Chapter

8 Sampling Distributions of Estimators

8.1 The Sampling Distribution of a Statistic

A statistic is a function of some observable random variables, and hence is itself a random variable with a distribution. That distribution is its sampling distribution, and it tells us what values the statistic is likely to assume and how likely it is to assume those values prior to observing our data. When the distribution of the observable data is indexed by a parameter, the sampling distribution is specified as the distribution of the statistic for a given value of the parameter.

Statistics and Estimators

Example 8.1.1

A Clinical Trial. In the clinical trial first introduced in Example 2.1.4, let θ stand for the proportion who do not relapse among all possible imipramine patients. We could use the observed proportion of patients without relapse in the imipramine group to estimate θ. Prior to observing the data, the proportion of sampled patients with no relapse is a random variable T that has a distribution and will not exactly equal the parameter θ. However, we hope that T will be close to θ with high probability. For example, we could try to compute the probability that $|T - \theta| < 0.1$. Such calculations require that we know the distribution of the random variable T. In the clinical trial, we modeled the responses of the 40 patients in the imipramine group as conditionally (given θ) i.i.d. Bernoulli random variables with parameter θ. It follows that the conditional distribution of $40T$ given θ is the binomial distribution with parameters 40 and θ. The distribution of T can be derived easily from this. Indeed T has the following p.f. given θ:

$$f(t|\theta) = \binom{40}{40t}\theta^{40t}(1-\theta)^{40(1-t)}, \quad \text{for } t = 0, \tfrac{1}{40}, \tfrac{2}{40}, \ldots, \tfrac{39}{40}, 1,$$

and $f(t|\theta) = 0$ otherwise. ◀

The distribution at the end of Example 8.1.1 is called the *sampling distribution* of the statistic T, and we can use it to help address questions such as how close we expect T to be to θ prior to observing the data. We can also use the sampling distribution of T to help to determine how much we will learn about θ by observing T. If we are

trying to decide which of two different statistics to use as an estimator, their sampling distributions can be useful for helping us to compare them.

The concept of sampling distribution applies to a larger class of random variables than statistics.

Definition 8.1.1 Sampling Distribution. Suppose that the random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from a distribution involving a parameter θ whose value is unknown. Let T be a function of $\boldsymbol{X}$ and possibly θ. That is, $T = r(X_1, \ldots, X_n, \theta)$. The distribution of T (given θ) is called the *sampling distribution* of T. We will use the notation $E_\theta(T)$ to denote the mean of T calculated from its sampling distribution.

The name "sampling distribution" comes from the fact that T depends on a random sample and so its distribution is derived from the distribution of the sample.

Often, the random variable T in Definition 8.1.1 will not depend on θ, and hence will be a statistic as defined in Definition 7.1.4. In particular, if T is an estimator of θ (as defined in Definition 7.4.1), then T is also a statistic because it is a function of $\boldsymbol{X}$. Therefore, in principle, it is possible to derive the sampling distribution of each estimator of θ. In fact, the distributions of many estimators and statistics have already been found in previous parts of this book.

Example 8.1.2 Sampling Distribution of the M.L.E. of the Mean of a Normal Distribution. Supppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. We found in Examples 7.5.5 and 7.5.6 that the sample mean $\overline{X}_n$ is the M.L.E. of μ. Furthermore, it was found in Corollary 5.6.2 that the distribution of $\overline{X}_n$ is the normal distribution with mean μ and variance σ^2/n. ◀

In this chapter, we shall derive, for random samples from a normal distribution, the distribution of the sample variance and the distributions of various functions of the sample mean and the sample variance. These derivations will lead us to the definitions of some new distributions that play important roles in problems of statistical inference. In addition, we shall study certain general properties of estimators and their sampling distributions.

Purpose of the Sampling Distribution

Example 8.1.3 Lifetimes of Electronic Components. Consider the company in Example 7.1.1 that sells electronic components. They model the lifetimes of these components as i.i.d. exponential random variables with parameter θ conditional on θ. They model θ as having the gamma distribution with parameters 1 and 2. Now, suppose that they are about to observe $n = 3$ lifetimes, and they will use the posterior mean of θ as an estimator. According to Theorem 7.3.4, the posterior distribution of θ will be the gamma distribution with parameters $1 + 3 = 4$ and $2 + \sum_{i=1}^{3} X_i$. The posterior mean will then be $\hat{\theta} = 4/(2 + \sum_{i=1}^{3} X_i)$.

Prior to observing the three lifetimes, the company may want to know how likely it is that $\hat{\theta}$ will be close to θ. For example, they may want to compute $\Pr(|\hat{\theta} - \theta| < 0.1)$. In addition, other interested parties such as customers might be interested in how close the estimator is going to be to θ. But these others might not wish to assign the same prior distribution to θ. Indeed, some of them may wish to assign no prior distribution at all. We shall soon see that all of these people will find it useful to determine the sampling distribution of $\hat{\theta}$. What they do with that sampling distribution will differ, but they will all be able to make use of the sampling distribution. ◀

In Example 8.1.3, after the company observes the three lifetimes, they will be interested only in the posterior distribution of θ. They could then compute the posterior probability that $|\hat{\theta} - \theta| < 0.1$. However, before the sample is taken, both $\hat{\theta}$ and θ are random and $\Pr(|\hat{\theta} - \theta| < 0.1)$ involves the joint distribution of $\hat{\theta}$ and θ. The sampling distribution is merely the conditional distribution of $\hat{\theta}$ given θ. Hence, the law of total probability says that

$$\Pr(|\hat{\theta} - \theta| < 0.1) = E\left[\Pr(|\hat{\theta} - \theta| < 0.1|\theta)\right].$$

In this way, the company makes use of the sampling distribution of $\hat{\theta}$ as an intermediate calculation on the way to computing $\Pr(|\hat{\theta} - \theta| < 0.1)$.

Example 8.1.4

Lifetimes of Electronic Components. In Example 8.1.3, the sampling distribution of $\hat{\theta}$ does not have a name, but it is easy to see that $\hat{\theta}$ is a monotone function of the statistic $T = \sum_{i=1}^{3} X_i$ that has the gamma distribution with parameters 3 and θ (conditional on θ). So, we can compute the c.d.f. $F(\cdot|\theta)$ for the sampling distribution of $\hat{\theta}$ from the c.d.f. $G(\cdot|\theta)$ of the distribution of T. Argue as follows. For $t > 0$,

$$\begin{aligned} F(t|\theta) &= \Pr(\hat{\theta} \le t|\theta) \\ &= \Pr\left(\frac{4}{2+T} \le t \middle| \theta\right) \\ &= \Pr\left(T \ge \frac{4}{t} - 2 \middle| \theta\right) \\ &= 1 - G\left(\frac{4}{t} - 2 \middle| \theta\right). \end{aligned}$$

For $t \le 0$, $F(t|\theta) = 0$. Most statistical computer packages include the function G, which is the c.d.f. of a gamma distribution. The company can now compute, for each θ,

$$\Pr(|\hat{\theta} - \theta| < 0.1|\theta) = F(\theta + 0.1|\theta) - F(\theta - 0.1|\theta). \tag{8.1.1}$$

Figure 8.1 shows a graph of this probability as a function of θ. To complete the calculation of $\Pr(|\hat{\theta} - \theta| < 0.1)$, we must integrate (8.1.1) with respect to the distribution of θ, that is, the gamma distribution with parameters 1 and 2. This integral cannot be performed in closed form and requires a numerical approximation. One such approximation would be a simulation, which will be discussed in Chapter 12. In this example, the approximation yields $\Pr(|\hat{\theta} - \theta| < 0.1) \approx 0.478$.

Also included in Fig. 8.1 is the calculation of $\Pr(|\hat{\theta} - \theta| < 0.1|\theta)$ using $\hat{\theta} = 3/T$, the M.L.E. of θ. The sampling distribution of the M.L.E. can be derived in Exercise 9 at the end of this section. Notice that the posterior mean has higher probability of being close to θ than does the M.L.E. when θ is near the mean of the prior distribution. When θ is far from the prior mean, the M.L.E. has higher probability of being close to θ. ◀

Another case in which the sampling distribution of an estimator is needed is when the statistician must decide which one of two or more available experiments should be performed in order to obtain the best estimator of θ. For example, if she must choose which sample size to use for an experiment, then she will typically base her decision on the sampling distributions of the different estimators that might be used for each sample size.

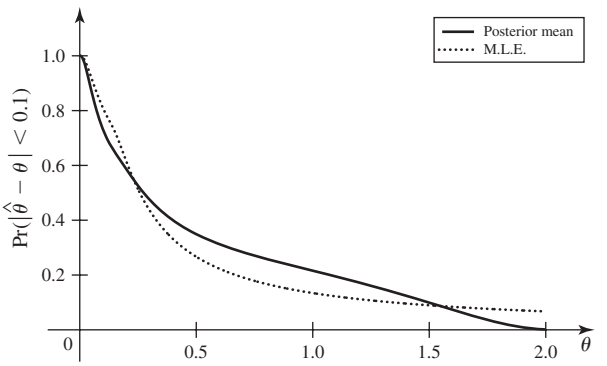

Figure 8.1 Plot of $\Pr(|\hat{\theta} - \theta| < 0.1|\theta)$ for both $\hat{\theta}$ equal to the posterior mean and $\hat{\theta}$ equal to the M.L.E. in Example 8.1.4.

As mentioned at the end of Example 8.1.3, there are statisticians who do not wish to assign a prior distribution to θ. Those statisticians would not be able to calculate a posterior distribution for θ. Instead, they would base all of their statistical inferences on the sampling distribution of whatever estimators they chose. For example, a statistician who chose to use the M.L.E. of θ in Example 8.1.4 would need to deal with the entire curve in Fig. 8.1 corresponding to the M.L.E. in order to decide how likely it is that the M.L.E. will be closer to θ than 0.1. Alternatively, she might choose a different measure of how close the M.L.E. is to θ.

Example 8.1.5

Lifetimes of Electronic Components. Suppose that a statistician chooses to estimate θ by the M.L.E., $\hat{\theta} = 3/T$ instead of the posterior mean in Example 8.1.4. This statistician may not find the graph in Fig. 8.1 very useful unless she can decide which θ values are most important to consider. Instead of calculating $\Pr(|\hat{\theta} - \theta| < 0.1|\theta)$, she might compute

$$\Pr\left(\left|\frac{\hat{\theta}}{\theta} - 1\right| < 0.1 \middle| \theta\right). \tag{8.1.2}$$

This is the probability that $\hat{\theta}$ is within 10% of the value of θ. The probability in (8.1.2) could be computed from the sampling distribution of the M.L.E. Alternatively, one can notice that $\hat{\theta}/\theta = 3/(\theta T)$, and the distribution of θT is the gamma distribution with parameters 3 and 1. Hence, $\hat{\theta}/\theta$ has a distribution that does not depend on θ. It follows that $\Pr(|\hat{\theta}/\theta - 1| < 0.1|\theta)$ is the same number for all θ. In the notation of Example 8.1.4, the c.d.f. of θT is $G(\cdot|1)$, and hence

$$\begin{aligned}
\Pr\left(\left|\frac{\hat{\theta}}{\theta} - 1\right| < 0.1 \middle| \theta\right) &= \Pr\left(\left|\frac{3}{\theta T} - 1\right| < 0.1 \middle| \theta\right) \\
&= \Pr\left(0.9 < \frac{3}{\theta T} < 1.1 \middle| \theta\right) \\
&= \Pr(2.73 < \theta T < 3.33|\theta) \\
&= G(3.33|1) - G(2.73|1) = 0.134.
\end{aligned}$$

The statistician can now claim that the probability is 0.134 that the M.L.E. of θ will be within 10% of the value of θ, no matter what θ is. ◀

The random variable $\hat{\theta}/\theta$ in Example 8.1.5 is an example of a *pivotal quantity,* which will be defined and used extensively in Sec. 8.5.

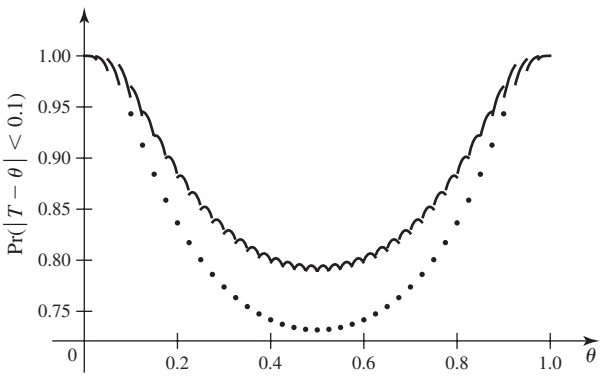

Figure 8.2 Plot of $\Pr(|T - \theta| < 0.1|\theta)$ in Example 8.1.6.

Example 8.1.6 A Clinical Trial. In Example 8.1.1, we found the sampling distribution of T, the proportion of patients without relapse in the imipramine group. Using that distribution, we can draw a plot similar to that in Fig. 8.1. That is, for each θ, we can compute $\Pr(|T - \theta| < 0.1|\theta)$. The plot appears in Fig. 8.2. The jumps and cyclic nature of the plot are due the discreteness of the distribution of T. The smallest probability is 0.7318 at $\theta = 0.5$. (The isolated points that appear below the main part of the graph at θ equal to each multiple of 1/40 would appear equally far above the main part of the graph, if we had plotted $\Pr(|T - \theta| \le 0.1|\theta)$ instead of $\Pr(|T - \theta| < 0.1|\theta)$.) ◀

Summary

The sampling distribution of an estimator $\hat{\theta}$ is the conditional distribution of the estimator given the parameter. The sampling distribution can be used as an intermediate calculation in assessing the properties of a Bayes estimator prior to observing data. More commonly, the sampling distribution is used by those statisticians who prefer not to use prior and posterior distributions. For example, before the sample has been taken, the statistician can use the sampling distribution of $\hat{\theta}$ to calculate the probability that $\hat{\theta}$ will be close to θ. If this probability is high for every possible value of θ, then the statistician can feel confident that the observed value of $\hat{\theta}$ will be close to θ. After the data are observed and a particular estimate is obtained, the statistician would like to continue feeling confident that the particular estimate is likely to be close to θ, even though explicit posterior probabilities cannot be given. It is not always safe to draw such a conclusion, however, as we shall illustrate at the end of Example 8.5.11.

Exercises

1. Suppose that a random sample $X_1, \ldots, X_n$ is to be taken from the uniform distribution on the interval $[0, \theta]$ and that θ is unknown. How large a random sample must be taken in order that

$$\Pr\bigl(|\max\{X_1, \ldots, X_n\} - \theta| \le 0.1\theta\bigr) \ge 0.95,$$

for all possible θ?

2. Suppose that a random sample is to be taken from the normal distribution with unknown mean θ and standard deviation 2. How large a random sample must be taken in order that $E_\theta(|\overline{X}_n - \theta|^2) \le 0.1$ for every possible value of θ?

3. For the conditions of Exercise 2, how large a random sample must be taken in order that $E_\theta(|\overline{X}_n - \theta|) \le 0.1$ for every possible value of θ?

4. For the conditions of Exercise 2, how large a random sample must be taken in order that $\Pr(|\overline{X}_n - \theta| \le 0.1) \ge 0.95$ for every possible value of θ?

5. Suppose that a random sample is to be taken from the Bernoulli distribution with unknown parameter p. Suppose also that it is believed that the value of p is in the neighborhood of 0.2. How large a random sample must be taken in order that $\Pr(|\overline{X}_n - p| \le 0.1) \ge 0.75$ when $p = 0.2$?

6. For the conditions of Exercise 5, use the central limit theorem in Sec. 6.3 to find approximately the size of a random sample that must be taken in order that $\Pr(|\overline{X}_n - p| \le 0.1) \ge 0.95$ when $p = 0.2$.

7. For the conditions of Exercise 5, how large a random sample must be taken in order that $E_p(|\overline{X}_n - p|^2) \le 0.01$ when $p = 0.2$?

8. For the conditions of Exercise 5, how large a random sample must be taken in order that $E_p(|\overline{X}_n - p|^2) \le 0.01$ for every possible value of p $(0 \le p \le 1)$?

9. Let $X_1, \ldots, X_n$ be a random sample from the exponential distribution with parameter θ. Find the c.d.f. for the sampling distribution of the M.L.E. of θ. (The M.L.E. itself was found in Exercise 7 in Sec. 7.5.)

8.2 The Chi-Square Distributions

The family of chi-square (χ^2) distributions is a subcollection of the family of gamma distributions. These special gamma distributions arise as sampling distributions of variance estimators based on random samples from a normal distribution.

Definition of the Distributions

Example 8.2.1 M.L.E. of the Variance of a Normal Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with known mean μ and unknown variance σ^2. The M.L.E. of σ^2 is found in Exercise 6 in Sec. 7.5. It is

$$\widehat{\sigma_0^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

The distributions of $\widehat{\sigma_0^2}$ and $\widehat{\sigma_0^2}/\sigma^2$ are useful in several statistical problems, and we shall derive them in this section. ◀

In this section, we shall introduce and discuss a particular class of gamma distributions known as the chi-square (χ^2) distributions. These distributions, which are closely related to random samples from a normal distribution, are widely applied in the field of statistics. In the remainder of this book, we shall see how they are applied in many important problems of statistical inference. In this section, we shall present the definition of the χ^2 distributions and some of their basic mathematical properties.

Definition 8.2.1 χ^2 Distributions. For each positive number m, the gamma distribution with parameters $\alpha = m/2$ and $\beta = 1/2$ is called the χ^2 *distribution with m degrees of freedom*. (See Definition 5.7.2 for the definition of the gamma distribution with parameters α and β.)

It is common to restrict the degrees of freedom m in Definition 8.2.1 to be an integer. However, there are situations in which it will be useful for the degrees of freedom to not be integers, so we will not make that restriction.

If a random variable X has the χ^2 distribution with m degrees of freedom, it follows from Eq. (5.7.13) that the p.d.f. of X for $x > 0$ is

$$f(x) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{(m/2)-1} e^{-x/2}. \tag{8.2.1}$$

Also, $f(x) = 0$ for $x \leq 0$.

A short table of p quantiles for the χ^2 distribution for various values of p and various degrees of freedom is given at the end of this book. Most statistical software packages include functions to compute the c.d.f. and the quantile function of an arbitrary χ^2 distribution.

It follows from Definition 8.2.1, and it can be seen from Eq. (8.2.1), that the χ^2 distribution with two degrees of freedom is the exponential distribution with parameter 1/2 or, equivalently, the exponential distribution for which the mean is 2. Thus, the following three distributions are all the same: the gamma distribution with parameters $\alpha = 1$ and $\beta = 1/2$, the χ^2 distribution with two degrees of freedom, and the exponential distribution for which the mean is 2.

Properties of the Distributions

The means and variances of χ^2 distributions follow immediately from Theorem 5.7.5, and are given here without proof.

Theorem 8.2.1 Mean and Variance. If a random variable X has the χ^2 distribution with m degrees of freedom, then $E(X) = m$ and $\text{Var}(X) = 2m$. ■

Furthermore, it follows from the moment generating function given in Eq. (5.7.15) that the m.g.f. of X is

$$\psi(t) = \left(\frac{1}{1-2t}\right)^{m/2} \quad \text{for } t < \frac{1}{2}.$$

The additivity property of the χ^2 distribution, which is presented without proof in the next theorem, follows directly from Theorem 5.7.7.

Theorem 8.2.2 If the random variables $X_1, \ldots, X_k$ are independent and if X_i has the χ^2 distribution with m_i degrees of freedom ($i = 1, \ldots, k$), then the sum $X_1 + \cdots + X_k$ has the χ^2 distribution with $m_1 + \cdots + m_k$ degrees of freedom. ■

We shall now establish the basic relation between the χ^2 distributions and the standard normal distribution.

Theorem 8.2.3 Let X have the standard normal distribution. Then the random variable $Y = X^2$ has the χ^2 distribution with one degree of freedom.

Proof Let $f(y)$ and $F(y)$ denote, respectively, the p.d.f. and the c.d.f. of Y. Also, since X has the standard normal distribution, we shall let $\phi(x)$ and $\Phi(x)$ denote the p.d.f. and the c.d.f. of X. Then for $y > 0$,

$$\begin{aligned} F(y) &= \Pr(Y \leq y) = \Pr(X^2 \leq y) = \Pr(-y^{1/2} \leq X \leq y^{1/2}) \\ &= \Phi(y^{1/2}) - \Phi(-y^{1/2}). \end{aligned}$$

Since $f(y) = F'(y)$ and $\phi(x) = \Phi'(x)$, it follows from the chain rule for derivatives that

$$f(y) = \phi(y^{1/2})\left(\frac{1}{2}y^{-1/2}\right) + \phi(-y^{1/2})\left(\frac{1}{2}y^{-1/2}\right).$$

Furthermore, since $\phi(y^{1/2}) = \phi(-y^{1/2}) = (2\pi)^{-1/2}e^{-y/2}$, it now follows that

$$f(y) = \frac{1}{(2\pi)^{1/2}}y^{-1/2}e^{-y/2} \quad \text{for } y > 0.$$

By comparing this equation with Eq. (8.2.1), it is seen that the p.d.f. of Y is indeed the p.d.f. of the χ^2 distribution with one degree of freedom. ■

We can now combine Theorem 8.2.3 with Theorem 8.2.2 to obtain the following result, which provides the main reason that the χ^2 distribution is important in statistics.

