sqrt{2}\frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}$$

where the gamma function $\Gamma(r) = \int_0^\infty y^{r-1}e^{-y}\,dy$. Then

$$\begin{aligned} E(S) &= \sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}\sigma \\ &= c_n\sigma \end{aligned}$$

Although S is a biased estimator of σ, the bias gets small fairly quickly as the sample size n increases. For example, note that $c_n = 0.94$ for a sample of $n = 5$, $c_n = 0.9727$ for a sample of $n = 10$, and $c_n = 0.9896$ or very nearly unity for a sample of $n = 25$.

7-2.5 Bootstrap Estimate of the Standard Error (CD Only)

There are situations in which the standard error of the point estimator is unknown. Usually, these are cases where the form of $\hat{\Theta}$ is complicated, and the standard expectation and variance operators are difficult to apply. A computer-intensive technique called the **bootstrap** that was developed in recent years can be used for this problem.

Suppose that we are sampling from a population that can be modeled by the probability distribution $f(x; \theta)$. The random sample results in data values $x_1, x_2, \ldots, x_n$ and we obtain $\hat{\theta}$ as the point estimate of θ. We would now use a computer to obtain *bootstrap samples* from the distribution $f(x; \hat{\theta})$, and for each of these samples we calculate the bootstrap estimate $\hat{\theta}^*$ of θ. This results in

Bootstrap Sample	Observations	Bootstrap Estimate
1	$x_1^*, x_2^*, \ldots, x_n^*$	$\hat{\theta}_1^*$
2	$x_1^*, x_2^*, \ldots, x_n^*$	$\hat{\theta}_2^*$
$\vdots$	$\vdots$	$\vdots$
B	$x_1^*, x_2^*, \ldots, x_n^*$	$\hat{\theta}_B^*$

Usually $B = 100$ or 200 of these bootstrap samples are taken. Let $\overline{\theta}^* = (1/B)\sum_{i=1}^{B} \hat{\theta}_i^*$ be the sample mean of the bootstrap estimates. The bootstrap estimate of the standard error of $\hat{\Theta}$ is just the sample standard deviation of the $\hat{\theta}_i^*$, or

$$s_{\hat{\Theta}} = \sqrt{\frac{\sum_{i=1}^{B} (\hat{\theta}_i^* - \overline{\theta}^*)^2}{B - 1}} \tag{S7-1}$$

In the bootstrap literature, $B - 1$ in Equation S7-1 is often replaced by B. However, for the large values usually employed for B, there is little difference in the estimate produced for $s_{\hat{\Theta}}$.

EXAMPLE S7-1

The time to failure of an electronic module used in an automobile engine controller is tested at an elevated temperature in order to accelerate the failure mechanism. The time to failure is exponentially distributed with unknown parameter λ. Eight units are selected at random and tested, with the resulting failure times (in hours): $x_1 = 11.96$, $x_2 = 5.03$, $x_3 = 67.40$, $x_4 = 16.07$, $x_5 = 31.50$, $x_6 = 7.73$, $x_7 = 11.10$, and $x_8 = 22.38$. Now the mean of an exponential distribution is $\mu = 1/\lambda$, so $E(X) = 1/\lambda$, and the expected value of the sample average is $E(\overline{X}) = 1/\lambda$. Therefore, a reasonable way to estimate λ is with $\hat{\lambda} = 1/\overline{X}$. For our sample, $\overline{x} = 21.65$, so our estimate of λ is $\hat{\lambda} = 1/21.65 = 0.0462$. To find the bootstrap standard error we would now obtain $B = 200$ (say) samples of $n = 8$ observations each from an exponential distribution with parameter $\lambda = 0.0462$. The following table shows some of these results:

Bootstrap Sample	Observations	Bootstrap Estimate
1	8.01, 28.85, 14.14, 59.12, 3.11, 32.19, 5.26, 14.17	$\hat{\lambda}_1^* = 0.0485$
2	33.27, 2.10, 40.17, 32.43, 6.94, 30.66, 18.99, 5.61	$\hat{\lambda}_2^* = 0.0470$
$\vdots$	$\vdots$	$\vdots$
200	40.26, 39.26, 19.59, 43.53, 9.55, 7.07, 6.03, 8.94	$\hat{\lambda}_{200}^* = 0.0459$

The sample average of the $\hat{\lambda}_i^*$ (the bootstrap estimates) is 0.0513, and the standard deviation of these bootstrap estimates is 0.020. Therefore, the bootstrap standard error of $\hat{\lambda}$ is 0.020. In this case, estimating the parameter λ in an exponential distribution, the variance of the estimator we used, $\hat{\lambda}$, is known. When n is large, $V(\hat{\lambda}) = \lambda^2/n$. Therefore the estimated standard error of $\hat{\lambda}$ is $\sqrt{\hat{\lambda}^2/n} = \sqrt{(0.0462)^2/8} = 0.016$. Notice that this result agrees reasonably closely with the bootstrap standard error.