Corollary 8.2.1 If the random variables $X_1, \ldots, X_m$ are i.i.d. with the standard normal distribution, then the sum of squares $X_1^2 + \cdots + X_m^2$ has the χ^2 distribution with m degrees of freedom. ■

Example 8.2.2 M.L.E. of the Variance of a Normal Distribution. In Example 8.2.1, the random variables $Z_i = (X_i - \mu)/\sigma$ for $i = 1, \ldots, n$ form a random sample from the standard normal distribution. It follows from Corollary 8.2.1 that the distribution of $\sum_{i=1}^n Z_i^2$ is the χ^2 distribution with n degrees of freedom. It is easy to see that $\sum_{i=1}^n Z_i^2$ is precisely the same as $n\widehat{\sigma_0^2}/\sigma^2$, which appears in Example 8.2.1. So the distribution of $n\widehat{\sigma_0^2}/\sigma^2$ is the χ^2 distribution with n degrees of freedom. The reader should also be able to see that the distribution of $\widehat{\sigma_0^2}$ itself is the gamma distribution with parameters $n/2$ and $n/(2\sigma^2)$ (Exercise 13). ◀

Example 8.2.3 Acid Concentration in Cheese. Moore and McCabe (1999, p. D-1) describe an experiment conducted in Australia to study the relationship between taste and the chemical composition of cheese. One chemical whose concentration can affect taste is lactic acid. Cheese manufacturers who want to establish a loyal customer base would like the taste to be about the same each time a customer purchases the cheese. The variation in concentrations of chemicals like lactic acid can lead to variation in the taste of cheese. Suppose that we model the concentration of lactic acid in several chunks of cheese as independent normal random variables with mean μ and variance σ^2. We are interested in how much these concentrations differ from the value μ. Let $X_1, \ldots, X_k$ be the concentrations in k chunks, and let $Z_i = (X_i - \mu)/\sigma$. Then

$$Y = \frac{1}{k}\sum_{i=1}^{k}|X_i - \mu|^2 = \frac{\sigma^2}{k}\sum_{i=1}^{k}Z_i^2$$

is one measure of how much the k concentrations differ from μ. Suppose that a difference of u or more in lactic acid concentration is enough to cause a noticeable difference in taste. We might then wish to calculate $\Pr(Y \le u^2)$. According to Corollary 8.2.1, the distribution of $W = kY/\sigma^2$ is χ^2 with k degrees of freedom. Hence, $\Pr(Y \le u^2) = \Pr(W \le ku^2/\sigma^2)$.

For example, suppose that $\sigma^2 = 0.09$, and we are interested in $k = 10$ cheese chunks. Furthermore, suppose that $u = 0.3$ is the critical difference of interest. We

can write

$$\Pr(Y \le 0.3^2) = \Pr\left(W \le \frac{10 \times 0.09}{0.09}\right) = \Pr(W \le 10). \tag{8.2.2}$$

Using the table of quantiles of the χ^2 distribution with 10 degrees of freedom, we see that 10 is between the 0.5 and 0.6 quantiles. In fact, the probability in Eq. (8.2.2) can be found by computer software to equal 0.56, so there is a 44 percent chance that the average squared difference between lactic acid concentration and mean concentration in 10 chunks will be more than the desired amount. If this probability is too large, the manufacturer might wish to invest some effort in reducing the variance of lactic acid concentration. ◀

Summary

The chi-square distribution with n degrees of freedom is the same as the gamma distribution with parameters $m/2$ and $1/2$. It is the distribution of the sum of squares of a sample of m independent standard normal random variables. The mean of the χ^2 distribution with m degrees of freedom is m, and the variance is $2m$.

Exercises

1. Suppose that we will sample 20 chunks of cheese in Example 8.2.3. Let $T = \sum_{i=1}^{20}(X_i - \mu)^2/20$, where X_i is the concentration of lactic acid in the ith chunk. Assume that $\sigma^2 = 0.09$. What number c satisfies $\Pr(T \le c) = 0.9$?

2. Find the mode of the χ^2 distribution with m degrees of freedom ($m = 1, 2, \ldots$).

3. Sketch the p.d.f. of the χ^2 distribution with m degrees of freedom for each of the following values of m. Locate the mean, the median, and the mode on each sketch. **(a)** $m = 1$; **(b)** $m = 2$; **(c)** $m = 3$; **(d)** $m = 4$.

4. Suppose that a point (X, Y) is to be chosen at random in the xy-plane, where X and Y are independent random variables and each has the standard normal distribution. If a circle is drawn in the xy-plane with its center at the origin, what is the radius of the smallest circle that can be chosen in order for there to be probability 0.99 that the point (X, Y) will lie inside the circle?

5. Suppose that a point (X, Y, Z) is to be chosen at random in three-dimensional space, where X, Y, and Z are independent random variables and each has the standard normal distribution. What is the probability that the distance from the origin to the point will be less than 1 unit?

6. When the motion of a microscopic particle in a liquid or a gas is observed, it is seen that the motion is irregular because the particle collides frequently with other particles. The probability model for this motion, which is called *Brownian motion,* is as follows: A coordinate system is chosen in the liquid or gas. Suppose that the particle is at the origin of this coordinate system at time $t = 0$, and let (X, Y, Z) denote the coordinates of the particle at any time $t > 0$. The random variables X, Y, and Z are i.i.d., and each of them has the normal distribution with mean 0 and variance $\sigma^2 t$. Find the probability that at time $t = 2$ the particle will lie within a sphere whose center is at the origin and whose radius is 4σ.

7. Suppose that the random variables $X_1, \ldots, X_n$ are independent, and each random variable X_i has a continuous c.d.f. F_i. Also, let the random variable Y be defined by the relation $Y = -2\sum_{i=1}^{n} \log F_i(X_i)$. Show that Y has the χ^2 distribution with $2n$ degrees of freedom.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, 1]$, and let W denote the range of the sample, as defined in Example 3.9.7. Also, let $g_n(x)$ denote the p.d.f. of the random variable $2n(1 - W)$, and let $g(x)$ denote the p.d.f. of the χ^2 distribution with four degrees of freedom. Show that

$$\lim_{n\to\infty} g_n(x) = g(x) \quad \text{for } x > 0.$$

9. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Find the distribution of

$$\frac{n(\overline{X}_n - \mu)^2}{\sigma^2}.$$

10. Suppose that six random variables $X_1, \ldots, X_6$ form a random sample from the standard normal distribution, and let

$$Y = (X_1 + X_2 + X_3)^2 + (X_4 + X_5 + X_6)^2.$$

Determine a value of c such that the random variable cY will have a χ^2 distribution.

11. If a random variable X has the χ^2 distribution with m degrees of freedom, then the distribution of $X^{1/2}$ is called a *chi* (χ) *distribution with m degrees of freedom*. Determine the mean of this distribution.

12. Consider again the situation described in Example 8.2.3. How small would σ^2 need to be in order for $\Pr(Y \leq 0.09) \geq 0.9$?

13. Prove that the distribution of $\widehat{\sigma_0^2}$ in Examples 8.2.1 and 8.2.2 is the gamma distribution with parameters $n/2$ and $n/(2\sigma^2)$.

8.3 Joint Distribution of the Sample Mean and Sample Variance

Suppose that our data form a random sample from a normal distribution. The sample mean $\hat{\mu}$ and sample variance $\widehat{\sigma^2}$ are important statistics that are computed in order to estimate the parameters of the normal distribution. Their marginal distributions help us to understand how good each of them is as an estimator of the corresponding parameter. However, the marginal distribution of $\hat{\mu}$ depends on σ. The joint distribution of $\hat{\mu}$ and $\widehat{\sigma^2}$ will allow us to make inferences about μ without reference to σ.

Independence of the Sample Mean and Sample Variance

Example 8.3.1

Rain from Seeded Clouds. Simpson, Olsen, and Eden (1975) describe an experiment in which a random sample of 26 clouds were seeded with silver nitrate to see if they produced more rain than unseeded clouds. Suppose that, on a log scale, unseeded clouds typically produced a mean rainfall of 4. In comparing the mean of the seeded clouds to the unseeded mean, one might naturally see how far the average log-rainfall of the seeded clouds $\hat{\mu}$ is from 4. But the variation in rainfall within the sample is also important. For example, if one compared two different samples of seeded clouds, one would expect the average rainfalls in the two samples to be different just due to variation between clouds. In order to be confident that seeding the clouds really produced more rain, we would want the average log-rainfall to exceed 4 by a large amount compared to the variation between samples, which is closely related to the variation within samples. Since we do not know the variance for seeded clouds, we compute the sample variance $\widehat{\sigma^2}$. Comparing $\hat{\mu} - 4$ to $\widehat{\sigma^2}$ requires us to consider the joint distribution of the sample mean and the sample variance. ◀

Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Then, as was shown in Example 7.5.6, the M.L.E.'s of μ and σ^2 are the sample mean $\overline{X}_n$ and the sample variance $(1/n)\sum_{i=1}^n (X_i - \overline{X}_n)^2$. In this section, we shall derive the joint distribution of these two estimators.

We already know from Corollary 5.6.2 that the sample mean itself has the normal distribution with mean μ and variance σ^2/n. We shall establish the noteworthy property that the sample mean and the sample variance are independent random variables, even though both are functions of the same random variables $X_1, \ldots, X_n$. Furthermore, we shall show that, except for a scale factor, the sample variance has the χ^2 distribution with $n - 1$ degrees of freedom. More precisely, we shall show that the random variable $\sum_{i=1}^n (X_i - \overline{X}_n)^2/\sigma^2$ has the χ^2 distribution with $n - 1$ degrees

of freedom. This result is also a rather striking property of random samples from a normal distribution, as the following discussion indicates.

Because the random variables $X_1, \ldots, X_n$ are independent, and because each has the normal distribution with mean μ and variance σ^2, the random variables $(X_1 - \mu)/\sigma, \ldots, (X_n - \mu)/\sigma$ are also independent, and each of these variables has the standard normal distribution. It follows from Corollary 8.2.1 that the sum of their squares $\sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2$ has the χ^2 distribution with n degrees of freedom. Hence, the striking property mentioned in the previous paragraph is that if the population mean μ is replaced by the sample mean $\overline{X}_n$ in this sum of squares, the effect is simply to reduce the degrees of freedom in the χ^2 distribution from n to $n - 1$. In summary, we shall establish the following theorem.

Theorem 8.3.1

Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Then the sample mean $\overline{X}_n$ and the sample variance $(1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ are independent random variables, $\overline{X}_n$ has the normal distribution with mean μ and variance σ^2/n, and $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2/\sigma^2$ has the χ^2 distribution with $n - 1$ degrees of freedom.

Furthermore, it can be shown that the sample mean and the sample variance are independent *only* when the random sample is drawn from a normal distribution. We shall not consider this result further in this book. However, it does emphasize the fact that the independence of the sample mean and the sample variance is indeed a noteworthy property of samples from a normal distribution.

The proof of Theorem 8.3.1 makes use of transformations of several variables as described in Sec. 3.9 and of the properties of orthogonal matrices. The proof appears at the end of this section.

Example 8.3.2

Rain from Seeded Clouds. Figure 8.3 is a histogram of the logarithms of the rainfalls from the seeded clouds in Example 8.3.1. Suppose that these logarithms $X_1, \ldots, X_{26}$ are modeled as i.i.d. normal random variables with mean μ and variance σ^2. If we are interested in how much variation there is in rainfall among the seeded clouds, we can compute the sample variance $\widehat{\sigma^2} = \sum_{i=1}^{26}(X_i - \overline{X}_n)^2/26$. The distribution of $U = 26\widehat{\sigma^2}/\sigma^2$ is the χ^2 distribution with 25 degrees of freedom. We can use this distribution to tell us how likely it is that $\widehat{\sigma^2}$ will overestimate or underestimate σ^2 by various amounts. For example, the χ^2 table in this book says that the 0.25 quantile of the χ^2 distribution with 25 degrees of freedom is 19.94, so $\Pr(U \leq 19.94) = 0.25$.

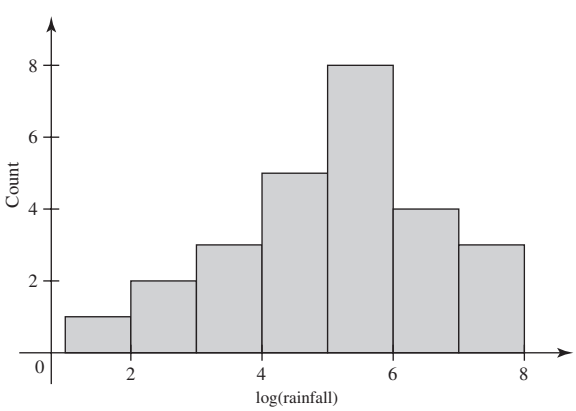

Figure 8.3 Histogram of log-rainfalls from seeded clouds.

It follows that

$$0.25 = \Pr\left(\frac{\widehat{\sigma^2}}{\sigma^2} \le \frac{19.94}{26}\right) = \Pr(\widehat{\sigma^2} \le 0.77\sigma^2). \tag{8.3.1}$$

That is, there is probability 0.25 that $\widehat{\sigma^2}$ will underestimate σ^2 by 23 percent or more. The observed value of $\widehat{\sigma^2}$ is 2.460 in this example. The probability calculated in Eq. (8.3.1) has nothing to do with how far 2.460 is from σ^2. Eq. (8.3.1) tells us the probability (prior to observing the data) that $\widehat{\sigma^2}$ would be at least 23% below σ^2. ◀

Estimation of the Mean and Standard Deviation

We shall assume that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown standard deviation σ. Also, as usual, we shall denote the M.L.E.'s of μ and σ by $\hat{\mu}$ and $\hat{\sigma}$. Thus,

$$\hat{\mu} = \overline{X}_n \quad \text{and} \quad \hat{\sigma} = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right)^{1/2}.$$

Notice that $\hat{\sigma}^2 = \widehat{\sigma^2}$, the M.L.E. of σ^2. For the remainder of this book, when referring to the M.L.E. of σ^2, we shall use whichever symbol $\hat{\sigma}^2$ or $\widehat{\sigma^2}$ is more convenient. As an illustration of the application of Theorem 8.3.1, we shall now determine the smallest possible sample size n such that the following relation will be satisfied:

$$\Pr\left(|\hat{\mu} - \mu| \le \frac{1}{5}\sigma \quad \text{and} \quad |\hat{\sigma} - \sigma| \le \frac{1}{5}\sigma\right) \ge \frac{1}{2}. \tag{8.3.2}$$

In other words, we shall determine the minimum sample size n for which the probability will be at least 1/2 that neither $\hat{\mu}$ nor $\hat{\sigma}$ will differ from the unknown value it is estimating by more than $(1/5)\sigma$.

Because of the independence of $\hat{\mu}$ and $\hat{\sigma}$, the relation (8.3.2) can be rewritten as follows:

$$\Pr\left(|\hat{\mu} - \mu| < \frac{1}{5}\sigma\right)\Pr\left(|\hat{\sigma} - \sigma| < \frac{1}{5}\sigma\right) \ge \frac{1}{2}. \tag{8.3.3}$$

If we let p_1 denote the first probability on the left side of the relation (8.3.3), and let U be a random variable that has the standard normal distribution, this probability can be written in the following form:

$$p_1 = \Pr\left(\frac{\sqrt{n}|\hat{\mu} - \mu|}{\sigma} < \frac{1}{5}\sqrt{n}\right) = \Pr\left(|U| < \frac{1}{5}\sqrt{n}\right).$$

Similarly, if we let p_2 denote the second probability on the left side of the relation (8.3.3), and let $V = n\hat{\sigma}^2/\sigma^2$, this probability can be written in the following form:

$$p_2 = \Pr\left(0.8 < \frac{\hat{\sigma}}{\sigma} < 1.2\right) = \Pr\left(0.64n < \frac{n\hat{\sigma}^2}{\sigma^2} < 1.44n\right)$$
$$= \Pr(0.64n < V < 1.44n).$$

By Theorem 8.3.1, the random variable V has the χ^2 distribution with $n - 1$ degrees of freedom.

For each specific value of n, the values of p_1 and p_2 can be found, at least approximately, from the table of the standard normal distribution and the table of the χ^2 distribution given at the end of this book. In particular, after various values

of n have been tried, it will be found that for $n = 21$ the values of p_1 and p_2 are $p_1 = 0.64$ and $p_2 = 0.78$. Hence, $p_1 p_2 = 0.50$, and it follows that the relation (8.3.2) will be satisfied for $n = 21$.

Proof of Theorem 8.3.1

We already knew from Corollary 5.6.2 that the distribution of the sample mean was as stated in Theorem 8.3.1. What remains to prove is the stated distribution of the sample variance and the independence of the sample mean and sample variance.

Orthogonal Matrices

We begin with some properties of orthogonal matrices that are essential for the proof.

Definition 8.3.1 Orthogonal Matrix. It is said that an $n \times n$ matrix $\boldsymbol{A}$ is *orthogonal* if $\boldsymbol{A}^{-1} = \boldsymbol{A}'$, where $\boldsymbol{A}'$ is the transpose of $\boldsymbol{A}$.

In other words, a matrix $\boldsymbol{A}$ is orthogonal if and only if $\mathbf{A}\mathbf{A}' = \mathbf{A}'\mathbf{A} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix. It follows from this latter property that a matrix is orthogonal if and only if the sum of the squares of the elements in each row is 1 and the sum of the products of the corresponding elements in every pair of different rows is 0. Alternatively, a matrix is orthogonal if and only if the sum of the squares of the elements in each column is 1 and the sum of the products of the corresponding elements in every pair of different columns is 0.

Properties of Orthogonal Matrices We shall now derive two important properties of orthogonal matrices.

Theorem 8.3.2 Determinant is 1. If $\boldsymbol{A}$ is orthogonal, then $|\det \boldsymbol{A}| = 1$.

Proof To prove this result, it should be recalled that $\det \boldsymbol{A} = \det \boldsymbol{A}'$ for every square matrix $\boldsymbol{A}$. Also recall that $\det \boldsymbol{AB} = (\det \boldsymbol{A})(\det \boldsymbol{B})$ for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Therefore,

$$\det(\boldsymbol{A}\boldsymbol{A}') = (\det \boldsymbol{A})(\det \boldsymbol{A}') = (\det \boldsymbol{A})^2.$$

Also, if $\boldsymbol{A}$ is orthogonal, then $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$, and it follows that

$$\det(\boldsymbol{A}\boldsymbol{A}') = \det \boldsymbol{I} = 1.$$

Hence $(\det \boldsymbol{A})^2 = 1$ or, equivalently, $|\det \boldsymbol{A}| = 1$. ■

Theorem 8.3.3 Squared Length Is Preserved. Consider two n-dimensional random vectors

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \tag{8.3.4}$$

and suppose that $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{A}$ is an orthogonal matrix. Then

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2. \tag{8.3.5}$$

Proof This result follows from the fact that $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}$, because

$$\sum_{i=1}^{N} Y_i^2 = \boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{X} = \sum_{i=1}^{n} X_i^2. \quad \blacksquare$$

Multiplication of a vector $\boldsymbol{X}$ by an orthogonal matrix $\boldsymbol{A}$ corresponds to a rotation of $\boldsymbol{X}$ in n-dimensional space possibly followed by changing the signs of some coordinates. Neither of these operations can change the length of the original vector $\boldsymbol{X}$, and that length equals $(\sum_{i=1}^{n} X_i^2)^{1/2}$.

Together, these two properties of orthogonal matrices imply that if a random vector $\boldsymbol{Y}$ is obtained from a random vector $\boldsymbol{X}$ by an *orthogonal* linear transformation $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, then the absolute value of the Jacobian of the transformation is 1 and $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2$.

We combine Theorems 8.3.2 and 8.3.3 to obtain a useful fact about orthogonal transformations of a random sample of standard normal random variables.

Theorem 8.3.4 Suppose that the random variables, $X_1, \ldots, X_n$ are i.i.d. and each has the standard normal distribution. Suppose also that $\boldsymbol{A}$ is an orthogonal $n \times n$ matrix, and $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$. Then the random variables $Y_1, \ldots, Y_n$ are also i.i.d., each also has the standard normal distribution, and $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i^2$.

Proof The joint p.d.f. of $X_1, \ldots, X_n$ is as follows, for $-\infty < x_i < \infty$ $(i = 1, \ldots, n)$:

$$f_n(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right). \tag{8.3.6}$$

If $\boldsymbol{A}$ is an orthogonal $n \times n$ matrix, and the random variables $Y_1, \ldots, Y_n$ are defined by the relation $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, where the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are as specified in Eq. (8.3.4). This is a linear transformation, so the joint p.d.f. of $Y_1, \ldots, Y_n$ is obtained from Eq. (3.9.20) and equals

$$g_n(\boldsymbol{y}) = \frac{1}{|\det \boldsymbol{A}|} f_n(\boldsymbol{A}^{-1}\boldsymbol{y}).$$

Let $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$. Since $\boldsymbol{A}$ is orthogonal, $|\det \boldsymbol{A}| = 1$ and $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} x_i^2$, as we just proved. So,

$$g_n(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i^2\right). \tag{8.3.7}$$

It can be seen from Eq. (8.3.7) that the joint p.d.f. of $Y_1, \ldots, Y_n$ is exactly the same as the joint p.d.f. of $X_1, \ldots, X_n$. $\blacksquare$

Proof of Theorem 8.3.1

Random Samples from the Standard Normal Distribution We shall begin by proving Theorem 8.3.1 under the assumption that $X_1, \ldots, X_n$ form a random sample from the standard normal distribution. Consider the n-dimensional row vector $\boldsymbol{u}$, in which each of the n components has the value $1/\sqrt{n}$:

$$\boldsymbol{u} = \left[\frac{1}{\sqrt{n}} \cdots \frac{1}{\sqrt{n}}\right]. \tag{8.3.8}$$

Since the sum of the squares of the n components of the vector $\boldsymbol{u}$ is 1, it is possible to construct an orthogonal matrix $\boldsymbol{A}$ such that the components of the vector $\boldsymbol{u}$ form

the first row of $\boldsymbol{A}$. This construction, called the *Gram-Schmidt method*, is described in textbooks on linear algebra such as Cullen (1972) and will not be discussed here. We shall assume that such a matrix $\boldsymbol{A}$ has been constructed, and we shall again define the random variables $Y_1, \ldots, Y_n$ by the transformation $\boldsymbol{Y} = \boldsymbol{AX}$.

Since the components of $\boldsymbol{u}$ form the first row of $\boldsymbol{A}$, it follows that

$$Y_1 = \boldsymbol{uX} = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} X_i = \sqrt{n}\, \overline{X}_n. \tag{8.3.9}$$

Furthermore, by Theorem 8.3.4, $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i^2$. Therefore,

$$\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

We have thus obtained the relation

$$\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2. \tag{8.3.10}$$

It is known from Theorem 8.3.4 that the random variables $Y_1, \ldots, Y_n$ are independent. Therefore, the two random variables Y_1 and $\sum_{i=2}^{n} Y_i^2$ are independent, and it follows from Eqs. (8.3.9) and (8.3.10) that $\overline{X}_n$ and $\sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ are independent. Furthermore, it is known from Theorem 8.3.4 that the $n - 1$ random variables $Y_2, \ldots, Y_n$ are i.i.d., and that each of these random variables has the standard normal distribution. Hence, by Corollary 8.2.1 the random variable $\sum_{i=2}^{n} Y_i^2$ has the χ^2 distribution with $n - 1$ degrees of freedom. It follows from Eq. (8.3.10) that $\sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ also has the χ^2 distribution with $n - 1$ degrees of freedom.

Random Samples from an Arbitrary Normal Distribution Thus far, in proving Theorem 8.3.1, we have considered only random samples from the standard normal distribution. Suppose now that the random variables $X_1, \ldots, X_n$ form a random sample from an arbitrary normal distribution with mean μ and variance σ^2.

If we let $Z_i = (X_i - \mu)/\sigma$ for $i = 1, \ldots, n$, then the random variables $Z_1, \ldots, Z_n$ are independent, and each has the standard normal distribution. In other words, the joint distribution of $Z_1, \ldots, Z_n$ is the same as the joint distribution of a random sample from the standard normal distribution. It follows from the results that have just been obtained that $\overline{Z}_n$ and $\sum_{i=1}^{n} (Z_i - \overline{Z}_n)^2$ are independent, and $\sum_{i=1}^{n} (Z_i - \overline{Z}_n)^2$ has the χ^2 distribution with $n - 1$ degrees of freedom. However, $\overline{Z}_n = (\overline{X}_n - \mu)/\sigma$ and

$$\sum_{i=1}^{n} (Z_i - \overline{Z}_n)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2. \tag{8.3.11}$$

We now conclude that the sample mean $\overline{X}_n$ and the sample variance $(1/n) \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ are independent, and that the random variable on the right side of Eq. (8.3.11) has the χ^2 distribution with $n - 1$ degrees of freedom. All the results stated in Theorem 8.3.1 have now been established.

Summary

Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. Then the sample mean $\hat{\mu} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and sample variance $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ are independent random variables. Furthermore, $\hat{\mu}$ has the normal distribution with mean μ and variance σ^2/n, and $n\widehat{\sigma^2}/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom.

Exercises

1. Assume that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Show that $\widehat{\sigma^2}$ has the gamma distribution with parameters $(n-1)/2$ and $n/(2\sigma^2)$.

2. Determine whether or not each of the five following matrices is orthogonal:

a. $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ **b.** $\begin{bmatrix} 0.8 & 0 & 0.6 \\ -0.6 & 0 & 0.8 \\ 0 & -1 & 0 \end{bmatrix}$

c. $\begin{bmatrix} 0.8 & 0 & 0.6 \\ -0.6 & 0 & 0.8 \\ 0 & 0.5 & 0 \end{bmatrix}$ **d.** $\begin{bmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{bmatrix}$

e. $\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$

3.a. Construct a 2×2 orthogonal matrix for which the first row is as follows:

$$\left[\, \tfrac{1}{\sqrt{2}} \quad \tfrac{1}{\sqrt{2}} \,\right].$$

b. Construct a 3×3 orthogonal matrix for which the first row is as follows:

$$\left[\, \tfrac{1}{\sqrt{3}} \quad \tfrac{1}{\sqrt{3}} \quad \tfrac{1}{\sqrt{3}} \,\right].$$

4. Suppose that the random variables X_1, X_2, and X_3 are i.i.d., and that each has the standard normal distribution. Also, suppose that

$$Y_1 = 0.8X_1 + 0.6X_2,$$
$$Y_2 = \sqrt{2}(0.3X_1 - 0.4X_2 - 0.5X_3),$$
$$Y_3 = \sqrt{2}(0.3X_1 - 0.4X_2 + 0.5X_3).$$

Find the joint distribution of Y_1, Y_2, and Y_3.

5. Suppose that the random variables X_1 and X_2 are independent, and that each has the normal distribution with mean μ and variance σ^2. Prove that the random variables $X_1 + X_2$ and $X_1 - X_2$ are independent.

6. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Assuming that the sample size n is 16, determine the values of the following probabilities:

a. $\Pr\left[\frac{1}{2}\sigma^2 \le \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \le 2\sigma^2\right]$

b. $\Pr\left[\frac{1}{2}\sigma^2 \le \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \le 2\sigma^2\right]$

7. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2, and let $\widehat{\sigma^2}$ denote the sample variance. Determine the smallest values of n for which the following relations are satisfied:

a. $\Pr\left(\frac{\widehat{\sigma^2}}{\sigma^2} \le 1.5\right) \ge 0.95$

b. $\Pr\left(|\widehat{\sigma^2} - \sigma^2| \le \frac{1}{2}\sigma^2\right) \ge 0.8$

8. Suppose that X has the χ^2 distribution with 200 degrees of freedom. Explain why the central limit theorem can be used to determine the approximate value of $\Pr(160 < X < 240)$ and find this approximate value.

9. Suppose that each of two statisticians, A and B, independently takes a random sample of 20 observations from the normal distribution with unknown mean μ and known variance 4. Suppose also that statistician A finds the sample variance in his random sample to be 3.8, and statistician B finds the sample variance in her random sample to be 9.4. For which random sample is the sample mean likely to be closer to the unknown value of μ?

8.4 The *t* Distributions

When our data are a sample from the normal distribution with mean μ and variance σ^2, the distribution of $Z = n^{1/2}(\hat{\mu} - \mu)/\sigma$ is the standard normal distribution, where $\hat{\mu}$ is the sample mean. If σ^2 is unknown, we can replace σ by an estimator (similar to the M.L.E.) in the formula for Z. The resulting random variable has the t distribution with $n - 1$ degrees of freedom and is useful for making inferences about μ alone even when both μ and σ^2 are unknown.

Definition of the Distributions

Example 8.4.1 Rain from Seeded Clouds. Consider the same sample of log-rainfall measurements from 26 seeded clouds from Example 8.3.2. Suppose now that we are interested in how far the sample average $\overline{X}_n$ of those measurements is from the mean μ. We know that $n^{1/2}(\overline{X}_n - \mu)/\sigma$ has the standard normal distribution, but we do not know σ. If we replace σ by an estimator $\hat{\sigma}$ such as the M.L.E., or something similar, what is the distribution of $n^{1/2}(\overline{X}_n - \mu)/\hat{\sigma}$, and how can we make use of this random variable to make inferences about μ? ◀

In this section, we shall introduce and discuss another family of distributions, called the *t* distributions, which are closely related to random samples from a normal distribution. The *t* distributions, like the χ^2 distributions, have been widely applied in important problems of statistical inference. The *t* distributions are also known as Student's distributions (see Student, 1908), in honor of W. S. Gosset, who published his studies of this distribution in 1908 under the pen name "Student." The distributions are defined as follows.

Definition 8.4.1 *t* Distributions. Consider two independent random variables Y and Z, such that Y has the χ^2 distribution with m degrees of freedom and Z has the standard normal distribution. Suppose that a random variable X is defined by the equation

$$X = \frac{Z}{\left(\frac{Y}{m}\right)^{1/2}}. \tag{8.4.1}$$

Then the distribution of X is called the *t distribution with m degrees of freedom*.

The derivation of the p.d.f. of the *t* distribution with m degrees of freedom makes use of the methods of Sec. 3.9 and will be given at the end of this section. But we state the result here.

Theorem 8.4.1 Probability Density Function. The p.d.f. of the *t* distribution with m degrees of freedom is

$$\frac{\Gamma\left(\frac{m+1}{2}\right)}{(m\pi)^{1/2}\Gamma\left(\frac{m}{2}\right)}\left(1 + \frac{x^2}{m}\right)^{-(m+1)/2} \quad \text{for } -\infty < x < \infty. \tag{8.4.2}$$

Moments of the t Distributions Although the mean of the *t* distribution does not exist when $m \leq 1$, the mean does exist for every value of $m > 1$. Of course, whenever the mean does exist, its value is 0 because of the symmetry of the *t* distribution.

In general, if a random variable X has the t distribution with m degrees of freedom ($m > 1$), then it can be shown that $E(|X|^k) < \infty$ for $k < m$ and that $E(|X|^k) = \infty$ for $k \geq m$. If m is an integer, the first $m - 1$ moments of X exist, but no moments of higher order exist. It follows, therefore, that the m.g.f. of X does not exist.

It can be shown (see Exercise 1 at the end of this section) that if X has the t distribution with m degrees of freedom ($m > 2$), then $\text{Var}(X) = m/(m-2)$.

Relation to Random Samples from a Normal Distribution

Example 8.4.2 Rain from Seeded Clouds. Return to Example 8.4.1. We have already seen that $Z = n^{1/2}(\overline{X}_n - \mu)/\sigma$ has the standard normal distribution. Furthermore, Theorem 8.3.1 says that $\overline{X}_n$ (and hence Z) is independent of $Y = n\widehat{\sigma^2}/\sigma^2$, which has the χ^2 distribution with $n - 1$ degrees of freedom. It follows that $Z/(Y/[n-1])^{1/2}$ has the t distribution with $n - 1$ degrees of freedom. We shall show how to use this fact after stating the general version of this result. ◀

Theorem 8.4.2 Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2. Let $\overline{X}_n$ denote the sample mean, and define

$$\sigma' = \left[\frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{n-1}\right]^{1/2}. \tag{8.4.3}$$

Then $n^{1/2}(\overline{X}_n - \mu)/\sigma'$ has the t distribution with $n - 1$ degrees of freedom.

Proof Define $S_n^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2$. Next, define $Z = n^{1/2}(\overline{X}_n - \mu)/\sigma$ and $Y = S_n^2/\sigma^2$. It follows from Theorem 8.3.1 that Y and Z are independent, Y has the χ^2 distribution with $n - 1$ degrees of freedom, and Z has the standard normal distribution. Finally, define U by

$$U = \frac{Z}{\left(\frac{Y}{n-1}\right)^{1/2}}.$$

It follows from the definition of the t distribution that U has the t distribution with $n - 1$ degrees of freedom. It is easily seen that U can be rewritten as

$$U = \frac{n^{1/2}(\overline{X}_n - \mu)}{\left(\frac{S_n^2}{n-1}\right)^{1/2}}. \tag{8.4.4}$$

The denominator of the expression on the right side of Eq. (8.4.4) is easily recognized as σ' defined in Eq. (8.4.3). ■

The first rigorous proof of Theorem 8.4.2 was given by R. A. Fisher in 1923.

One important aspect of Eq. (8.4.4) is that neither the value of U nor the distribution of U depends on the value of the variance σ^2. In Example 8.4.1, we tried replacing σ in the random variable $Z = n^{1/2}(\overline{X}_n - \mu)/\sigma$ by $\hat{\sigma}$. Instead, Theorem 8.4.2 suggests that we should replace σ by σ' defined in Eq. (8.4.3). If we replace σ by σ', we produce the random variable U in Eq. (8.4.4) that does not involve σ and also has a distribution that does not depend on σ.

The reader should notice that σ' differs from the M.L.E. $\hat{\sigma}$ of σ by a constant factor,

$$\sigma' = \left[\frac{S_n^2}{n-1}\right]^{1/2} = \left(\frac{n}{n-1}\right)^{1/2} \hat{\sigma}. \tag{8.4.5}$$

It can be seen from Eq. (8.4.5) that for large values of n the estimators σ' and $\hat{\sigma}$ will be very close to each other. The estimator σ' will be discussed further in Sec. 8.7.

If the sample size n is large, the probability that the estimator σ' will be close to σ is high. Hence, replacing σ by σ' in the random variable Z will not greatly change the standard normal distribution of Z. For this reason, it is plausible that the t distribution with $n-1$ degrees of freedom should be close to the standard normal distribution if n is large. We shall return to this point more formally later in this section.

Example 8.4.3

Rain from Seeded Clouds. Return to Example 8.4.2. Under the assumption that the observations $X_1, \ldots, X_n$ (log-rainfalls) are independent with common normal distribution, the distribution of $U = n^{1/2}(\overline{X}_n - \mu)/\sigma'$ is the t distribution with $n-1$ degrees of freedom. With $n = 26$, the table of the t distribution tells us that the 0.9 quantile of the t distribution with 25 degrees of freedom is 1.316, so $\Pr(U \le 1.316) = 0.9$. It follows that

$$\Pr\left(\overline{X}_n \le \mu + 0.2581\sigma'\right) = 0.9,$$

because $1.316/(26)^{1/2} = 0.2581$. That is, the probability is 0.9 that $\overline{X}_n$ will be no more than 0.2581 times σ' above μ. Of course, σ' is a random variable as well as $\overline{X}_n$, so this result is not as informative as we might have hoped. In Sections 8.5 and 8.6, we will show how to make use of the t distribution to make some standard inferences about the unknown mean μ. ◀

Relation to the Cauchy Distribution and to the Standard Normal Distribution

It can be seen from Eq. (8.4.2) (and Fig. 8.4) that the p.d.f. $g(x)$ is a symmetric, bell-shaped function with its maximum value at $x = 0$. Thus, its general shape is similar to that of the p.d.f. of a normal distribution with mean 0. However, as $x \to \infty$ or $x \to -\infty$, the tails of the p.d.f. $g(x)$ approach 0 much more slowly than do the tails of the p.d.f. of a normal distribution. In fact, it can be seen from Eq. (8.4.2) that the t distribution with one degree of freedom is the Cauchy distribution, which was defined in Example 4.1.8. The p.d.f. of the Cauchy distribution was sketched in Fig. 4.3. It was shown in Example 4.1.8 that the mean of the Cauchy distribution does not exist, because the integral that specifies the value of the mean is not absolutely convergent. It follows that, although the p.d.f. of the t distribution with one degree of freedom is symmetric with respect to the point $x = 0$, the mean of this distribution does not exist.

It can also be shown from Eq. (8.4.2) that, as $n \to \infty$, the p.d.f. $g(x)$ converges to the p.d.f. $\phi(x)$ of the standard normal distribution for every value of x $(-\infty < x < \infty)$. This follows from Theorem 5.3.3 and the following result:

$$\lim_{m \to \infty} \frac{\Gamma\left(m + \frac{1}{2}\right)}{\Gamma(m) m^{1/2}} = 1. \tag{8.4.6}$$

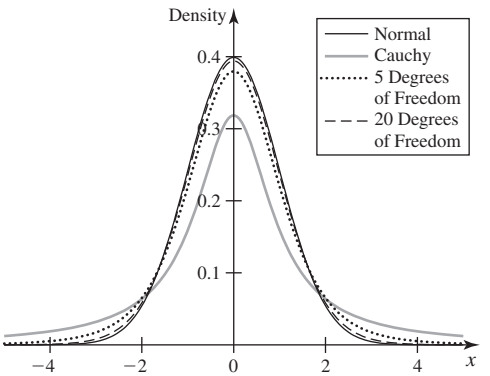

Figure 8.4 p.d.f.'s of standard normal and t distributions.

(See Exercise 7 for a way to prove the above result.) Hence, when n is large, the t distribution with n degrees of freedom can be approximated by the standard normal distribution. Figure 8.4 shows the p.d.f. of the standard normal distribution together with the p.d.f.'s of the t distributions with 1, 5, and 20 degrees of freedom so that the reader can see how the t distributions get closer to normal as the degrees of freedom increase.

A short table of p quantiles for the t distribution with m degrees of freedom for various values of p and m is given at the end of this book. The probabilities in the first line of the table, corresponding to $m = 1$, are those for the Cauchy distribution. The probabilities in the bottom line of the table corresponding to $m = \infty$ are those for the standard normal distribution. Most statistical packages include a function to compute the c.d.f. and the quantile function of an arbitrary t distribution.

Derivation of the p.d.f.

Suppose that the joint distribution of Y and Z is as specified in Definition 8.4.1. Then, because Y and Z are independent, their joint p.d.f. is equal to the product $f_1(y)\,f_2(z)$, where $f_1(y)$ is the p.d.f. of the χ^2 distribution with m degrees of freedom and $f_2(z)$ is the p.d.f. of the standard normal distribution. Let X be defined by Eq. (8.4.1) and, as a convenient device, let $W = Y$. We shall determine first the joint p.d.f. of X and W.

From the definitions of X and W,

$$Z = X\left(\frac{W}{m}\right)^{1/2} \quad \text{and} \quad Y = W. \tag{8.4.7}$$

The Jacobian of the transformation (8.4.7) from X and W to Y and Z is $(W/m)^{1/2}$. The joint p.d.f. $f(x, w)$ of X and W can be obtained from the joint p.d.f. $f_1(y)\,f_2(z)$ by replacing y and z by the expressions given in (8.4.7) and then multiplying the result by $(w/m)^{1/2}$. It is then found that the value of $f(x, w)$ is as follows, for $-\infty < x < \infty$ and $w > 0$:

$$\begin{aligned} f(x, w) &= f_1(w)\,f_2\left(x\left[\frac{w}{m}\right]^{1/2}\right)\left(\frac{w}{m}\right)^{1/2} \\ &= cw^{(m+1)/2-1}\exp\left[-\frac{1}{2}\left(1+\frac{x^2}{m}\right)w\right], \end{aligned} \tag{8.4.8}$$

where

$$c = \left[2^{(m+1)/2}(m\pi)^{1/2}\Gamma\left(\frac{m}{2}\right)\right]^{-1}.$$

The marginal p.d.f. $g(x)$ of X can be obtained from Eq. (8.4.8) by using the relation

$$g(x) = \int f(x, w)\, dw$$
$$= c\int_0^\infty w^{(m+1)/2-1}\exp[-wh(x)]\, dw,$$

where $h(x) = [1 + x^2/m]/2$. It follows from Eq. (5.7.10) that

$$g(x) = c\frac{\Gamma((m+1)/2)}{h(x)^{(m+1)/2}}.$$

Substituting the formula for c into this yields the function in (8.4.2).

Summary

Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\sigma' = \left(\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right)^{1/2}$. Then the distribution of $n^{1/2}(\overline{X}_n - \mu)/\sigma'$ is the t distribution with $n-1$ degrees of freedom.

Exercises

1. Suppose that X has the t distribution with m degrees of freedom ($m > 2$). Show that $\mathrm{Var}(X) = m/(m-2)$. *Hint:* To evaluate $E(X^2)$, restrict the integral to the positive half of the real line and change the variable from x to

$$y = \frac{\frac{x^2}{m}}{1 + \frac{x^2}{m}}.$$

Compare the integral with the p.d.f. of a beta distribution. Alternatively, use Exercise 21 in Sec. 5.7.

2. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown standard deviation σ, and let $\hat{\mu}$ and $\hat{\sigma}$ denote the M.L.E.'s of μ and σ. For the sample size $n = 17$, find a value of k such that

$$\Pr(\hat{\mu} > \mu + k\hat{\sigma}) = 0.95.$$

3. Suppose that the five random variables $X_1, \ldots, X_5$ are i.i.d. and that each has the standard normal distribution. Determine a constant c such that the random variable

$$\frac{c(X_1 + X_2)}{(X_3^2 + X_4^2 + X_5^2)^{1/2}}$$

will have a t distribution.

4. By using the table of the t distribution given in the back of this book, determine the value of the integral

$$\int_{-\infty}^{2.5} \frac{dx}{(12 + x^2)^2}.$$

5. Suppose that the random variables X_1 and X_2 are independent and that each has the normal distribution with mean 0 and variance σ^2. Determine the value of

$$\Pr\left[\frac{(X_1 + X_2)^2}{(X_1 - X_2)^2} < 4\right].$$

Hint:

$$(X_1 - X_2)^2 = 2\left[\left(X_1 - \frac{X_1 + X_2}{2}\right)^2 + \left(X_2 - \frac{X_1 + X_2}{2}\right)^2\right].$$

6. In Example 8.2.3, suppose that we will observe $n = 20$ cheese chunks with lactic acid concentrations $X_1, \ldots, X_{20}$. Find a number c so that $\Pr(\overline{X}_{20} \le \mu + c\sigma') = 0.95$.

7. Prove the limit formula Eq. (8.4.6). *Hint:* Use Theorem 5.7.4.

8. Let X have the standard normal distribution, and let Y have the t distribution with five degrees of freedom. Explain why $c = 1.63$ provides the largest value of the difference $\Pr(-c < X < c) - \Pr(-c < Y < c)$. *Hint:* Start by looking at Fig. 8.4.

8.5 Confidence Intervals

Confidence intervals provide a method of adding more information to an estimator $\hat{\theta}$ when we wish to estimate an unknown parameter θ. We can find an interval (A, B) that we think has high probability of containing θ. The length of such an interval gives us an idea of how closely we can estimate θ.