Sometimes we want to use the bootstrap in situations in which the form of the probability distribution is unknown. In these cases, we take the n observations in the sample as the *population* and select B random samples each of size n, with replacement, from this population. Then Equation S7-1 can be applied as described above. The book by Efron and Tibshirani (1993) is an excellent introduction to the bootstrap.

7-3.3 Bayesian Estimation of Parameters (CD Only)

This book uses methods of statistical inference based on the information in the sample data. In effect, these methods interpret probabilities as relative frequencies. Sometimes we call probabilities that are interpreted in this manner **objective probabilities.** There is another approach to statistical inference, called the **Bayesian approach,** that combines sample information with other information that may be available prior to collecting the sample. In this section we briefly illustrate how this approach may be used in parameter estimation.

Suppose that the random variable X has a probability distribution that is a function of one parameter θ. We will write this probability distribution as $f(x \mid \theta)$. This notation implies that the exact form of the distribution of X is conditional on the value assigned to θ. The classical approach to estimation would consist of taking a random sample of size n from this distribution and then substituting the sample values x_i into the estimator for θ. This estimator could have been developed using the maximum likelihood approach, for example.

Suppose that we have some additional information about θ and that we can summarize that information in the form of a probability distribution for θ, say, $f(\theta)$. This probability distribution is often called the **prior distribution** for θ, and suppose that the mean of the prior is μ_0 and the variance is σ_0^2. This is a very novel concept insofar as the rest of this book is concerned because we are now viewing the parameter θ as a random variable. The probabilities associated with the prior distribution are often called **subjective probabilities,** in that they usually reflect the analyst's degree of belief regarding the true value of θ. The Bayesian approach to estimation uses the prior distribution for θ, $f(\theta)$, and the joint probability distribution of the sample, say $f(x_1, x_2, \ldots, x_n \mid \theta)$, to find a **posterior distribution** for θ, say, $f(\theta \mid x_1, x_2, \ldots, x_n)$. This posterior distribution contains information both from the sample and the prior distribution for θ. In a sense, it expresses our degree of belief regarding the true value of θ after observing the sample data. It is easy conceptually to find the posterior distribution. The joint probability distribution of the sample $X_1, X_2, \ldots, X_n$ and the parameter θ (remember that θ is a random variable) is

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

and the marginal distribution of $X_1, X_2, \ldots, X_n$ is

$$f(x_1, x_2, \ldots, x_n) = \begin{cases} \displaystyle\sum_{\theta} f(x_1, x_2, \ldots, x_n, \theta), & \theta \text{ discrete} \\ \displaystyle\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n, \theta)\, d\theta, & \theta \text{ continuous} \end{cases}$$

Therefore, the desired distribution is

$$f(\theta \mid x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n, \theta)}{f(x_1, x_2, \ldots, x_n)}$$

We define the **Bayes estimator** of θ as the value $\tilde{\theta}$ that corresponds to the mean of the posterior distribution $f(\theta \mid x_1, x_2, \ldots, x_n)$.

Sometimes, the mean of the posterior distribution of θ can be determined easily. As a function of θ, $f(\theta \mid x_1, \ldots, x_n)$ is a probability density function and $x_1, \ldots, x_n$ are just constants. Because θ enters into $f(\theta \mid x_1, \ldots, x_n)$ only through $f(x_1, \ldots, x_n, \theta)$ if $f(x_1, \ldots, x_n, \theta)$, as a function of θ is recognized as a well-known probability function, the posterior mean of θ can be deduced from the well-known distribution without integration or even calculation of $f(x_1, \ldots, x_n)$.

EXAMPLE S7-2

Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2, where μ is unknown and σ^2 is known. Assume that the prior distribution for μ is normal with mean μ_0 and variance σ_0^2; that is

$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-(\mu-\mu_0)^2/(2\sigma_0^2)} = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(\mu^2 - 2\mu_0\mu + \mu_0^2)/(2\sigma_0^2)}$$

The joint probability distribution of the sample is

$$f(x_1, x_2, \ldots, x_n \mid \mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)(\sum x_i^2 - 2\mu\sum x_i + n\mu^2)}$$

Thus, the joint probability distribution of the sample and μ is

$$f(x_1, x_2, \ldots, x_n, \mu) = \frac{1}{(2\pi\sigma^2)^{n/2}\sqrt{2\pi}\sigma_0} e^{-(1/2)[(1/\sigma_0^2 + n/\sigma^2)\mu^2 - (2\mu_0/\sigma_0^2 + 2\sum x_i/\sigma^2)\mu + \sum x_i^2/\sigma^2 + \mu_0^2/\sigma_0^2]}$$

$$= e^{-(1/2)\left[\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}\right)\mu^2 - 2\left(\frac{\mu_0}{\sigma_0^2} + \frac{\bar{x}}{\sigma^2/n}\right)\mu\right]} h_1(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2)$$

Upon completing the square in the exponent

$$f(x_1, x_2, \ldots, x_n, \mu) = e^{-(1/2)\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}\right)\left[\mu^2 - \left(\frac{(\sigma^2/n)\mu_0}{\sigma_0^2 + \sigma^2/n} + \frac{\bar{x}\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\right)\right]^2} h_2(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2)$$

where $h_i(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2)$ is a function of the observed values, σ^2, μ_0, and σ_0^2.