Confidence Intervals for the Mean of a Normal Distribution

Example 8.5.1

Rain from Seeded Clouds. In Example 8.3.2, the average of the $n = 26$ log-rainfalls from the seeded clouds is $\overline{X}_n$. This may be a sensible estimator of the μ, the mean log-rainfall from a seeded cloud, but it doesn't give any idea how much stock we should place in the estimator. The standard deviation of $\overline{X}_n$ is $\sigma/(26)^{1/2}$, and we could estimate σ by an estimator like σ' from Eq. (8.4.3). Is there a sensible way to combine these two estimators into an inference that tells us both what we should estimate for μ and how much confidence we should place in the estimator? ◀

Assume that $X_1, \ldots, X_n$, form a random sample from the normal distribution with mean μ and variance σ^2. Construct the estimators $\overline{X}_n$ of μ and σ' of σ . We shall now show how to make use of the random variable

$$U = \frac{n^{1/2}(\overline{X}_n - \mu)}{\sigma'} \tag{8.5.1}$$

from Eq. (8.4.4) to address the question at the end of Example 8.5.1. We know that U has the t distribution with $n - 1$ degrees of freedom. Hence, we can calculate the c.d.f. of U and/or quantiles of U using either statistical software or tables such as those in the back of this book. In particular, we can compute $\Pr(-c < U < c)$ for every $c > 0$. The inequalities $-c < U < c$ can be translated into inequalities involving μ by making use of the formula for U in Eq. (8.5.1). Simple algebra shows that $-c < U < c$ is equivalent to

$$\overline{X}_n - \frac{c\sigma'}{n^{1/2}} < \mu < \overline{X}_n + \frac{c\sigma'}{n^{1/2}}. \tag{8.5.2}$$

Whatever probability we can assign to the event $\{-c < U < c\}$ we can also assign to the event that Eq. (8.5.2) holds. For example, if $\Pr(-c < U < c) = \gamma$, then

$$\Pr\left(\overline{X}_n - \frac{c\sigma'}{n^{1/2}} < \mu < \overline{X}_n + \frac{c\sigma'}{n^{1/2}}\right) = \gamma. \tag{8.5.3}$$

One must be careful to understand the probability statement in Eq. (8.5.3) as being a statement about the joint distribution of the random variables $\overline{X}_n$ and σ' for fixed values of μ and σ. That is, it is a statement about the sampling distribution of $\overline{X}_n$ and

σ', and is conditional on μ and σ. In particular, it is *not* a statement about μ even if we treat μ as a random variable.

The most popular version of the calculation above is to choose γ and then figure out what c must be in order to make (8.5.3) true. That is, what value of c makes $\Pr(-c < U < c) = \gamma$? Let T_{n-1} denote the c.d.f. of the t distribution with $n-1$ degrees of freedom. Then

$$\gamma = \Pr(-c < U < c) = T_{n-1}(c) - T_{n-1}(-c).$$

Since the t distributions are symmetric around 0, $T_{n-1}(-c) = 1 - T_{n-1}(c)$, so $\gamma = 2T_{n-1}(c) - 1$ or, equivalently, $c = T_{n-1}^{-1}([1+\gamma]/2)$. That is, c must be the $(1+\gamma)/2$ quantile of the t distribution with $n-1$ degrees of freedom.

Example 8.5.2 Rain from Seeded Clouds. In Example 8.3.2, we have $n = 26$. If we want $\gamma = 0.95$ in Eq. (8.5.3), then we need c to be the $1.95/2 = 0.975$ quantile of the t distribution with 25 degrees of freedom. This can be found in the table of t distribution quantiles in the back of the book to be $c = 2.060$. We can plug this value into Eq. (8.5.3) and combine the constants $c/n^{1/2} = 2.060/26^{1/2} = 0.404$. Then Eq. (8.5.3) states that regardless of the unknown values of μ and σ, the probability is 0.95 that the two random variables $A = \overline{X}_n - 0.404\sigma'$ and $B = \overline{X}_n + 0.404\sigma'$ will lie on opposite sides of μ. ◀

The interval (A, B), whose endpoints were computed at the end of Example 8.5.2, is called a *confidence interval*.

Definition 8.5.1 Confidence Interval. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or parameter vector) θ. Let $g(\theta)$ be a real-valued function of θ. Let $A \le B$ be two statistics that have the property that for all values of θ,

$$\Pr(A < g(\theta) < B) \ge \gamma. \tag{8.5.4}$$

Then the random interval (A, B) is called a *coefficient γ confidence interval for $g(\theta)$* or a *100γ percent confidence interval for $g(\theta)$*. If the inequality "$\ge \gamma$" in Eq. (8.5.4) is an equality for all θ, the confidence interval is called *exact*. After the values of the random variables $X_1, \ldots, X_n$ in the random sample have been observed, the values of $A = a$ and $B = b$ are computed, and the interval (a, b) is called the observed value of the confidence interval.

In Example 8.5.2, $\theta = (\mu, \sigma^2)$, and the interval (A, B) found in that example is an exact 95% confidence interval for $g(\theta) = \mu$.

Based on the discussion preceding Definition 8.5.1, we have established the following.

Theorem 8.5.1 Confidence Interval for the Mean of a Normal Distribution. Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. For each $0 < \gamma < 1$, the interval (A, B) with the following endpoints is an exact coefficient γ confidence interval for μ:

$$A = \overline{X}_n - T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma'}{n^{1/2}},$$

$$B = \overline{X}_n + T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma'}{n^{1/2}}.$$ ■

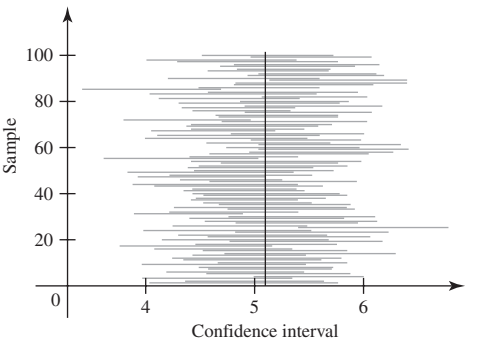

Figure 8.5 A sample of one hundred observed 95% confidence intervals based on samples of size 26 from the normal distribution with mean $\mu = 5.1$ and standard deviation $\sigma = 1.6$. In this figure, 94% of the intervals contain the value of μ.

Example 8.5.3

Rain from Seeded Clouds. In Example 8.5.2, the average of the 26 log-rainfalls from the seeded clouds is $\overline{X}_n = 5.134$. The observed value of σ' is 1.600. The observed values of A and B are, respectively, $a = 5.134 - 0.404 \times 1.600 = 4.488$ and $b = 5.134 + 0.404 \times 1.600 = 5.780$. The observed value of the 95% confidence interval is then (4.488, 5.780). For comparison, the mean unseeded level of 4 is a bit below the lower endpoint of this interval. ◀

Interpretaton of Confidence Intervals The interpretation of the confidence interval (A, B) defined in Definition 8.5.1 is straightforward, so long as one remembers that $\Pr(A < g(\theta) < B) = \gamma$ is a probability statement about the joint distribution of the two random variables A and B given a particular value of θ. Once we compute the observed values a and b, the observed interval (a, b) is not so easy to interpret. For example, some people would like to interpret the interval in Example 8.5.3 as meaning that we are 95% confident that μ is between 4.488 and 5.780. Later in this section, we shall show why such an interpretation is not safe in general. Before observing the data, we can be 95% confident that the random interval (A, B) will contain μ, but after observing the data, the safest interpretation is that (a, b) is simply the observed value of the random interval (A, B). One way to think of the random interval (A, B) is to imagine that the sample that we observed is one of many possible samples that we could have observed (or may yet observe in the future). Each such sample would allow us to compute an observed interval. Prior to observing the samples, we would expect 95% of the intervals to contain μ. Even if we observed many such intervals, we won't know which ones contain μ and which ones don't. Figure 8.5 contains a plot of 100 observed values of confidence intervals, each computed from a sample of size $n = 26$ from the normal distribution with mean $\mu = 5.1$ and standard deviation $\sigma = 1.6$. In this example, 94 of the 100 intervals contain the value of μ.

Example 8.5.4

Acid Concentration in Cheese. In Example 8.2.3, we discussed a random sample of 10 lactic acid measurements from cheese. Suppose that we desire to compute a 90% confidence interval for μ, the unknown mean lactic acid concentration. The number c that we need in Eq. (8.5.3) when $n = 10$ and $\gamma = 0.9$ is the $(1 + 0.9)/2 = 0.95$ quantile of the t distribution with nine degrees of freedom, $c = 1.833$. According to Eq. (8.5.3), the endpoints will be $\overline{X}_n$ plus and minus $1.833\sigma'/(10)^{1/2}$. Suppose that we observe the following 10 lactic acid concentrations as reported by Moore and McCabe (1999, p. D-1):

0.86, 1.53, 1.57, 1.81, 0.99, 1.09, 1.29, 1.78, 1.29, 1.58.

The average of these 10 values is $\bar{x}_n = 1.379$, and the value of $\sigma' = 0.3277$. The endpoints of the observed value of our 90% confidence interval are then $1.379 - 1.833 \times 0.3277/(10)^{1/2} = 1.189$ and $1.379 + 1.833 \times 0.3277/(10)^{1/2} = 1.569$. ◀

Note: Alternative Definitions of Confidence Interval. Many authors define confidence intervals precisely as we have done here. Some others define the confidence interval to be what we called the observed value of the confidence interval, namely, (a, b), and they need another name for the random interval (A, B). Throughout this book, we shall stay with the definition we have given, but the reader who studies statistics further might encounter the other definition at a later date. Also, some authors define confidence intervals to be closed intervals rather than open intervals.

One-Sided Confidence Intervals

Example 8.5.5 Rain from Seeded Clouds. Suppose that we are interested only in obtaining a lower bound on μ, the mean log-rainfall of seeded clouds. In the spirit of confidence intervals, we could then seek a random variable A such that $\Pr(A < \mu) = \gamma$. If we let $B = \infty$ in Definition 8.5.1, we see that (A, ∞) is then a coefficient γ confidence interval for μ. ◀

For a given confidence coefficient γ, it is possible to construct many different confidence intervals for μ. For example, let $\gamma_2 > \gamma_1$ be two numbers such that $\gamma_2 - \gamma_1 = \gamma$, and let U be as in Eq. (8.5.1). Then

$$\Pr\left(T_{n-1}^{-1}(\gamma_1) < U < T_{n-1}^{-1}(\gamma_2)\right) = \gamma,$$

and the following statistics are the endpoints of a coefficient γ confidence interval for μ:

$$A = \overline{X}_n + T_{n-1}^{-1}(\gamma_1)\frac{\sigma'}{n^{1/2}} \quad \text{and} \quad B = \overline{X}_n + T_{n-1}^{-1}(\gamma_2)\frac{\sigma'}{n^{1/2}}.$$

Among all such coefficient γ confidence intervals, the symmetric interval with $\gamma_1 = 1 - \gamma_2$ is the shortest one.

Nevertheless, there are cases, such as Example 8.5.5, in which an asymmetric confidence interval is useful. In general, it is a simple matter to extend Definition 8.5.1 to allow either $A = -\infty$ or $B = \infty$ so that the confidence interval either has the form $(-\infty, B)$ or (A, ∞).

Definition 8.5.2 One-Sided Confidence Intervals/Limits. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or parameter vector) θ. Let $g(\theta)$ be a real-valued function of θ. Let A be a statistic that has the property that for all values of θ,

$$\Pr(A < g(\theta)) \geq \gamma. \tag{8.5.5}$$

Then the random interval (A, ∞) is called a *one-sided coefficient γ confidence interval for $g(\theta)$* or a *one-sided 100γ percent confidence interval for $g(\theta)$*. Also, A is called a *coefficient γ lower confidence limit for $g(\theta)$* or a *100γ percent lower confidence limit for $g(\theta)$*. Similarly, if B is a statistic such that

$$\Pr(g(\theta) < B) \geq \gamma, \tag{8.5.6}$$

then $(-\infty, B)$ is a *one-sided coefficient γ confidence interval for $g(\theta)$* or a *one-sided 100γ percent confidence interval for $g(\theta)$* and B is a *coefficient γ upper confidence limit*

for $g(\theta)$ or a *100γ percent upper confidence limit for $g(\theta)$*. If the inequality "$\geq \gamma$" in either Eq. (8.5.5) or Eq. (8.5.6) is equality for all θ, the corresponding confidence interval and confidence limit are called *exact*.

The following result follows in much the same way as Theorem 8.5.1.

Theorem 8.5.2 One-Sided Confidence Intervals for the Mean of a Normal Distribution. Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. For each $0 < \gamma < 1$, the following statistics are, respectively, exact lower and upper coefficient γ confidence limits for μ:

$$A = \overline{X}_n - T_{n-1}^{-1}(\gamma)\frac{\sigma'}{n^{1/2}},$$

$$B = \overline{X}_n + T_{n-1}^{-1}(\gamma)\frac{\sigma'}{n^{1/2}}.$$

■

Example 8.5.6 Rain from Seeded Clouds. In Example 8.5.5, suppose that we want a 90% lower confidence limit for μ. We find $T_{25}^{-1}(0.9) = 1.316$. Using the observed data from Example 8.5.3, we compute the observed lower confidence limit as

$$a = 5.134 - 1.316\frac{1.600}{26^{1/2}} = 4.727.$$

◀

Confidence Intervals for Other Parameters

Example 8.5.7 Lifetimes of Electronic Components. Recall the company in Example 8.1.3 that is estimating the failure rate θ of electronic components based on a sample of $n = 3$ observed lifetimes X_1, X_2, X_3. The statistic $T = \sum_{i=1}^{3} X_i$ was used in Examples 8.1.4 and 8.1.5 to make some inferences. We can use the distribution of T to construct confidence intervals for θ. Recall from Example 8.1.5 that θT has the gamma distribution with parameters 3 and 1 for all θ. Let G stand for the c.d.f. of this gamma distribution. Then $\Pr(\theta T < G^{-1}(\gamma)) = \gamma$ for all θ. It follows that $\Pr(\theta < G^{-1}(\gamma)/T) = \gamma$ for all θ, and $G^{-1}(\gamma)/T$ is an exact coefficient γ upper confidence limit for θ. For example, if the company would like to have a random variable B so that they can be 98% confident that the failure rate θ is bounded above by B, they can find $G^{-1}(0.98) = 7.516$. Then $B = 7.516/T$ is the desired upper confidence limit. ◀

In Example 8.5.7, the random variable θT has the property that its distribution is the same for all θ. The random variable U in Eq. (8.5.1) has the property that its distribution is the same for all μ and σ. Such random variables greatly facilitate the construction of confidence intervals.

Definition 8.5.3 Pivotal. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or vector of parameters) θ. Let $V(\boldsymbol{X}, \theta)$ be a random variable whose distribution is the same for all θ. Then V is called a *pivotal quantity* (or simply a *pivotal*).

In order to be able to use a pivotal to construct a confidence interval for $g(\theta)$, one needs to be able to "invert" the pivotal. That is, one needs a function $r(v, \boldsymbol{x})$ such that

$$r\left(V(\boldsymbol{X}, \theta), \boldsymbol{X}\right) = g(\theta). \tag{8.5.7}$$

If such a function exists, then one can use it to construct confidence intervals.

Theorem 8.5.3 Confidence Interval from a Pivotal. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or vector of parameters) θ. Suppose that a pivotal V exists. Let G be the c.d.f. of V, and assume that G is continuous. Assume that a function r exists as in Eq. (8.5.7), and assume that $r(v, \boldsymbol{x})$ is strictly increasing in v for each $\boldsymbol{x}$. Let $0 < \gamma < 1$ and let $\gamma_2 > \gamma_1$ be such that $\gamma_2 - \gamma_1 = \gamma$. Then the following statistics are the endpoints of an exact coefficient γ confidence interval for $g(\theta)$:

$$A = r\left(G^{-1}(\gamma_1), \boldsymbol{X}\right),$$

$$B = r\left(G^{-1}(\gamma_2), \boldsymbol{X}\right).$$

If $r(v, \boldsymbol{x})$ is strictly decreasing in v for each $\boldsymbol{x}$, then switch the definitions of A and B.

Proof If $r(v, \boldsymbol{x})$ is strictly increasing in v for each $\boldsymbol{x}$, we have

$$V(\boldsymbol{X}, \theta) < c \text{ if and only if } g(\theta) < r(c, \boldsymbol{X}). \tag{8.5.8}$$

Let $c = G^{-1}(\gamma_i)$ in Eq. (8.5.8) for each of $i = 1, 2$ to obtain

$$\begin{aligned} &\Pr(g(\theta) < A) = \gamma_1, \\ &\Pr(g(\theta) < B) = \gamma_2. \end{aligned} \tag{8.5.9}$$

Because V has a continuous distribution and r is strictly increasing,

$$\Pr(A = g(\theta)) = \Pr(V(\boldsymbol{X}, \theta) = G^{-1}(\gamma_1)) = 0.$$

Similarly, $\Pr(B = g(\theta)) = 0$. The two equations in (8.5.9) combine to give $\Pr(A < g(\theta) < B) = \gamma$. The proof when r is strictly decreasing is similar and is left to the reader. ■

Example 8.5.8 Pivotal for Estimating the Variance of a Normal Distribution. Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean μ and variance σ^2 In Theorem 8.3.1, we found that the random variable $V(\boldsymbol{X}, \theta) = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2/\sigma^2$ has the χ^2 distribution with $n - 1$ degrees of freedom for all $\theta = (\mu, \sigma^2)$. This makes V a pivotal. The reader can use this pivotal in Exercise 5 in this section to find a confidence interval of $g(\theta) = \sigma^2$. ◀

Sometimes pivotals do not exist. This is common when the data have a discrete distribution.

Example 8.5.9 A Clinical Trial. Consider the imipramine treatment group in the clinical trial in Example 2.1.4. Let θ stand for the proportion of successes among a very large population of imipramine patients. Suppose that the clinicians desire a random variable A such that, for all θ, $\Pr(A < \theta) \geq 0.9$. That is, they want to be 90% confident that the success proportion is at least A. The observable data consist of the number X of successes in a random sample of $n = 40$ patients. No pivotal exists in this example, and confidence intervals are more difficult to construct. In Example 9.1.16, we shall see a method that applies to this case. ◀

Even with discrete data, if the sample size is large enough to apply the central limit theorem, one can find approximate confidence intervals.

Example 8.5.10 Approximate Confidence Interval for Poisson Mean. Suppose that $X_1, \ldots, X_n$ have the Poisson distribution with unknown mean θ. Suppose that n is large enough so that

$\overline{X}_n$ has approximately a normal distribution. In Example 6.3.8 on page 365, we found that

$$\Pr\left(|2\overline{X}_n^{1/2} - 2\theta^{1/2}| < c\right) \approx 2\Phi(cn^{1/2}) - 1. \tag{8.5.10}$$

After we observe $\overline{X}_n = x$, Eq. (8.5.10) says that

$$\left(-c + 2x^{1/2},\, c + 2x^{1/2}\right) \tag{8.5.11}$$

is the observed value of an approximate confidence interval for $2\theta^{1/2}$ with coefficient $2\Phi(cn^{1/2}) - 1$. For example, if $c = 0.196$ and $n = 100$, then $2\Phi(cn^{1/2}) - 1 = 0.95$. The inverse of $g(\theta) = 2\theta^{1/2}$ is $g^{-1}(y) = y^2/4$, which is an increasing function of y for $y \geq 0$. If both endpoints of (8.5.11) are nonnegative, then we know that $2\theta^{1/2}$ is in the interval (8.5.11) if and only if θ is in the interval

$$\left(\frac{1}{4}[-c + 2x^{1/2}]^2,\, \frac{1}{4}[c + 2x^{1/2}]^2\right). \tag{8.5.12}$$

If $-c + 2x^{1/2} < 0$, the left endpoints of (8.5.11) and (8.5.12) should be replaced by 0. With this modification, (8.5.12) is the observed value of an approximate coefficient $2\Phi(cn^{1/2}) - 1$ confidence interval for θ. ◀

Shortcoming of Confidence Intervals

Interpretation of Confidence Intervals Let (A, B) be a coefficient γ confidence interval for a parameter θ, and let (a, b) be the observed value of the interval. It is important to understand that it is *not* correct to say that θ lies in the interval (a, b) with *probability* γ. We shall explain this point further here. *Before* the values of the statistics $A(X_1, \ldots, X_n)$ and $B(X_1, \ldots, X_n)$ are observed, these statistics are random variables. It follows, therefore, from Definition 8.5.1 that θ will lie in the random interval having endpoints $A(X_1, \ldots, X_n)$ and $B(X_1, \ldots, X_n)$ with probability γ. *After* the specific values $A(X_1, \ldots, X_n) = a$ and $B(X_1, \ldots, X_n) = b$ have been observed, it is not possible to assign a probability to the event that θ lies in the specific interval (a, b) without regarding θ as a random variable, which itself has a probability distribution. In order to calculate the probability that θ lies in the interval (a, b), it is necessary first to assign a prior distribution to θ and then use the resulting posterior distribution. Instead of assigning a prior distribution to the parameter θ, many statisticians prefer to state that there is *confidence* γ, rather than probability γ, that θ lies in the interval (a, b). Because of this distinction between confidence and probability, the meaning and the relevance of confidence intervals in statistical practice is a somewhat controversial topic.

Information Can Be Ignored In accordance with the preceding explanation, the interpretation of a confidence coefficient γ for a confidence interval is as follows: Before a sample is taken, there is probability γ that the interval that will be constructed from the sample will include the unknown value of θ. After the sample values are observed, however, there might be additional information about whether or not the interval formed from these particular values actually does include θ. How to adjust the confidence coefficient γ in the light of this information is another controversial topic.

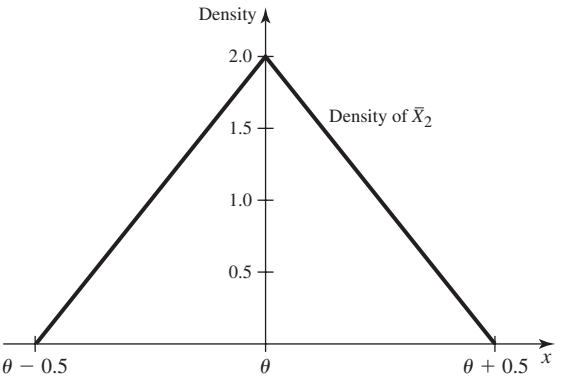

Figure 8.6 p.d.f. of $\overline{X}_2$ in Example 8.5.11.

Example 8.5.11

Uniforms on an Interval of Length One. Suppose that two observations X_1 and X_2 are taken at random from the uniform distribution on the interval $\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$, where the value of θ is unknown ($-\infty < \theta < \infty$). If we let $Y_1 = \min\{X_1, X_2\}$ and $Y_2 = \max\{X_1, X_2\}$, then

$$\begin{aligned}\Pr(Y_1 < \theta < Y_2) &= \Pr(X_1 < \theta < X_2) + \Pr(X_2 < \theta < X_1)\\ &= \Pr(X_1 < \theta)\Pr(X_2 > \theta) + \Pr(X_2 < \theta)\Pr(X_1 > \theta)\\ &= (1/2)(1/2) + (1/2)(1/2) = 1/2. \end{aligned} \tag{8.5.13}$$

It follows from Eq. (8.5.13) that (Y_1, Y_2) is a confidence interval for θ with confidence coefficient 1/2. However, the analysis can be carried further.