Now, because $f(x_1, \ldots, x_n)$ does not depend on μ,

$$f(\mu \mid x_1, \ldots, x_n) = e^{-(1/2)\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}\right)\left[\mu^2 - \left(\frac{(\sigma^2/n)\mu_0 + \sigma_0^2\bar{x}}{\sigma_0^2 + \sigma^2/n}\right)\right]} h_3(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2)$$

This is recognized as a normal probability density function with posterior mean

$$\frac{(\sigma^2/n)\mu_0 + \sigma_0^2\bar{x}}{\sigma_0^2 + \sigma^2/n}$$

and posterior variance

$$\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n}\right)^{-1} = \frac{\sigma_0^2(\sigma^2/n)}{\sigma_0^2 + \sigma^2/n}$$

Consequently, the Bayes estimate of μ is a weighted average of μ_0 and $\bar{x}$. For purposes of comparison, note that the maximum likelihood estimate of μ is $\hat{\mu} = \bar{x}$.

To illustrate, suppose that we have a sample of size $n = 10$ from a normal distribution with unknown mean μ and variance $\sigma^2 = 4$. Assume that the prior distribution for μ is normal with mean $\mu_0 = 0$ and variance $\sigma_0^2 = 1$. If the sample mean is 0.75, the Bayes estimate of μ is

$$\frac{(4/10)0 + 1(0.75)}{1 + (4/10)} = \frac{0.75}{1.4} = 0.536$$

Note that the maximum likelihood estimate of μ is $\bar{x} = 0.75$.

There is a relationship between the Bayes estimator for a parameter and the maximum likelihood estimator of the same parameter. For large sample sizes, the two are nearly equivalent. In general, the difference between the two estimators is small compared to $1/\sqrt{n}$. In practical problems, a moderate sample size will produce approximately the same estimate by either the Bayes or maximum likelihood method, if the sample results are consistent with the assumed prior information. If the sample results are inconsistent with the prior assumptions, the Bayes estimate may differ considerably from the maximum likelihood estimate. In these circumstances, if the sample results are accepted as being correct, the prior information must be incorrect. The maximum likelihood estimate would then be the better estimate to use.

If the sample results are very different from the prior information, the Bayes estimator will always tend to produce an estimate that is between the maximum likelihood estimate and the prior assumptions. If there is more inconsistency between the prior information and the sample, there will be more difference between the two estimates.

EXERCISES FOR SECTION 7-3.3

S7-1. Suppose that X is a normal random variable with unknown mean μ and known variance σ^2. The prior distribution for μ is a normal distribution with mean μ_0 and variance σ_0^2. Show that the Bayes estimator for μ becomes the maximum likelihood estimator when the sample size n is large.

S7-2. Suppose that X is a normal random variable with unknown mean μ and known variance σ^2. The prior distribution for μ is a uniform distribution defined over the interval $[a, b]$.

(a) Find the posterior distribution for μ.
(b) Find the Bayes estimator for μ.

S7-3. Suppose that X is a Poisson random variable with parameter λ. Let the prior distribution for λ be a gamma distribution with parameters $m + 1$ and $(m + 1)/\lambda_0$.

(a) Find the posterior distribution for λ.
(b) Find the Bayes estimator for λ.

S7-4. Suppose that X is a normal random variable with unknown mean and known variance $\sigma^2 = 9$. The prior distribution for μ is normal with $\mu_0 = 4$ and $\sigma_0^2 = 1$. A random sample of $n = 25$ observations is taken, and the sample mean is $\bar{x} = 4.85$.

(a) Find the Bayes estimate of μ.

(b) Compare the Bayes estimate with the maximum likelihood estimate.

S7-5. The weight of boxes of candy is a normal random variable with mean μ and variance 1/10 pound. The prior distribution for μ is normal with mean 5.03 pound. and variance 1/25 pound. A random sample of 10 boxes gives a sample mean of $\bar{x} = 5.05$ pound.

(a) Find the Bayes estimate of μ.

(b) Compare the Bayes estimate with the maximum likelihood estimate.

S7-6. The time between failures of a machine has an exponential distribution with parameter λ. Suppose that the prior distribution for λ is exponential with mean 100 hours. Two machines are observed, and the average time between failures is $\bar{x} = 1125$ hours.

(a) Find the Bayes estimate for λ.

(b) What proportion of the machines do you think will fail before 1000 hours?