Since both observations X_1 and X_2 must be at least $\theta - (1/2)$, and both must be at most $\theta + (1/2)$, we know with certainty that $Y_1 \geq \theta - (1/2)$ and $Y_2 \leq \theta + (1/2)$. In other words, we know with certainty that

$$Y_2 - (1/2) \leq \theta \leq Y_1 + (1/2). \tag{8.5.14}$$

Suppose now that $Y_1 = y_1$ and $Y_2 = y_2$ are observed such that $(y_2 - y_1) > 1/2$. Then $y_1 < y_2 - (1/2)$, and it follows from Eq. (8.5.14) that $y_1 < \theta$. Moreover, because $y_1 + (1/2) < y_2$, it also follows from Eq. (8.5.14) that $\theta < y_2$. Thus, if $(y_2 - y_1) > 1/2$, then $y_1 < \theta < y_2$. In other words, if $(y_2 - y_1) > 1/2$, then we know with certainty that the observed value (y_1, y_2) of the confidence interval includes the unknown value of θ, even though the confidence coefficient of this interval is only 1/2.

Indeed, even when $(y_2 - y_1) \leq 1/2$, the closer the value of $(y_2 - y_1)$ is to 1/2, the more certain we feel that the interval (y_1, y_2) includes θ. Also, the closer the value of $(y_2 - y_1)$ is to 0, the more certain we feel that the interval (y_1, y_2) does not include θ. However, the confidence coefficient necessarily remains 1/2 and does not depend on the observed values y_1 and y_2.

This example also helps to illustrate the statement of caution made at the end of Sec. 8.1. In this problem, it might seem natural to estimate θ by $\overline{X}_2 = 0.5(X_1 + X_2)$. Using the methods of Sec. 3.9, we can find the p.d.f. of $\overline{X}_2$:

$$g(x) = \begin{cases} 4x - 4\theta + 2 & \text{if } \theta - \frac{1}{2} < x \leq \theta,\\ 4\theta - 4x + 2 & \text{if } \theta < x < \theta + \frac{1}{2},\\ 0 & \text{otherwise.}\end{cases}$$

Figure 8.6 shows the p.d.f. g, which is triangular. This makes it fairly simple to compute the probability that $\overline{X}_2$ is close to θ:

$$\Pr(|\overline{X}_2 - \theta| < c) = 4c(1-c),$$

for $0 < c < 1/2$, and the probability is 1 for $c \geq 1/2$. For example, if $c = 0.3$, $\Pr(|\overline{X}_2 - \theta| < 0.3) = 0.84$. However, the random variable $Z = Y_2 - Y_1$ contains useful information that is not accounted for in this calculation. Indeed, the conditional distribution of $\overline{X}_2$ given $Z = z$ is uniform on the interval $\left[\theta - \frac{1}{2}(1-z), \theta + \frac{1}{2}(1-z)\right]$. We see that the larger the observed value of z, the shorter the range of possible values of $\overline{X}_2$. In particular, the conditional probability that $\overline{X}_2$ is close to θ given $Z = z$ is

$$\Pr(|\overline{X}_2 - \theta| < c | Z = z) = \begin{cases} \frac{2c}{1-z} & \text{if } c \leq (1-z)/2, \\ 1 & \text{if } c > (1-z)/2. \end{cases} \tag{8.5.15}$$

For example, if $z = 0.1$, then $\Pr(|\overline{X}_2 - \theta| < 0.3 | Z = 0.1) = 0.6667$, which is quite a bit smaller than the marginal probability of 0.84. This illustrates why it is not always safe to assume that our estimate is close to the parameter just because the sampling distribution of the estimator had high probability of being close. There may be other information available that suggests to us that the estimate is not as close as the sampling distribution suggests, or that it is closer than the sampling distribution suggests. (The reader should calculate $\Pr(|\overline{X}_2 - \theta| < 0.3 | Z = 0.9)$ for the other extreme.) ◀

In the next section, we shall discuss Bayesian methods for analyzing a random sample from a normal distribution for which both the mean μ and the variance σ^2 are unknown. We shall assign a joint prior distribution to μ and σ^2, and shall then calculate the posterior probability that μ belongs to any given interval (a, b). It can be shown [see, e.g., DeGroot (1970)] that if the joint prior p.d.f. of μ and σ^2 is fairly smooth and does not assign high probability to any particular small set of values of μ and σ^2, and if the sample size n is large, then the confidence coefficient assigned to a particular confidence interval (A, B) for the mean μ will be approximately equal to the posterior probability that μ lies in the observed interval (a, b). An example of this approximate equality is included in the next section. Therefore, under these conditions, the differences between the results obtained by the practical application of methods based on confidence intervals and methods based on prior probabilities will be small. Nevertheless interpretations of these methods will differ. As an aside, a Bayesian analysis of Example 8.5.11 will necessarily take into account the extra information contained in the random variable Z. See Exercise 10 for an example.

Summary

Let $X_1, \ldots, X_n$ be a random sample of independent random variables from the normal distribution with mean μ and variance σ^2. Let the observed values be $x_1, \ldots, x_n$. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\sigma'^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. The interval $(\overline{X}_n - c\sigma'/n^{1/2}, \overline{X} + c\sigma'/n^{1/2})$ is a coefficient γ confidence interval for μ, where c is the $(1+\gamma)/2$ quantile of the t distribution with $n-1$ degrees of freedom.

Exercises

1. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance σ^2. Let Φ stand for the c.d.f. of the standard normal distribution, and let Φ^{-1} be its inverse. Show that the following interval is a coefficient γ confidence interval for μ if $\overline{X}_n$ is the observed average of the data values:

$$\left(\overline{X}_n - \Phi^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma}{n^{1/2}}, \overline{X}_n + \Phi^{-1}\left(\frac{1+\gamma}{2}\right)\frac{\sigma}{n^{1/2}}\right).$$

2. Suppose that a random sample of eight observations is taken from the normal distribution with unknown mean μ and unknown variance σ^2, and that the observed values are 3.1, 3.5, 2.6, 3.4, 3.8, 3.0, 2.9, and 2.2. Find the shortest confidence interval for μ with each of the following three confidence coefficients: **(a)** 0.90, **(b)** 0.95, and **(c)** 0.99.

3. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2, and let the random variable L denote the length of the shortest confidence interval for μ that can be constructed from the observed values in the sample. Find the value of $E(L^2)$ for the following values of the sample size n and the confidence coefficient γ:

a. $n = 5, \gamma = 0.95$ **d.** $n = 8, \gamma = 0.90$

b. $n = 10, \gamma = 0.95$ **e.** $n = 8, \gamma = 0.95$

c. $n = 30, \gamma = 0.95$ **f.** $n = 8, \gamma = 0.99$

4. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance σ^2. How large a random sample must be taken in order that there will be a confidence interval for μ with confidence coefficient 0.95 and length less than 0.01σ?

5. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Describe a method for constructing a confidence interval for σ^2 with a specified confidence coefficient γ $(0 < \gamma < 1)$. *Hint:* Determine constants c_1 and c_2 such that

$$\Pr\left[c_1 < \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{\sigma^2} < c_2\right] = \gamma.$$

6. Suppose that $X_1, \ldots, X_n$ form a random sample from the exponential distribution with unknown mean μ. Describe a method for constructing a confidence interval for μ with a specified confidence coefficient γ $(0 < \gamma < 1)$. *Hint:* Determine constants c_1 and c_2 such that $\Pr[c_1 < (1/\mu)\sum_{i=1}^n X_i < c_2] = \gamma$.

7. In the June 1986 issue of *Consumer Reports*, some data on the calorie content of beef hot dogs is given. Here are the numbers of calories in 20 different hot dog brands:

186, 181, 176, 149, 184, 190, 158, 139, 175, 148,
152, 111, 141, 153, 190, 157, 131, 149, 135, 132.

Assume that these numbers are the observed values from a random sample of twenty independent normal random variables with mean μ and variance σ^2, both unknown. Find a 90% confidence interval for the mean number of calories μ.

8. At the end of Example 8.5.11, compute the probability that $|\overline{X}_2 - \theta| < 0.3$ given $Z = 0.9$. Why is it so large?

9. In the situation of Example 8.5.11, suppose that we observe $X_1 = 4.7$ and $X_2 = 5.3$.

a. Find the 50% confidence interval described in Example 8.5.11.

b. Find the interval of possible θ values that are consistent with the observed data.

c. Is the 50% confidence interval larger or smaller than the set of possible θ values?

d. Calculate the value of the random variable $Z = Y_2 - Y_1$ as described in Example 8.5.11.

e. Use Eq. (8.5.15) to compute the conditional probability that $|\overline{X}_2 - \theta| < 0.1$ given Z equal to the value computed in part (d).

10. In the situation of Exercise 9, suppose that a prior distribution is used for θ with p.d.f. $\xi(\theta) = 0.1\exp(-0.1\theta)$ for $\theta > 0$. (This is the exponential distribution with parameter 0.1.)

a. Prove that the posterior p.d.f. of θ given the data observed in Exercise 9 is

$$\xi(\theta|\boldsymbol{x}) = \begin{cases} 4.122\exp(-0.1\theta) & \text{if } 4.8 < \theta < 5.2, \\ 0 & \text{otherwise.} \end{cases}$$

b. Calculate the posterior probability that $|\theta - \overline{x}_2| < 0.1$, where $\overline{x}_2$ is the observed average of the data values.

c. Calculate the posterior probability that θ is in the confidence interval found in part (a) of Exercise 9.

d. Can you explain why the answer to part (b) is so close to the answer to part (e) of Exercise 9? *Hint:* Compare the posterior p.d.f. in part (a) to the function in Eq. (8.5.15).

11. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with parameter p. Let $\overline{X}_n$ be the sample average. Use the variance stabilizing transformation found in Exercise 5 of Section 6.5 to construct an approximate coefficient γ confidence interval for p.

12. Complete the proof of Theorem 8.5.3 by dealing with the case in which $r(v, \boldsymbol{x})$ is strictly decreasing in v for each $\boldsymbol{x}$.

⋆ 8.6 Bayesian Analysis of Samples from a Normal Distribution

When we are interested in constructing a prior distribution for the parameters μ and σ^2 of a normal distribution, it is more convenient to work with $\tau = 1/\sigma^2$, called the precision. *A conjugate family of prior distributions is introduced for μ and τ, and the posterior distribution is derived. Interval estimates of μ can be constructed from the posterior and these are similar to confidence intervals in form, but they are interpreted differently.*

The Precision of a Normal Distribution

Example 8.6.1 Rain from Seeded Clouds. In Example 8.3.1, we mentioned that it was of interest whether the mean log-rainfall μ from seeded clouds exceeded the mean log-rainfall of unseeded clouds, namely, 4. Although we were able to find an estimator of μ and we were able to construct a confidence interval for μ, we have not yet directly addressed the question of whether or not $\mu > 4$ or how likely it is that $\mu > 4$. If we construct a joint prior distribution for both μ and σ^2, we can then find the posterior distribution of μ and finally provide direct answers to these questions. ◀

Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. In this section, we shall consider the assignment of a joint prior distribution to the parameters μ and σ^2 and study the posterior distribution that is then derived from the observed values in the sample. Manipulating prior and posterior distributions for the parameters of a normal distribution turns out to be simpler if we reparameterize from μ and σ^2 to μ and $\tau = 1/\sigma^2$.

Definition 8.6.1 Precision of a Normal Distribution. The *precision* τ of a normal distribution is defined as the reciprocal of the variance; that is, $\tau = 1/\sigma^2$.

If a random variable has the normal distribution with mean μ and precision τ, then its p.d.f. $f(x|\mu, \tau)$ is specified as follows, for $-\infty < x < \infty$:

$$f(x|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left[-\frac{1}{2}\tau(x-\mu)^2\right].$$

Similarly, if $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and precision τ, then their joint p.d.f. $f_n(\boldsymbol{x}|\mu, \tau)$ is as follows, for $-\infty < x_i < \infty$ $(i = 1, \ldots, n)$:

$$f_n(x|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left[-\frac{1}{2}\tau \sum_{i=1}^{n}(x_i-\mu)^2\right].$$

A Conjugate Family of Prior Distributions

We shall now describe a conjugate family of joint prior distributions for μ and τ. We shall specify the joint distribution of μ and τ by specifying both the conditional distribution of μ given τ and the marginal distribution of τ. In particular, we shall assume that the conditional distribution of μ for each given value of τ is a normal distribution for which the precision is proportional to the given value of τ, and also

that the marginal distribution of τ is a gamma distribution. The family of all joint distributions of this type is a conjugate family of joint prior distributions. If the joint prior distribution of μ and τ belongs to this family, then for every possible set of observed values in the random sample, the joint posterior distribution of μ and τ will also belong to the family. This result is established in Theorem 8.6.1. We shall use the following notation in the theorem and the remainder of this section:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_n^2 = \sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

Theorem 8.6.1

Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown precision τ ($-\infty < \mu < \infty$ and $\tau > 0$). Suppose also that the joint prior distribution of μ and τ is as follows: The conditional distribution of μ given τ is the normal distribution with mean μ_0 and precision $\lambda_0\tau$ ($-\infty < \mu_0 < \infty$ and $\lambda_0 > 0$), and the marginal distribution of τ is the gamma distribution with parameters α_0 and β_0 ($\alpha_0 > 0$ and $\beta_0 > 0$). Then the joint posterior distribution of μ and τ, given that $X_i = x_i$ for $i = 1, \ldots, n$, is as follows: The conditional distribution of μ given τ is the normal distribution with mean μ_1 and precision $\lambda_1\tau$, where

$$\mu_1 = \frac{\lambda_0\mu_0 + n\overline{x}_n}{\lambda_0 + n} \quad \text{and} \quad \lambda_1 = \lambda_0 + n, \tag{8.6.1}$$

and the marginal distribution of τ is the gamma distribution with parameters α_1 and β_1, where

$$\alpha_1 = \alpha_0 + \frac{n}{2} \quad \text{and} \quad \beta_1 = \beta_0 + \frac{1}{2}s_n^2 + \frac{n\lambda_0(\overline{x}_n - \mu_0)^2}{2(\lambda_0 + n)}. \tag{8.6.2}$$

Proof The joint prior p.d.f. $\xi(\mu, \tau)$ of μ and τ can be found by multiplying the conditional p.d.f. $\xi_1(\mu|\tau)$ of μ given τ by the marginal p.d.f. $\xi_2(\tau)$ of τ. By the conditions of the theorem, we have, for $-\infty < \mu < \infty$ and $\tau > 0$,

$$\xi_1(\mu|\tau) \propto \tau^{1/2}\exp\left[-\frac{1}{2}\lambda_0\tau(\mu - \mu_0)^2\right]$$

and

$$\xi_2(\tau) \propto \tau^{\alpha_0 - 1}e^{-\beta_0\tau}.$$

A constant factor involving neither μ nor τ has been dropped from the right side of each of these relations.

The joint posterior p.d.f. $\xi(\mu, \tau|\boldsymbol{x})$ for μ and τ satisfies the relation

$$\begin{aligned}\xi(\mu, \tau|\boldsymbol{x}) &\propto f_n(\boldsymbol{x}|\mu, \tau)\xi_1(\mu|\tau)\xi_2(\tau) \qquad (8.6.3)\\ &\propto \tau^{\alpha_0 + (n+1)/2 - 1}\exp\left[-\frac{\tau}{2}\left(\lambda_0[\mu - \mu_0]^2 + \sum_{i=1}^{n}(x_i - \mu)^2\right) - \beta_0\tau\right].\end{aligned}$$

By adding and subtracting $\overline{x}_n$ inside the $(x_i - \mu)^2$ terms, we can prove that

$$\sum_{i=1}^{n}(x_i - \mu)^2 = s_n^2 + n(\overline{x}_n - \mu)^2. \tag{8.6.4}$$

Next, combine the last term in Eq. (8.6.4) with the term $\lambda_0(\mu - \mu_0)^2$ in (8.6.3) by completing the square (see Exercise 24 in Sec. 5.6) to get

$$n(\overline{x}_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2 = (\lambda_0 + n)(\mu - \mu_1)^2 + \frac{n\lambda_0(\overline{x}_n - \mu_0)^2}{\lambda_0 + n}, \tag{8.6.5}$$

where μ_1 is defined in Eq. (8.6.1). Combining (8.6.4) with (8.6.5) yields

$$\sum_{i=1}^{n}(x_i-\mu)^2+\lambda_0(\mu-\mu_0)^2=(\lambda_0+n)(\mu-\mu_1)^2+s_n^2 + \frac{n\lambda_0(\overline{x}_n-\mu_0)^2}{\lambda_0+n}. \tag{8.6.6}$$

Using (8.6.2) and $\lambda_1=\lambda_0+n$ together with (8.6.6) allows us to write Eq. (8.6.3) in the form

$$\xi(\mu,\tau|\boldsymbol{x})\propto\left\{\tau^{1/2}\exp\left[-\frac{1}{2}\lambda_1\tau(\mu-\mu_1)^2\right]\right\}(\tau^{\alpha_1-1}e^{-\beta_1\tau}), \tag{8.6.7}$$

where λ_1, α_1, and β_1 are defined by Eqs. (8.6.1) and (8.6.2).

When the expression inside the braces on the right side of Eq. (8.6.7) is regarded as a function of μ for a fixed value of τ, this expression can be recognized as being (except for a factor that depends on neither μ nor τ) the p.d.f. of the normal distribution with mean μ_1 and precision $\lambda_1\tau$. Since the variable μ does not appear elsewhere on the right side of Eq. (8.6.7), it follows that this p.d.f. must be the conditional posterior p.d.f. of μ given τ. It now follows in turn that the expression outside the braces on the right side of Eq. (8.6.7) must be proportional to the marginal posterior p.d.f. of τ. This expression can be recognized as being (except for a constant factor) the p.d.f. of the gamma distribution with parameters α_1 and β_1. Hence, the joint posterior distribution of μ and τ is as specified in the theorem. ■

We shall give a name to the family of joint distributions described in Theorem 8.6.1.

Definition 8.6.2 Normal-Gamma Family of Distributions. Let μ and τ be random variables. Suppose that the conditional distribution of μ given τ is the normal distribution with mean μ_0 and precision $\lambda_0\tau$. Suppose also that the marginal distribution of τ is the gamma distribution with parameters α_0 and β_0. Then we say that the joint distribution of μ and τ is the *normal-gamma distribution with hyperparameters* μ_0, λ_0, α_0, and β_0.

The prior distribution in Theorem 8.6.1 is the normal-gamma distribution with hyperparameters μ_0, λ_0, α_0, and β_0. The posterior distribution derived in that theorem is the normal-gamma distribution with hyperparameters μ_1, λ_1, α_1, and β_1. As in Sec. 7.3, we shall refer to the hyperparameters of the prior distribution as *prior hyperparameters*, and we shall refer to the hyperparameters of the posterior distribution as *posterior hyperparameters*.

By choosing appropriate values of the prior hyperparameters, it is usually possible in a particular problem to find a normal-gamma distribution that approximates an experimenter's actual prior distribution of μ and τ sufficiently well. It should be emphasized, however, that if the joint distribution of μ and τ is a normal-gamma distribution, then μ and τ are not independent. Thus, it is not possible to use a normal-gamma distribution as a joint prior distribution of μ and τ in a problem in which the experimenter wishes μ and τ to be independent in the prior. Although this characteristic of the family of normal-gamma distributions is a deficiency, it is not an important deficiency, because of the following fact: Even if a joint prior distribution under which μ and τ are independent is chosen from outside the conjugate family, it will be found that after just a single value of X has been observed, μ and τ will have a posterior distribution under which they are dependent. In other words, it is not possible for μ

and τ to remain independent in the light of even one observation from the underlying normal distribution.

Example 8.6.2 Acid Concentration in Cheese. Consider again the example of lactic acid concentration in cheese as discussed in Example 8.5.4. Suppose that the concentrations are independent normal random variables with mean μ and precision τ. Suppose that the prior opinion of the experimenters could be expressed as a normal-gamma distribution with hyperparameters $\mu_0 = 1$, $\lambda_0 = 1$, $\alpha_0 = 0.5$, and $\beta_0 = 0.5$. We can use the data on page 487 to find the posterior distribution of μ and τ. In this case, $n = 10$, $\overline{x}_n = 1.379$, and $s_n^2 = 0.9663$. Applying the formulas in Theorem 8.6.1, we get

$$\mu_1 = \frac{1 \times 1 + 10 \times 1.379}{1 + 10} = 1.345, \qquad \lambda_1 = 1 + 10 = 11, \qquad \alpha_1 = 0.5 + \frac{10}{2} = 5.5,$$

$$\beta_1 = 0.5 + \frac{1}{2}0.9663 + \frac{10 \times 1 \times (1.379 - 1)^2}{2(1 + 10)} = 1.0484.$$

So, the posterior distribution of μ and τ is the normal-gamma distribution with these four hyperparameters. In particular, we can now address the issue of variation in lactic acid concentration more directly. For example, we can compute the posterior probability that $\sigma = \tau^{-1/2}$ is larger than some value such as 0.3:

$$\Pr(\sigma > 0.3|\boldsymbol{x}) = \Pr(\tau < 11.11|\boldsymbol{x}) = 0.984.$$

This can be found using any computer program that calculates the c.d.f. of a gamma distribution. Alternatively, we can use the relationship between the gamma and χ^2 distributions that allows us to say that the posterior distribution of $U = 2 \times 1.0484 \times \tau$ is the χ^2 distribution with $2 \times 5.5 = 11$ degrees of freedom. (See Exercise 1 in Sec. 5.7.) Then $\Pr(\tau < 11.11|\boldsymbol{x}) = \Pr(U \leq 23.30|\boldsymbol{x}) \approx 0.982$ by interpolating in the table of the χ^2 distributions in the back of the book. If $\sigma > 0.3$ is considered a large standard deviation, the cheese manufacturer might wish to look into better quality-control measures. ◀

The Marginal Distribution of the Mean

When the joint distribution of μ and τ is a normal-gamma distribution of the type described in Theorem 8.6.1, then the conditional distribution of μ for a given value of τ is a normal distribution and the marginal distribution of τ is a gamma distribution. It is not clear from this specification, however, what the marginal distribution of μ will be. We shall now derive this marginal distribution.

Theorem 8.6.2 Marginal Distribution of the Mean. Suppose that the prior distribution of μ and τ is the normal-gamma distribution with hyperparameters μ_0, λ_0, α_0, and β_0. Then the marginal distribution of μ is related to a t distribution in the following way:

$$\left(\frac{\lambda_0 \alpha_0}{\beta_0}\right)^{1/2} (\mu - \mu_0)$$

has the t distribution with $2\alpha_0$ degrees of freedom.

Proof Since the conditional distribution of μ given τ is the normal distribution with mean μ_0 and variance $(\lambda_0\tau)^{-1}$, we can use Theorem 5.6.4 to conclude that the conditional distribution of $Z = (\lambda_0\tau)^{1/2}(\mu - \mu_0)$ given τ is the standard normal distribution. We shall continue to let $\xi_2(\tau)$ be the marginal p.d.f. of τ, and let $\xi_1(\mu|\tau)$

be the conditional p.d.f. of μ given τ. Then the joint p.d.f. of Z and τ is

$$f(z, \tau) = (\lambda_0\tau)^{-1/2}\xi_1((\lambda_0\tau)^{-1/2}z + \mu_0|\tau)\xi_2(\tau) = \phi(z)\xi_2(\tau), \tag{8.6.8}$$

where ϕ is the standard normal p.d.f. of Eq. (5.6.6). We see from Eq. (8.6.8) that Z and τ are independent with Z having the standard normal distribution. Next, let $Y = 2\beta_0\tau$. Using the result of Exercise 1 in Sec. 5.7, we find that the distribution of Y is the gamma distribution with parameters α_0 and 1/2, which is also known as the χ^2 distribution with $2\alpha_0$ degrees of freedom. In summary, Y and Z are independent with Z having the standard normal distribution and Y having the χ^2 distribution with $2\alpha_0$ degrees of freedom. It follows from the definition of the t distributions in Sec. 8.4 that

$$U = \frac{Z}{\left(\frac{Y}{2\alpha_0}\right)^{1/2}} = \frac{(\lambda_0\tau)^{1/2}(\mu - \mu_0)}{\left(\frac{2\beta_0\tau}{2\alpha_0}\right)^{1/2}} = \left(\frac{\lambda_0\alpha_0}{\beta_0}\right)^{1/2}(\mu - \mu_0) \tag{8.6.9}$$

has the t distribution with $2\alpha_0$ degrees of freedom. ■

Theorem 8.6.2 can also be used to find the posterior distribution of μ after data are observed. To do that, just replace μ_0 by μ_1, λ_0 by λ_1, α_0 by α_1, and β_0 by β_1 in the statement of the theorem. The reason for this is that the prior and posterior distributions both have the same form, and the theorem depends only on that form. This same reasoning applies to the discussion that follows, including Theorem 8.6.3.

An alternative way to describe the marginal distribution of μ starts by rewriting (8.6.9) as

$$\mu = \left(\frac{\beta_0}{\lambda_0\alpha_0}\right)^{1/2} U + \mu_0. \tag{8.6.10}$$

Now we see that the distribution of μ can be obtained from a t distribution by translating the t distribution so that it is centered at μ_0 rather than at 0, and also changing the scale factor. This makes it straightforward to find the moments (if they exist) of the distribution of μ.

Theorem 8.6.3 Suppose that μ and τ have the joint normal-gamma distribution with hyperparameters μ_0, λ_0, α_0, and β_0. If $\alpha_0 > 1/2$, then $E(\mu) = \mu_0$. If $\alpha_0 > 1$, then

$$\text{Var}(\mu) = \frac{\beta_0}{\lambda_0(\alpha_0 - 1)}. \tag{8.6.11}$$

Proof The mean and the variance of the marginal distribution of μ can easily be obtained from the mean and the variance of the t distributions that are given in Sec. 8.4. Since U in Eq. (8.6.9) has the t distribution with $2\alpha_0$ degrees of freedom, it follows from Section 8.4 that $E(U) = 0$ if $\alpha_0 > 1/2$ and that $\text{Var}(U) = \alpha_0/(\alpha_0 - 1)$ if $\alpha_0 > 1$. Now use Eq. (8.6.10) to see that if $\alpha_0 > 1/2$, then $E(\mu) = \mu_0$. Also, if $\alpha_0 > 1$, then

$$\text{Var}(\mu) = \left(\frac{\beta_0}{\lambda_0\alpha_0}\right)\text{Var}(U).$$

Eq. (8.6.11) now follows directly. ■

Furthermore, the probability that μ lies in any specified interval can, in principle, be obtained from a table of the t distribution or appropriate software. Most statistical packages include functions that can compute the c.d.f. and the quantile function of

a t distribution with arbitrary degrees of freedom, not just integers. Tables typically deal solely with integer degrees of freedom. If necessary, one can interpolate between adjacent degrees of freedom.

As we pointed out already, we can change the prior hyperparameters to posterior hyperparameters in Theorems 8.6.2 and 8.6.3 and translate them into results concerning the posterior marginal distribution of μ. In particular, the posterior distribution of the following random variable is the t distribution with $2\alpha_1$ degrees of freedom:

$$\left(\frac{\lambda_1\alpha_1}{\beta_1}\right)^{1/2}(\mu - \mu_1). \tag{8.6.12}$$

A Numerical Example

Example 8.6.3

Nursing Homes in New Mexico. In 1988, the New Mexico Department of Health and Social Services recorded information from many of its licensed nursing homes. The data were analyzed by Smith, Piland, and Fisher (1992). In this example, we shall consider the annual medical in-patient days X (measured in hundreds) for a sample of 18 nonrural nursing homes. Prior to observing the data, we shall model the value of X for each nursing home as a normal random variable with mean μ and precision τ. To choose a prior mean and variance for μ and τ, we could speak with experts in the field, but for simplicity, we shall just base these on some additional information we have about the numbers of beds in these nursing homes. There are, on average, 111 beds with a sample standard deviation of 43.5 beds. Suppose that our prior opinion is that there is a 50 percent occupancy rate. Then we can naïvely scale up the mean and standard deviation by a factor of 0.5×365 to obtain a prior mean and standard deviation for the number of in-patient days in a year. In units of hundreds of in-patient days per year, this gives us a mean of $0.5 \times 365 \times 1.11 \approx 200$ and a standard deviation of $0.5 \times 365 \times 0.435 \approx 6300^{1/2}$. To map these values into prior hyperparameters, we shall split the variance of 6300 so that half of it is due to variance between the nursing homes and half is the variance of μ. That is, we shall set $\mathrm{Var}(\mu) = 3150$ and $E(\tau) = 1/3150$. We choose $\alpha_0 = 2$ to reflect only a small amount of prior information. Then, since $E(\tau) = \alpha_0/\beta_0$, we find that $\beta_0 = 6300$. Using $E(\mu) = \mu_0$ and (8.6.11), we get $\mu_0 = 200$ and $\lambda_0 = 2$.

Next, we shall determine an interval for μ centered at the point $\mu_0 = 200$ such that the probability that μ lies in this interval is 0.95. Since the random variable U defined by Eq. (8.6.9) has the t distribution with $2\alpha_0$ degrees of freedom, it follows that, for the numerical values just obtained, the random variable $0.025(\mu - 200)$ has the t distribution with four degrees of freedom. The table of the t distribution gives the 0.975 quantile of the t distribution with four degrees of freedom as 2.776. So,

$$\Pr[-2.776 < 0.025(\mu - 200) < 2.776] = 0.95. \tag{8.6.13}$$

An equivalent statement is that

$$\Pr(89 < \mu < 311) = 0.95. \tag{8.6.14}$$

Thus, under the prior distribution assigned to μ and τ, there is probability 0.95 that μ lies in the interval (89, 311).

Suppose now that the following is our sample of 18 observed numbers of medical in-patient days (in hundreds):

128 281 291 238 155 148 154 232 316 96 146 151 100 213 208 157 48 217.

For these observations, which we denote $\boldsymbol{x}$, $\overline{x}_n = 182.17$ and $s_n^2 = 88678.5$. Then, it follows from Theorem 8.6.1 that the joint posterior distribution of μ and τ is the normal-gamma distribution with hyperparameters

$$\mu_1 = 183.95, \qquad \lambda_1 = 20, \qquad \alpha_1 = 11, \qquad \beta_1 = 50925.37. \tag{8.6.15}$$

Hence, the values of the means and the variances of μ and τ, as found from this joint posterior distribution, are

$$\begin{aligned} &E(\mu|\boldsymbol{x}) = \mu_1 = 183.95, && \operatorname{Var}(\mu|\boldsymbol{x}) = \frac{\beta_1}{\lambda_1(\alpha_1 - 1)} = 254.63, \\ &E(\tau|\boldsymbol{x}) = \frac{\alpha_1}{\beta_1} = 2.16 \times 10^{-4}, && \operatorname{Var}(\tau|\boldsymbol{x}) = \frac{\alpha_1}{\beta_1^2} = 4.24 \times 10^{-9}. \end{aligned} \tag{8.6.16}$$

It follows from Eq. (8.6.1) that the mean μ_1 of the posterior distribution of μ is a weighted average of μ_0 and $\overline{x}_n$. In this numerical example, it is seen that μ_1 is quite close to $\overline{x}_n$.

Next, we shall determine the marginal posterior distribution of μ. Let U be the random variable in Eq. (8.6.12), and use the values computed in (8.6.15). Then $U = (0.0657)(\mu - 183.95)$, and the posterior distribution of U is the t distribution with $2\alpha_1 = 22$ degrees of freedom. The 0.975 quantile of this t distribution is 2.074, so

$$\Pr(-2.074 < U < 2.074|\boldsymbol{x}) = 0.95. \tag{8.6.17}$$

An equivalent statement is that

$$\Pr(152.38 < \mu < 215.52|\boldsymbol{x}) = 0.95. \tag{8.6.18}$$

In other words, under the posterior distribution of μ and τ, the probability that μ lies in the interval (152.38, 215.52) is 0.95.

It should be noted that the interval in Eq. (8.6.18) determined from the posterior distribution of μ is much shorter than the interval in Eq. (8.6.14) determined from the prior distribution. This result reflects the fact that the posterior distribution of μ is much more concentrated around its mean than was the prior distribution. The variance of the prior distribution of μ was 3150, and the variance of the posterior distribution is 254.63. Graphs of the prior and posterior p.d.f.'s of μ are in Fig. 8.7 together with the posterior interval (8.6.18). ◀

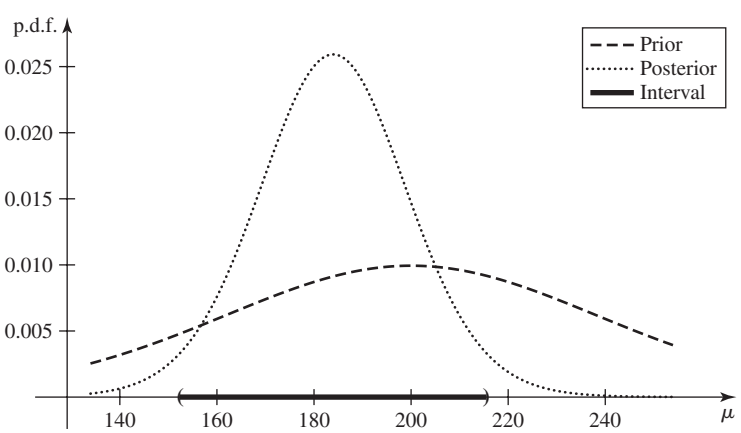

Figure 8.7 Plots of prior and posterior p.d.f.'s of μ in Example 8.6.3. The posterior probability interval (8.6.18) is indicated at the bottom of the graph. The corresponding prior probability interval (8.6.14) would extend far beyond both sides of the plot.

Comparison with Confidence Intervals Continue using the nursing home data from Example 8.6.3. We shall now construct a confidence interval for μ with confidence coefficient 0.95 and compare this interval with the interval in Eq. (8.6.18) for which the posterior probability is 0.95. Since the sample size n in Example 8.6.3 is 18, the random variable U defined by Eq. (8.4.4) on page 481 has the t distribution with 17 degrees of freedom. The 0.975 quantile of this t distribution is 2.110. It now follows from Theorem 8.5.1 that the endpoints of a confidence interval for μ with confidence coefficient 0.95 will be

$$A = \overline{X}_n - 2.110\frac{\sigma'}{n^{1/2}},$$

$$B = \overline{X}_n + 2.110\frac{\sigma'}{n^{1/2}}.$$

When the observed values of $\overline{x}_n = 182.17$ and $s_n^2 = 88678.5$ are used here, we get $\sigma' = (88678.5/17)^{1/2} = 72.22$. The observed confidence interval for μ is then (146.25, 218.09).

This interval is close to the interval (152.38, 215.52) in Eq. (8.6.18), for which the posterior probability is 0.95. The similarity of the two intervals illustrates the statement made at the end of Sec. 8.5. That is, in many problems involving the normal distribution, the method of confidence intervals and the method of using posterior probabilities yield similar results, even though the interpretations of the two methods are quite different.

Improper Prior Distributions

As we discussed at the end of Sec. 7.3 on page 402, it is often convenient to use improper priors that are not real distributions, but do lead to posteriors that are real distributions. These improper priors are chosen more for convenience than to represent anyone's beliefs. When there is a sizeable amount of data, the posterior distribution that results from use of an improper prior is often very close to one that would result from a proper prior distribution. For the case that we have been considering in this section, we can combine the improper prior that we introduced for a location parameter like μ together with the improper prior for a scale parameter like $\sigma = \tau^{-1/2}$ into the usual improper prior for μ and τ. The typical improper prior "p.d.f." for a location parameter was found (in Example 7.3.15) to be the constant function $\xi_1(\mu) = 1$. The typical improper prior "p.d.f." for a scale parameter σ is $g(\sigma) = 1/\sigma$. Since $\sigma = \tau^{-1/2}$, we can apply the techniques of Sec. 3.8 to find the improper "p.d.f." of $\tau = \sigma^{-2}$. The derivative of the inverse function is $-\frac{1}{2}\tau^{-3/2}$, so the improper "p.d.f." of τ would be

$$\left|\frac{1}{2}\tau^{-3/2}\right| g(1/\tau^{-1/2}) = \frac{1}{2}\tau^{-1},$$

for $\tau > 0$. Since this function has infinite integral, we shall drop the factor 1/2 and set $\xi_2(\tau) = \tau^{-1}$. If we act as if μ and τ were independent, then the joint improper prior "p.d.f." for μ and τ is

$$\xi(\mu, \tau) = \frac{1}{\tau}, \quad \text{for } -\infty < \mu < \infty, \tau > 0.$$

If we were to pretend as if this function were a p.d.f., the posterior p.d.f. $\xi(\mu, \tau|\boldsymbol{x})$ would be proportional to

$$\xi(\mu, \tau) f_n(\boldsymbol{x}|\mu, \tau) \propto \tau^{-1}\tau^{n/2} \exp\left(-\frac{\tau}{2}s_n^2 - \frac{n\tau}{2}(\mu - \bar{x}_n)^2\right) \tag{8.6.19}$$

$$= \left\{\tau^{1/2} \exp\left[-\frac{n\tau}{2}(\mu - \bar{x}_n)^2\right]\right\} \tau^{(n-1)/2-1} \exp\left[-\tau\frac{s_n^2}{2}\right].$$

When the expression inside the braces on the far right side of (8.6.19) is regarded as a function of μ for fixed value of τ, this expression can be recognized as being (except for a factor that depends on neither μ nor τ) the p.d.f. of the normal distribution with mean $\bar{x}_n$ and precision $n\tau$. Since the variable μ does not appear elsewhere, it follows that this p.d.f. must be the conditional posterior p.d.f. of μ given τ. It now follows in turn that the expression outside the braces on the far right side of (8.6.19) must be proportional to the marginal posterior p.d.f. of τ. This expression can be recognized as being (except for a constant factor) the p.d.f. of the gamma distribution with parameters $(n-1)/2$ and $s_n^2/2$. This joint distribution would be in precisely the same form as the distribution in Theorem 8.6.1 if our prior distribution had been of the normal-gamma form with hyperparameters $\mu_0 = \beta_0 = \lambda_0 = 0$ and $\alpha_0 = -1/2$. That is, if we pretend as if $\mu_0 = \beta_0 = \lambda_0 = 0$ and $\alpha_0 = -1/2$, and then we apply Theorem 8.6.1, we get the posterior hyperparameters $\mu_1 = \bar{x}_n$, $\lambda_1 = n$, $\alpha_1 = (n-1)/2$, and $\beta_1 = s_n^2/2$.

There is no probability distribution in the normal-gamma family with $\mu_0 = \beta_0 = \lambda_0 = 0$ and $\alpha_0 = -1/2$; however, if we pretend as if this were our prior, then we are said to be using the *usual improper prior distribution*. Notice that the posterior distribution of μ and τ is a real member of the normal-gamma family so long as $n \geq 2$.

Example 8.6.4

An Improper Prior for Seeded Cloud Rainfall. Suppose that we use the usual improper prior for the parameters in Examples 8.3.2 and 8.5.3 with prior hyperparameters $\mu_0 = \beta_0 = \lambda_0 = 0$ and $\alpha_0 = -1/2$. The data summaries are $\bar{x}_n = 5.134$ and $s_n^2 = 63.96$. The posterior distribution will then be the normal-gamma distribution with hyperparameters $\mu_1 = \bar{x}_n = 5.134$, $\lambda_1 = n = 26$, $\alpha_1 = (n-1)/2 = 12.5$, and $\beta_1 = s_n^2/2 = 31.98$. Also, the marginal posterior distribution of μ is given by (7.6.12). In particular,

$$U = \left(\frac{26 \times 12.5}{31.98}\right)^{1/2} (\mu - 5.134) = 3.188(\mu - 5.134) \tag{8.6.20}$$

has the t distribution with 25 degrees of freedom. Suppose that we want an interval (a, b) such that the posterior probability of $a < \mu < b$ is 0.95. The 0.975 quantile of the t distribution with 25 degrees of freedom is 2.060. So, we have that $\Pr(-2.060 < U < 2.060) = 0.95$. Combining this with (8.6.20), we get

$$\Pr(5.134 - 2.060/3.188 < \mu < 5.134 + 2.060/3.188|\boldsymbol{x}) = 0.95.$$

The interval we need runs from $a = 5.134 - 2.060/3.188 = 4.488$ to $b = 5.134 + 2.060/3.188 = 5.780$. Notice that the interval (4.488, 5.780) is precisely the same as the 95% confidence interval for μ that was computed in Example 8.5.3.

Another calculation that we can do with this posterior distribution is to see how likely it is that $\mu > 4$, where 4 is the mean of log-rainfall for unseeded clouds:

$$\Pr(\mu > 4|\boldsymbol{x}) = \Pr(U > 3.188(4 - 5.134)|\boldsymbol{x}) = 1 - T_{25}(-3.615) = 0.9993,$$

where the final value is calculated using statistical software that includes the c.d.f.'s of all t distributions. It appears quite likely, after observing the data, that the mean log-rainfall of seeded clouds is more than 4. ◀

Note: Improper Priors Lead to Confidence Intervals. Example 8.6.4 illustrates one of the more interesting properties of the usual improper prior. If one uses the usual

improper prior with normal data, then the posterior probability is γ that μ is in the observed value of a coefficient γ confidence interval. In general, if we apply (8.6.9) after using an improper prior, we find that the posterior distribution of

$$U = \left(\frac{n(n-1)}{s_n^2}\right)^{1/2} (\mu - \overline{x}_n) \tag{8.6.21}$$

is the t distribution with $n-1$ degrees of freedom. It follows that if $\Pr(-c < U < c) = \gamma$, then

$$\Pr\left(\overline{x}_n - c\frac{\sigma'}{n^{1/2}} < \mu < \overline{x}_n + c\frac{\sigma'}{n^{1/2}} \middle| \boldsymbol{x}\right) = \gamma. \tag{8.6.22}$$

The reader will notice the striking similarity between (8.6.22) and (8.5.3). The difference between the two is that (8.6.22) is a statement about the posterior distribution of μ *after* observing the data, while (8.5.3) is a statement about the conditional distribution of the random variables $\overline{X}_n$ and σ' given μ and σ *before* observing the data. That these two probabilities are the same for all possible data and all possible values of γ follows from the fact that they are both equal to $\Pr(-c < U < c)$ where U is defined either in Eq. (8.4.4) or Eq. (8.6.21). The sampling distribution (conditional on μ and τ) of U is the t distribution with $n-1$ degrees of freedom, as we found in Eq. (8.4.4). The posterior distribution from the improper prior (conditional on the data) of U is also the t distribution with $n-1$ degrees of freedom.

The same kind of thing happens when we try to estimate $\sigma^2 = 1/\tau$. The sampling distribution (conditional on μ and τ) of $V = (n-1)\sigma'^2\tau = (n-1)\sigma'^2/\sigma^2$ is the χ^2 distribution with $n-1$ degrees of freedom, as we saw in Eq. (8.3.11). The posterior distribution from the improper prior (conditional on the data) of V is also the χ^2 distribution with $n-1$ degrees of freedom (see Exercise 4). Therefore, a coefficient γ confidence interval (a, b) for σ^2 based on the sampling distribution of V will satisfy $\Pr(a < \sigma^2 < b|\boldsymbol{x}) = \gamma$ as a posterior probability statement given the data if we used an improper prior.

There are many situations in which the sampling distribution of a pivotal quantity like U above is the same as its posterior distribution when an improper prior is used. A very mathematical treatment of these situations can be found in Schervish (1995, chapter 6). The most common situations are those involving location parameters (like μ) and/or scale parameters (like σ).

Summary

We introduced a family of conjugate prior distributions for the parameters μ and $\tau = 1/\sigma^2$ of a normal distribution. The conditional distribution of μ given τ is normal with mean μ_0 and precision $\lambda_0\tau$, and the marginal distribution of τ is the gamma distribution with parameters α_0 and β_0. If $X_1 = x_1, \ldots, X_n = x_n$ is an observed sample of size n from the normal distribution with mean μ and precision τ, then the posterior distribution of μ given τ is the normal distribution with mean μ_1 and precision $\lambda_1\tau$, and the posterior distribution of τ is the gamma distribution with parameters α_1 and β_1 where the values of μ_1, λ_1, α_1, and β_1 are given in Eq. (8.6.1) and (8.6.2). The marginal posterior distribution of μ is given by saying that $(\lambda_1\alpha_1/\beta_1)^{1/2}(\mu - \mu_1)$ has the t distribution with $2\alpha_1$ degrees of freedom. An interval containing probability $1-\alpha$ of the posterior distribution of μ is

$$\left(\mu_1 - T_{2\alpha_1}^{-1}(1-\alpha/2)\left[\frac{\beta_1}{\alpha_1\lambda_1}\right]^{1/2}, \mu_1 + T_{2\alpha_1}^{-1}(1-\alpha/2)\left[\frac{\beta_1}{\alpha_1\lambda_1}\right]^{1/2}\right).$$

If we use the improper prior with prior hyperparameters $\alpha_0 = -1/2$ and $\mu_0 = \lambda_0 = \beta_0 = 0$, then the random variable $n^{1/2}(\overline{X}_n - \mu)/\sigma'$ has the t distribution with $n - 1$ degrees of freedom both as its posterior distribution given the data and as its sampling distribution given μ and σ. Also, $(n-1)\sigma'^2/\sigma^2$ has the χ^2 distribution with $n - 1$ degrees of freedom both as its posterior distribution given the data and as its sampling distribution given μ and σ. Hence, if we use the improper prior, interval estimates of μ or σ based on the posterior distribution will also be confidence intervals, and vice versa.

Exercises

1. Suppose that a random variable X has the normal distribution with mean μ and precision τ. Show that the random variable $Y = aX + b$ $(a \neq 0)$ has the normal distribution with mean $a\mu + b$ and precision τ/a^2.

2. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ $(-\infty < \mu < \infty)$ and known precision τ. Suppose also that the prior distribution of μ is the normal distribution with mean μ_0 and precision λ_0. Show that the posterior distribution of μ, given that $X_i = x_i$ $(i = 1, \ldots, n)$ is the normal distribution with mean

$$\frac{\lambda_0\mu_0 + n\tau\overline{x}_n}{\lambda_0 + n\tau}$$

and precision $\lambda_0 + n\tau$.

3. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with known mean μ and unknown precision τ $(\tau > 0)$. Suppose also that the prior distribution of τ is the gamma distribution with parameters α_0 and β_0 $(\alpha_0 > 0$ and $\beta_0 > 0)$. Show that the posterior distribution of τ given that $X_i = x_i$ $(i = 1, \ldots, n)$ is the gamma distribution with parameters $\alpha_0 + (n/2)$ and

$$\beta_0 + \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2.$$

4. Suppose that $X_1, \ldots, X_n$ are i.i.d. having the normal distribution with mean μ and precision τ given (μ, τ). Let (μ, τ) have the usual improper prior. Let $\sigma'^2 = s_n^2/(n-1)$. Prove that the posterior distribution of $V = (n-1)\sigma'^2\tau$ is the χ^2 distribution with $n - 1$ degrees of freedom.

5. Suppose that two random variables μ and τ have the joint normal-gamma distribution such that $E(\mu) = -5$. $\text{Var}(\mu) = 1$, $E(\tau) = 1/2$, and $\text{Var}(\tau) = 1/8$. Find the prior hyperparameters μ_0, λ_0, α_0, and β_0 that specify the normal-gamma distribution.

6. Show that two random variables μ and τ cannot have a joint normal-gamma distribution such that $E(\mu) = 0$, $\text{Var}(\mu) = 1$, $E(\tau) = 1/2$, and $\text{Var}(\tau) = 1/4$.

7. Show that two random variables μ and τ cannot have the joint normal-gamma distribution such that $E(\mu) = 0$, $E(\tau) = 1$, and $\text{Var}(\tau) = 4$.

8. Suppose that two random variables μ and τ have the joint normal-gamma distribution with hyperparameters $\mu_0 = 4$, $\lambda_0 = 0.5$, $\alpha_0 = 1$, and $\beta_0 = 8$. Find the values of **(a)** $\Pr(\mu > 0)$ and **(b)** $\Pr(0.736 < \mu < 15.680)$.

9. Using the prior and data in the numerical example on nursing homes in New Mexico in this section, find **(a)** the shortest possible interval such that the posterior probability that μ lies in the interval is 0.90, and **(b)** the shortest possible confidence interval for μ for which the confidence coefficient is 0.90.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown precision τ, and also that the joint prior distribution of μ and τ is the normal-gamma distribution satisfying the following conditions: $E(\mu) = 0$, $E(\tau) = 2$, $E(\tau^2) = 5$, and $\Pr(|\mu| < 1.412) = 0.5$. Determine the prior hyperparameters μ_0, λ_0, α_0, and β_0.

11. Consider again the conditions of Exercise 10. Suppose also that in a random sample of size $n = 10$, it is found that $\overline{x}_n = 1$ and $s_n^2 = 8$. Find the shortest possible interval such that the posterior probability that μ lies in the interval is 0.95.

12. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown precision τ, and also that the joint prior distribution of μ and τ is the normal-gamma distribution satisfying the following conditions: $E(\tau) = 1$, $\text{Var}(\tau) = 1/3$, $\Pr(\mu > 3) = 0.5$, and $\Pr(\mu > 0.12) = 0.9$. Determine the prior hyperparameters μ_0, λ_0, α_0, and β_0.

13. Consider again the conditions of Exercise 12. Suppose also that in a random sample of size $n = 8$, it is found that $\sum_{i=1}^{n} x_i = 16$ and $\sum_{i=1}^{n} x_1^2 = 48$. Find the shortest possible interval such that the posterior probability that μ lies in the interval is 0.99.

14. Continue the analysis in Example 8.6.2 on page 498. Compute an interval (a, b) such that the posterior probability is 0.9 that $a < \mu < b$. Compare this interval with the 90% confidence interval from Example 8.5.4 on page 487.

15. We will draw a sample of size $n = 11$ from the normal distribution with mean μ and precision τ. We will use a natural conjugate prior for the parameters (μ, τ) from the normal-gamma family with hyperparameters $\alpha_0 = 2$, $\beta_0 = 1$, $\mu_0 = 3.5$, and $\lambda_0 = 2$. The sample yields an average of $\overline{x}_n = 7.2$ and $s_n^2 = 20.3$.

a. Find the posterior hyperparameters.

b. Find an interval that contains 95% of the posterior distribution of μ.

16. The study on acid concentration in cheese included a total of 30 lactic acid measurements, the 10 given in Example 8.5.4 on page 487 and the following additional 20:

1.68, 1.9, 1.06, 1.3, 1.52, 1.74, 1.16, 1.49, 1.63, 1.99,
1.15, 1.33, 1.44, 2.01, 1.31, 1.46, 1.72, 1.25, 1.08, 1.25.

a. Using the same prior as in Example 8.6.2 on page 498, compute the posterior distribution of μ and τ based on all 30 observations.

b. Use the posterior distribution found in Example 8.6.2 on page 498 as if it were the prior distribution before observing the 20 observations listed in this problem. Use these 20 new observations to find the posterior distribution of μ and τ and compare the result to the answer to part (a).

17. Consider the analysis performed in Example 8.6.2. This time, use the usual improper prior to compute the posterior distribution of the parameters.

18. Treat the posterior distribution conditional on the first 10 observations found in Exercise 17 as a prior and then observe the 20 additional observations in Exercise 16. Find the posterior distribution of the parameters after observing all of the data and compare it to the distribution found in part (b) of Exercise 16.

19. Consider the situation described in Exercise 7 of Sec. 8.5. Use a prior distribution from the normal-gamma family with values $\alpha_0 = 1$, $\beta_0 = 4$, $\mu_0 = 150$, and $\lambda_0 = 0.5$.

a. Find the posterior distribution of μ and $\tau = 1/\sigma^2$.

b. Find an interval (a, b) such that the posterior probability is 0.90 that $a < \mu < b$.

20. Consider the calorie count data described in Example 7.3.10 on page 400. Now assume that each observation has the normal distribution with unknown mean μ and unknown precision τ given the parameter (μ, τ). Use the normal-gamma conjugate prior distribution with prior hyperparameters $\mu_0 = 0$, $\lambda_0 = 1$, $\alpha_0 = 1$, and $\beta_0 = 60$. The value of s_n^2 is 2102.9.

a. Find the posterior distribution of (μ, τ).

b. Compute $\Pr(\mu > 1|\boldsymbol{x})$.

8.7 Unbiased Estimators

Let δ be an estimator of a function g of a parameter θ. We say that δ is unbiased *if $E_\theta[\delta(\boldsymbol{X})] = g(\theta)$ for all values of θ. This section provides several examples of unbiased estimators.*

Definition of an Unbiased Estimator

Example 8.7.1 Lifetimes of Electronic Components. Consider the company in Example 8.1.3 that wants to estimate the failure rate θ of electronic components. Based on a sample X_1, X_2, X_3 of lifetimes, the M.L.E. of θ is $\hat{\theta} = 3/T$, where $T = X_1 + X_2 + X_3$. The company hopes that $\hat{\theta}$ will be close to θ. The mean of a random variable, such as $\hat{\theta}$, is one measure of where we expect the random variable to be. The mean of $3/T$ is (according to Exercise 21 in Sec. 5.7) $3\theta/2$. If the mean tells us where we expect the estimator to be, we expect this estimator to be 50% larger than θ. ◀

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that involves a parameter (or parameter vector) θ whose value is unknown. Suppose that we wish to estimate a function $g(\theta)$ of the parameter. In a problem of this type, it is desirable to use an estimator $\delta(\boldsymbol{X})$ that, with high probability, will be close to $g(\theta)$. In other words,

it is desirable to use an estimator δ whose distribution changes with the value of θ in such a way that no matter what the true value of θ is, the probability distribution of δ is concentrated around $g(\theta)$.

For example, suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from a normal distribution for which the mean θ is unknown and the variance is 1. In this case, the M.L.E. of θ is the sample mean $\overline{X}_n$. The estimator $\overline{X}_n$ is a reasonably good estimator of θ because its distribution is the normal distribution with mean θ and variance $1/n$. This distribution is concentrated around the unknown value of θ, no matter how large or how small θ is.

These considerations lead to the following definition.

Definition 8.7.1 Unbiased Estimator/Bias. An estimator $\delta(\boldsymbol{X})$ is an *unbiased estimator* of a function $g(\theta)$ of the parameter θ if $E_\theta[\delta(\boldsymbol{X})] = g(\theta)$ for every possible value of θ. An estimator that is not unbiased is called a *biased estimator*. The difference between the expectation of an estimator and $g(\theta)$ is called the *bias* of the estimator. That is, the bias of δ as an estimator of $g(\theta)$ is $E_\theta[\delta(\boldsymbol{X})] - g(\theta)$, and δ is unbiased if and only if the bias is 0 for all θ.

In the case of a sample from a normal distribution with unknown mean θ, $\overline{X}_n$ is an unbiased estimator of θ because $E_\theta(\overline{X}_n) = \theta$ for $-\infty < \theta < \infty$.

Example 8.7.2 Lifetimes of Electronic Components. In Example 8.7.1, the bias of $\hat{\theta} = 3/T$ as an estimator of θ is $3\theta/2 - \theta = \theta/2$. It is easy to see that an unbiased estimator of θ is $\delta(\boldsymbol{X}) = 2/T$. ◀

If an estimator δ of some nonconstant function $g(\theta)$ of the parameter is unbiased, then the distribution of δ must indeed change with the value of θ, since the mean of this distribution is $g(\theta)$. It should be emphasized, however, that this distribution might be either closely concentrated around $g(\theta)$ or widely spread out. For example, an estimator that is equally likely to underestimate $g(\theta)$ by 1,000,000 units or to overestimate $g(\theta)$ by 1,000,000 units would be an unbiased estimator, but it would never yield an estimate close to $g(\theta)$. Therefore, the mere fact that an estimator is unbiased does not necessarily imply that the estimator is good or even reasonable. However, if an unbiased estimator also has a small variance, it follows that the distribution of the estimator will necessarily be concentrated around its mean $g(\theta)$, and there will be high probability that the estimator will be close to $g(\theta)$.

For the reasons just mentioned, the study of unbiased estimators is largely devoted to the search for an unbiased estimator that has a small variance. However, if an estimator δ is unbiased, then its M.S.E. $E_\theta[(\delta - g(\theta))^2]$ is equal to its variance $\text{Var}_\theta(\delta)$. Therefore, the search for an unbiased estimator with a small variance is equivalent to the search for an unbiased estimator with a small M.S.E. The following result is a simple corollary to Exercise 4 in Sec. 4.3.

Corollary 8.7.1 Let δ be an estimator with finite variance. Then the M.S.E. of δ as an estimator of $g(\theta)$ equals its variance plus the square of its bias.

Example 8.7.3 Lifetimes of Electronic Components. We can compare the two estimators $\hat{\theta}$ and $\delta(\boldsymbol{X})$ in Example 8.7.2 using M.S.E. According to Exercise 21 in Sec. 5.7, the variance of $1/T$ is $\theta^2/4$. So, the M.S.E. of $\delta(\boldsymbol{X})$ is θ^2. For $\hat{\theta}$, the variance is $9\theta^2/4$ and the square of the bias is $\theta^2/4$, so the M.S.E. is $5\theta^2/2$, which is 2.5 times as large as the M.S.E. of $\delta(\boldsymbol{X})$. If M.S.E. were the sole concern, the estimator $\delta^*(\boldsymbol{X}) = 1/T$ has variance

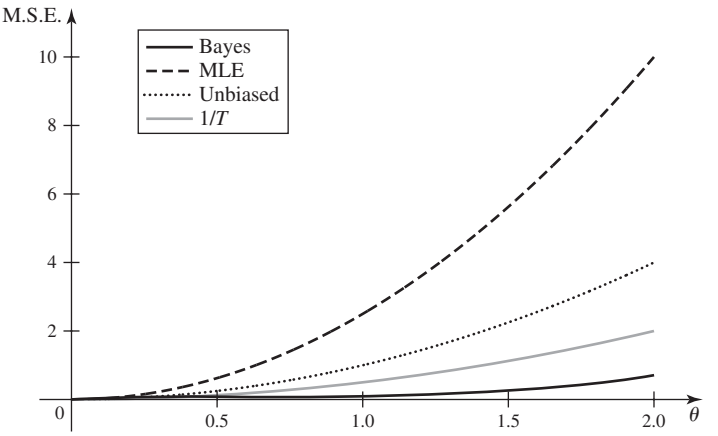

Figure 8.8 M.S.E. for each of the four estimators in Example 8.7.3.

and squared bias both equal to $\theta^2/4$, so the M.S.E. is $\theta^2/2$, half the M.S.E. of the unbiased estimator. Figure 8.8 plots the M.S.E. for each of these estimators together with the M.S.E. of the Bayes estimator $4/(2+T)$ found in Example 8.1.3. Calculation of the M.S.E. of the Bayes estimator required simulation. Eventually (above $\theta = 3.1$), the M.S.E. of the Bayes estimator crosses above the M.S.E. of $1/T$, but it stays below the other two for all θ. ◀

Example 8.7.4 Unbiased Estimation of the Mean. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or parameter vector) θ. Assume that the mean and variance of the distribution are finite. Define $g(\theta) = E_\theta(X_1)$. The sample mean $\overline{X}_n$ is obviously an unbiased estimator of $g(\theta)$. Its M.S.E. is $\mathrm{Var}_\theta(X_1)/n$. In Example 8.7.1, $g(\theta) = 1/\theta$ and $\overline{X}_n = 1/\hat{\theta}$ is an unbiased estimator the mean. ◀

Unbiased Estimation of the Variance

Theorem 8.7.1 Sampling from a General Distribution. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that depends on a parameter (or parameter vector) θ. Assume that the variance of the distribution is finite. Define $g(\theta) = \mathrm{Var}_\theta(X_1)$. The following statistic is an unbiased estimator of the variance $g(\theta)$:

$$\hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Proof Let $\mu = E_\theta(X_1)$, and let σ^2 stand for $g(\theta) = \mathrm{Var}_\theta(X_1)$. Since the sample mean is an unbiased estimator of μ, it is more or less natural to consider first the sample variance $\hat{\sigma}_0^2 = (1/n)\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ and to attempt to determine if it is an unbiased estimator of the variance σ^2. We shall use the identity

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + n(\overline{X}_n - \mu)^2.$$

Then it follows that

$$E(\hat{\sigma}_0^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right]$$
$$= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] - E[(\overline{X}_n - \mu)^2]. \tag{8.7.1}$$

Since each observation X_i has mean μ and variance σ^2, then $E[(X_i - \mu)^2] = \sigma^2$ for $i = 1, \ldots, n$. Therefore,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}E[(X_i - \mu)^2] = \frac{1}{n}n\sigma^2 = \sigma^2. \tag{8.7.2}$$

Furthermore, the sample mean $\overline{X}_n$ has mean μ and variance σ^2/n. Therefore,

$$E[(\overline{X}_n - \mu)^2] = \text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}. \tag{8.7.3}$$

It now follows from Eqs. (8.7.1), (8.7.2), and (8.7.3) that

$$E(\hat{\sigma}_0^2) = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2. \tag{8.7.4}$$

It can be seen from Eq. (8.7.4) that the sample variance $\hat{\sigma}_0^2$ is not an unbiased estimator of σ^2, because its expectation is $[(n-1)/n]\sigma^2$, rather than σ^2. However, if $\hat{\sigma}_0^2$ is multiplied by the factor $n/(n-1)$ to obtain the statistic $\hat{\sigma}_1^2$, then the expectation of $\hat{\sigma}_1^2$ will indeed be σ^2. Therefore, $\hat{\sigma}_1^2$ is an unbiased estimator of σ^2. ■

In light of Theorem 8.7.1, many textbooks define the sample variance as $\hat{\sigma}_1^2$, rather than as $\hat{\sigma}_0^2$.

Note: Special Case of Normal Random Sample. The estimator $\hat{\sigma}_0^2$ is the same as the maximum likelihood estimator $\widehat{\sigma^2}$ of σ^2 when $X_1, \ldots, X_n$ have the normal distribution with mean μ and variance σ^2. Also, $\hat{\sigma}_1^2$ is the same as the random variable σ'^2 that appears in confidence intervals for μ. We have chosen to use different names for these estimators in this section because we are discussing general distributions for which σ^2 might be some function $g(\theta)$ whose M.L.E. is completely different from $\hat{\sigma}_0^2$. (See Exercise 1 for one such example.)

Sampling from a Specific Family of Distributions When it can be assumed that $X_1, \ldots, X_n$ form a random sample from a specific family of distributions, such as the family of Poisson distributions, it will generally be desirable to consider not only $\hat{\sigma}_1^2$ but also other unbiased estimators of the variance.

Example 8.7.5

Sample from a Poisson Distribution. Suppose that we observe a random sample from the Poisson distribution for which the mean θ is unknown. We have already seen that $\overline{X}_n$ will be an unbiased estimator of the mean θ. Moreover, since the variance of a Poisson distribution is also equal to θ, it follows that $\overline{X}_n$ is also an unbiased estimator of the variance. In this example, therefore, both $\overline{X}_n$ and $\hat{\sigma}_1^2$ are unbiased estimators of the unknown variance θ. Furthermore, any combination of $\overline{X}_n$ and $\hat{\sigma}_1^2$ having the form $\alpha\overline{X}_n + (1-\alpha)\hat{\sigma}_1^2$, where α is a given constant $(-\infty < \alpha < \infty)$, will also be an unbiased estimator of θ because its expectation will be

$$E[\alpha\overline{X}_n + (1-\alpha)\hat{\sigma}_1^2] = \alpha E(\overline{X}_n) + (1-\alpha)E(\hat{\sigma}_1^2) = \alpha\theta + (1-\alpha)\theta = \theta. \tag{8.7.5}$$

Other unbiased estimators of θ can also be constructed. ◀

If an unbiased estimator is to be used, the problem is to determine which one of the possible unbiased estimators has the smallest variance or, equivalently, has the smallest M.S.E. We shall not derive the solution to this problem right now. However, it will be shown in Sec. 8.8 that in Example 8.7.5, for every possible value of θ, the estimator $\overline{X}_n$ has the smallest variance among all unbiased estimators of θ. This result is not surprising. We know from Example 7.7.2 that $\overline{X}_n$ is a sufficient statistic for θ, and it was argued in Sec. 7.9 that we can restrict our attention to estimators that are functions of the sufficient statistic alone. (See also Exercise 13 at the end of this section.)

Example 8.7.6

Sampling from a Normal Distribution. Assume that $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. We shall consider the problem of estimating σ^2. We know from Theorem 8.7.1 that the estimator $\hat{\sigma}_1^2$ is an unbiased estimator of σ^2. Moreover, we know from Example 7.5.6 that the sample variance $\hat{\sigma}_0^2$ is the M.L.E. of σ^2. We want to determine whether the M.S.E. $E[(\hat{\sigma}_i^2 - \sigma^2)^2]$ is smaller for the estimator $\hat{\sigma}_0^2$ or for the estimator $\hat{\sigma}_1^2$, and also whether or not there is some other estimator of σ^2 that has a smaller M.S.E. than both $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$.

Both the estimator $\hat{\sigma}_0^2$ and the estimator $\hat{\sigma}_1^2$ have the following form:

$$T_c = c \sum_{i=1}^{n} (X_i - \overline{X}_n)^2, \tag{8.7.6}$$

where $c = 1/n$ for $\hat{\sigma}_0^2$ and $c = 1/(n-1)$ for $\hat{\sigma}_1^2$. We shall now determine the M.S.E. for an arbitrary estimator having the form in Eq. (8.7.6) and shall then determine the value of c for which this M.S.E. is minimum. We shall demonstrate the striking property that the same value of c minimizes the M.S.E. for all possible values of the parameters μ and σ^2. Therefore, among all estimators having the form in Eq. (8.7.6), there is a single one that has the smallest M.S.E. for all possible values of μ and σ^2.

It was shown in Sec. 8.3 that when $X_1, \ldots, X_n$ form a random sample from a normal distribution, the random variable $\sum_{i=1}^{n} (X_i - \overline{X}_n)^2/\sigma^2$ has the χ^2 distribution with $n-1$ degrees of freedom. By Theorem 8.2.1, the mean of this variable is $n-1$, and the variance is $2(n-1)$. Therefore, if T_c is defined by Eq. (8.7.6), then

$$E(T_c) = (n-1)c\sigma^2 \quad \text{and} \quad \text{Var}(T_c) = 2(n-1)c^2\sigma^4. \tag{8.7.7}$$

Thus, by Corollary 8.7.1, the M.S.E. of T_c can be found as follows:

$$\begin{aligned} E[(T_c - \sigma^2)^2] &= [E(T_c) - \sigma^2]^2 + \text{Var}(T^2) \\ &= [(n-1)c - 1]^2\sigma^4 + 2(n-1)c^2\sigma^4 \\ &= [(n^2-1)c^2 - 2(n-1)c + 1]\sigma^4. \end{aligned} \tag{8.7.8}$$

The coefficient of σ^4 in Eq. (8.7.8) is simply a quadratic function of c. Hence, no matter what σ^2 equals, the minimizing value of c is found by elementary differentiation to be $c = 1/(n+1)$.

In summary, we have established the following fact: Among all estimators of σ^2 having the form in Eq. (8.7.6), the estimator that has the smallest M.S.E. for all possible values of μ and σ^2 is $T_{1/(n+1)} = [1/(n+1)] \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$. In particular, $T_{1/(n+1)}$ has a smaller M.S.E. than both the M.L.E. $\hat{\sigma}_0^2$ and the unbiased estimator $\hat{\sigma}_1^2$. Therefore, the estimators $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$, as well as all other estimators having the form in Eq. (8.7.6) with $c \neq 1/(n+1)$, are inadmissible. Furthermore, it was shown

by C. Stein in 1964 that even the estimator $T_{1/(n+1)}$ is dominated by other estimators and that $T_{1/(n+1)}$ itself is therefore inadmissible.

The estimators $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ are compared in Exercise 6 at the end of this section. Of course, when the sample size n is large, it makes little difference whether n, $n-1$, or $n+1$ is used as the divisor in the estimate of σ^2; all three estimators $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$, and $T_{1/(n+1)}$ will be approximately equal. ◀

Limitations of Unbiased Estimation

The concept of unbiased estimation has played an important part in the historical development of statistics, and the feeling that an unbiased estimator should be preferred to a biased estimator is prevalent in current statistical practice. Indeed, what scientist wishes to be biased or to be accused of being biased? The very terminology of the theory of unbiased estimation seems to make the use of unbiased estimators highly desirable.

However, as explained in this section, the quality of an unbiased estimator must be evaluated in terms of its variance or its M.S.E. Examples 8.7.3 and 8.7.6 illustrate the following fact: In many problems, there exist biased estimators that have smaller M.S.E. than every unbiased estimator for every possible value of the parameter. Furthermore, it can be shown that a Bayes estimator, which makes use of all relevant prior information about the parameter and which minimizes the overall M.S.E., is unbiased only in trivial problems in which the parameter can be estimated perfectly.

Some other limitations of the theory of unbiased estimation will now be described.

Nonexistence of an Unbiased Estimator In many problems, there does not exist any unbiased estimator of the function of the parameter that must be estimated. For example, suppose that $X_1, \ldots, X_n$ form n Bernoulli trials for which the parameter p is unknown ($0 \le p \le 1$). Then the sample mean $\overline{X}_n$ will be an unbiased estimator of p, but it can be shown that there will be no unbiased estimator of $p^{1/2}$. (See Exercise 7.) Furthermore, if it is known in this example that p must lie in the interval $\frac{1}{3} \le p \le \frac{2}{3}$, then there is no unbiased estimator of p whose possible values are confined to that same interval.

Inappropriate Unbiased Estimators Consider an infinite sequence of Bernoulli trials for which the parameter p is unknown ($0 < p < 1$), and let X denote the number of failures that occur before the first success is obtained. Then X has the geometric distribution with parameter p whose p.f. is given by Eq. (5.5.3). If it is desired to estimate the value of p from the observation X, then it can be shown (see Exercise 8) that the *only* unbiased estimator of p yields the estimate 1 if $X = 0$ and yields the estimate 0 if $X > 0$. This estimator seems inappropriate. For example, if the first success is obtained on the second trial, that is, if $X = 1$, then it is silly to estimate that the probability of success p is 0. Similarly, if $X = 0$ (the first trial is success), it seems silly to estimate p to be as large as 1.

As another example of an inappropriate unbiased estimator, suppose that the random variable X has the Poisson distribution with unknown mean λ ($\lambda > 0$), and suppose also that it is desired to estimate the value of $e^{-2\lambda}$. It can be shown (see Exercise 9) that the *only* unbiased estimator of $e^{-2\lambda}$ yields the estimate 1 if X is an even integer and the estimate -1 if X is an odd integer. This estimator is inappropriate for two reasons. First, it yields the estimate 1 or -1 for a parameter $e^{-2\lambda}$, which must

lie between 0 and 1. Second, the value of the estimate depends only on whether X is odd or even, rather than on whether X is large or small.

Ignoring Information One more criticism of the concept of unbiased estimation is that the principle of always using an unbiased estimator for a parameter θ (when such exists) sometimes ignores valuable information that is available. As an example, suppose that the average voltage θ in a certain electric circuit is unknown; this voltage is to be measured by a voltmeter for which the reading X has the normal distribution with mean θ and known variance σ^2. Suppose also that the observed reading on the voltmeter is 2.5 volts. Since X is an unbiased estimator of θ in this example, a scientist who wished to use an unbiased estimator would estimate the value of θ to be 2.5 volts.

However, suppose also that after the scientist reported the value 2.5 as his estimate of θ, he discovered that the voltmeter actually truncates all readings at 3 volts, just as in Example 3.2.7 on page 106. That is, the reading of the voltmeter is accurate for any voltage less than 3 volts, but a voltage greater than 3 volts would be reported as 3 volts. Since the actual reading was 2.5 volts, this reading was unaffected by the truncation. Nevertheless, the observed reading would no longer be an unbiased estimator of θ because the distribution of the truncated reading X is not a normal distribution with mean θ. Therefore, if the scientist still wished to use an unbiased estimator, he would have to change his estimate of θ from 2.5 volts to a different value.

Ignoring the fact that the observed reading was accurate seems unacceptable. Since the actual observed reading was only 2.5 volts, it is the same as what would have been observed if there had been no truncation. Since the observed reading is untruncated, it would seem that the fact that there might have been a truncated reading is irrelevant to the estimation of θ. However, since this possibility does change the sample space of X and its probability distribution, it will also change the form of the unbiased estimator of θ.

Summary

An estimator $\delta(\boldsymbol{X})$ of $g(\theta)$ is unbiased if $E_\theta[\delta(\boldsymbol{X})] = g(\theta)$ for all possible values of θ. The bias of an estimator of $g(\theta)$ is $E_\theta[\delta(\boldsymbol{X})] - g(\theta)$. The M.S.E. of an estimator equals its variance plus the square of its bias. The M.S.E. of an unbiased estimator equals its variance.

Exercises

1. Let $X_1, \ldots, X_n$ be a random sample from the Poisson distribution with mean θ.

a. Express the $\text{Var}_\theta(X_i)$ as a function $\sigma^2 = g(\theta)$.

b. Find the M.L.E. of $g(\theta)$ and show that it is unbiased.

2. Suppose that X is a random variable whose distribution is completely unknown, but it is known that all the moments $E(X^k)$, for $k = 1, 2, \ldots$, are finite. Suppose also that $X_1, \ldots, X_n$ form a random sample from this distribution. Show that for $k = 1, 2, \ldots$, the kth sample moment $(1/n)\sum_{i=1}^n X_i^k$ is an unbiased estimator of $E(X^k)$.

3. For the conditions of Exercise 2, find an unbiased estimator of $[E(X)]^2$. *Hint:* $[E(X)]^2 = E(X^2) - \text{Var}(X)$.

4. Suppose that a random variable X has the geometric distribution with unknown parameter p. (See Sec. 5.5.) Find a statistic $\delta(X)$ that will be an unbiased estimator of $1/p$.

5. Suppose that a random variable X has the Poisson distribution with unknown mean λ ($\lambda > 0$). Find a statistic $\delta(X)$ that will be an unbiased estimator of e^{λ}. *Hint:* If $E[\delta(X)] = e^{\lambda}$, then

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

Multiply both sides of this equation by e^{λ}, expand the right side in a power series in λ, and then equate the coefficients of λ^x on both sides of the equation for $x = 0, 1, 2, \ldots$.

6. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Let $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ be the two estimators of σ^2, which are defined as follows:

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \text{ and } \hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i-1}^{n}(X_i - \overline{X}_n)^2.$$

Show that the M.S.E. of $\hat{\sigma}_0^2$ is smaller than the M.S.E. of $\hat{\sigma}_1^2$ for all possible values of μ and σ^2.

7. Suppose that $X_1, \ldots, X_n$ form n Bernoulli trials for which the parameter p is unknown ($0 \le p \le 1$). Show that the expectation of every function $\delta(X_1, \ldots, X_n)$ is a polynomial in p whose degree does not exceed n.

8. Suppose that a random variable X has the geometric distribution with unknown parameter p ($0 < p < 1$). Show that the only unbiased estimator of p is the estimator $\delta(X)$ such that $\delta(0) = 1$ and $\delta(X) = 0$ for $X > 0$.

9. Suppose that a random variable X has the Poisson distribution with unknown mean λ ($\lambda > 0$). Show that the only unbiased estimator of $e^{-2\lambda}$ is the estimator $\delta(X)$ such that $\delta(X) = 1$ if X is an even integer and $\delta(X) = -1$ if X is an odd integer.

10. Consider an infinite sequence of Bernoulli trials for which the parameter p is unknown ($0 < p < 1$), and suppose that sampling is continued until exactly k successes have been obtained, where k is a fixed integer ($k \ge 2$). Let N denote the total number of trials that are needed to obtain the k successes. Show that the estimator $(k-1)/(N-1)$ is an unbiased estimator of p.

11. Suppose that a certain drug is to be administered to two different types of animals A and B. It is known that the mean response of animals of type A is the same as the mean response of animals of type B, but the common value θ of this mean is unknown and must be estimated. It is also known that the variance of the response of animals of type A is four times as large as the variance of the response of animals of type B. Let $X_1, \ldots, X_m$ denote the responses of a random sample of m animals of type A, and let $Y_1, \ldots, Y_n$ denote the responses of an independent random sample of n animals of type B. Finally, consider the estimator $\hat{\theta} = \alpha\overline{X}_m + (1-\alpha)\overline{Y}_n$.

a. For what values of α, m, and n is $\hat{\theta}$ an unbiased estimator of θ?

b. For fixed values of m and n, what value of α yields an unbiased estimator with minimum variance?

12. Suppose that a certain population of individuals is composed of k different strata ($k \ge 2$), and that for $i = 1, \ldots, k$, the proportion of individuals in the total population who belong to stratum i is p_i, where $p_i > 0$ and $\sum_{i=1}^{k} p_i = 1$. We are interested in estimating the mean value μ of a certain characteristic among the total population. Among the individuals in stratum i, this characteristic has mean μ_i and variance σ_i^2, where the value of μ_i is unknown and the value of σ_i^2 is known. Suppose that a *stratified sample* is taken from the population as follows: From each stratum i, a random sample of n_i individuals is taken, and the characteristic is measured for each of these individuals. The samples from the k strata are taken independently of each other. Let $\overline{X}_i$ denote the average of the n_i measurements in the sample from stratum i.

a. Show that $\mu = \sum_{i=1}^{k} p_i\mu_i$, and show also that $\hat{\mu} = \sum_{i=1}^{k} p_i\overline{X}_i$ is an unbiased estimator of μ.

b. Let $n = \sum_{i=1}^{k} n_i$ denote the total number of observations in the k samples. For a fixed value of n, find the values of $n_1, \ldots, n_k$ for which the variance of $\hat{\mu}$ will be a minimum.

13. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, where the value of the parameter θ is unknown. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$, and let T be a statistic. Assume that $\delta(\boldsymbol{X})$ is an unbiased estimator of θ such that $E_\theta[\delta(\boldsymbol{X})|T]$ does not depend on θ. (If T is a sufficient statistic, as defined in Sec. 7.7, then this will be true for every estimator δ. The condition also holds in other examples.) Let $\delta_0(T)$ denote the conditional mean of $\delta(\boldsymbol{X})$ given T.

a. Show that $\delta_0(T)$ is also an unbiased estimator of θ.

b. Show that $\mathrm{Var}_\theta(\delta_0) \le \mathrm{Var}_\theta(\delta)$ for every possible value of θ. *Hint:* Use the result of Exercise 11 in Sec. 4.7.

14. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where the value of the parameter θ is unknown; and let $Y_n = \max(X_1, \ldots, X_n)$. Show that $[(n+1)/n]Y_n$ is an unbiased estimator of θ.

15. Suppose that a random variable X can take only the five values $x = 1, 2, 3, 4, 5$ with the following probabilities:

$$f(1|\theta) = \theta^3, \quad f(2|\theta) = \theta^2(1-\theta),$$
$$f(3|\theta) = 2\theta(1-\theta), \; f(4|\theta) = \theta(1-\theta)^2,$$
$$f(5|\theta) = (1-\theta)^3.$$

Here, the value of the parameter θ is unknown ($0 \le \theta \le 1$).

a. Verify that the sum of the five given probabilities is 1 for every value of θ.

b. Consider an estimator $\delta_c(X)$ that has the following form:

$$\delta_c(1) = 1,\ \delta_c(2) = 2 - 2c,\ \delta_c(3) = c,$$
$$\delta_c(4) = 1 - 2c,\ \delta_c(5) = 0.$$

Show that for each constant c, $\delta_c(X)$ is an unbiased estimator of θ.

c. Let θ_0 be a number such that $0 < \theta_0 < 1$. Determine a constant c_0 such that when $\theta = \theta_0$, the variance of $\delta_{c_0}(X)$ is smaller than the variance of $\delta_c(X)$ for every other value of c.

16. Reconsider the conditions of Exercise 3. Suppose that $n = 2$, and we observe $X_1 = 2$ and $X_2 = -1$. Compute the value of the unbiased estimator of $[E(X)]^2$ found in Exercise 3. Describe a flaw that you have discovered in the estimator.

⋆ 8.8 Fisher Information

This section introduces a method for measuring the amount of information that a sample of data contains about an unknown parameter. This measure has the intuitive properties that more data provide more information, and more precise data provide more information. The information measure can be used to find bounds on the variances of estimators, and it can be used to approximate the variances of estimators obtained from large samples.

Definition and Properties of Fisher Information

Example 8.8.1

Studying Customer Arrivals. A store owner is interested in learning about customer arrivals. She models arrivals during the day as a Poisson process (see Definition 5.4.2) with unknown rate θ. She thinks of two different possible sampling plans to obtain information about customer arrivals. One plan is to choose a fixed number, n, of customers and to see how long, X, it takes until n customers arrive. The other plan is to observe for a fixed length of time, t, and count how many customers, Y, arrive during time t. That is, the store owner can either observe a Poisson random variable, Y, with mean $t\theta$ or observe a gamma random variable, X, with parameters n and θ. Is there any way to address the question of which sampling plan is likely to be more informative? ◀

The Fisher information is one property of a distribution that can be used to measure how much information one is likely to obtain from a random variable or a random sample.

The Fisher Information in a Single Random Variable In this section, we shall introduce a concept, called the Fisher information, that enters various aspects of the theory of statistical inference, and we shall describe a few uses of this concept.

Consider a random variable X for which the p.f. or the p.d.f. is $f(x|\theta)$. It is assumed that $f(x|\theta)$ involves a parameter θ whose value is unknown but must lie in a given open interval Ω of the real line. Furthermore, it is assumed that X takes values in a specified sample space S, and $f(x|\theta) > 0$ for each value of $x \in S$ and each value of $\theta \in \Omega$. This assumption eliminates from consideration the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown, because, for that distribution, $f(x|\theta) > 0$ only when $x < \theta$ and $f(x|\theta) = 0$ when $x > \theta$. The assumption does not eliminate any distribution where the set of values of x for which $f(x|\theta) > 0$ is a fixed set that does not depend on θ.

Next, we define $\lambda(x|\theta)$ as follows:

$$\lambda(x|\theta) = \log f(x|\theta).$$

It is assumed that for each value of $x \in S$, the p.f. or p.d.f. $f(x|\theta)$ is a twice differentiable function of θ, and we let

$$\lambda'(x|\theta) = \frac{\partial}{\partial\theta}\lambda(x|\theta) \quad \text{and} \quad \lambda''(x|\theta) = \frac{\partial^2}{\partial\theta^2}\lambda(x|\theta).$$

Definition 8.8.1 Fisher Information in a Random Variable. Let X be a random variable whose distribution depends on a parameter θ that takes values in an open interval Ω of the real line. Let the p.f. or p.d.f. of X be $f(x|\theta)$. Assume that the set of x such that $f(x|\theta) > 0$ is the same for all θ and that $\lambda(x|\theta) = \log f(x|\theta)$ is twice differentiable as a function of θ. The *Fisher information* $I(\theta)$ in the random variable X is defined as

$$I(\theta) = E_\theta\{[\lambda'(X|\theta)]^2\}. \tag{8.8.1}$$

Thus, if $f(x|\theta)$ is a p.d.f., then

$$I(\theta) = \int_S [\lambda'(x|\theta)]^2 f(x|\theta)\, dx. \tag{8.8.2}$$

If $f(x|\theta)$ is a p.f., the integral in Eq. (8.8.2) is replaced by a sum over the points in S. In the discussion that follows, we shall assume for convenience that $f(x|\theta)$ is a p.d.f. However, all the results hold also when $f(x|\theta)$ is a p.f.

An alternative method for calculating the Fisher information sometimes proves more useful.

Theorem 8.8.1 Assume the conditions of Definition 8.8.1. Also, assume that two derivatives of $\int_S f(x|\theta)dx$ with respect to θ can be calculated by reversing the order of integration and differentiation. Then the Fisher information also equals

$$I(\theta) = -E_\theta[\lambda''(X|\theta)]. \tag{8.8.3}$$

Another expression for the Fisher information is

$$I(\theta) = \text{Var}_\theta[\lambda'(X|\theta)]. \tag{8.8.4}$$

Proof We know that $\int_S f(x|\theta)\, dx = 1$ for every value of $\theta \in \Omega$. Therefore, if the integral on the left side of this equation is differentiated with respect to θ, the result will be 0. We have assumed that we can reverse the order in which we perform the integration with respect to x, and the differentiation with respect to θ, and will still obtain the value 0. In other words, we shall assume that we can take the derivative inside the integral sign and obtain

$$\int_S f'(x|\theta)\, dx = 0 \quad \text{for } \theta \in \Omega. \tag{8.8.5}$$

Furthermore, we have assumed that we can take a second derivative with respect to θ "inside the integral sign" and obtain

$$\int_S f''(x|\theta)\, dx = 0 \quad \text{for } \theta \in \Omega. \tag{8.8.6}$$

Since $\lambda'(x|\theta) = f'(x|\theta)/f(x|\theta)$, then

$$E_\theta[\lambda'(X|\theta)] = \int_S \lambda'(x|\theta) f(x|\theta)\, dx = \int_S f'(x|\theta)\, dx.$$

Hence, it follows from Eq. (8.8.5) that

$$E_\theta[\lambda'(X|\theta)] = 0. \tag{8.8.7}$$

Since the mean of $\lambda'(X|\theta)$ is 0, it follows from Eq. (8.8.1) that Eq. (8.8.4) holds.

Next, note that

$$\lambda''(x|\theta) = \frac{f(x|\theta)\, f''(x|\theta) - [f'(x|\theta)]^2}{[f(x|\theta)]^2}$$
$$= \frac{f''(x|\theta)}{f(x|\theta)} - [\lambda'(x|\theta)]^2.$$

Therefore,

$$E_\theta[\lambda''(X|\theta)] = \int_S f''(x|\theta)\, dx - I(\theta). \tag{8.8.8}$$

It follows from Eqs. (8.8.8) and (8.8.6) that Eq. (8.8.3) holds. ■

In many problems, it is easier to determine the value of $I(\theta)$ from Eq. (8.8.3) than from Eqs. (8.8.1) or (8.8.4).

Example 8.8.2 The Bernoulli Distributions. Suppose that X has the Bernoulli distribution with parameter p. We shall determine the Fisher information $I(p)$ in X.

In this example, the possible values of X are the two values 0 and 1. For $x = 0$ or 1,

$$\lambda(x|p) = \log f(x|p) = x \log p + (1 - x) \log(1 - p).$$

Hence,

$$\lambda'(x|p) = \frac{x}{p} - \frac{1 - x}{1 - p}$$

and

$$\lambda''(x|p) = -\left[\frac{x}{p^2} + \frac{1 - x}{(1 - p)^2}\right].$$

Since $E(X) = p$, the Fisher information is

$$I(p) = -E[\lambda''(X|p)] = \frac{1}{p} + \frac{1}{1 - p} = \frac{1}{p(1 - p)}.$$

Recall from Eq. (4.3.3) that $\mathrm{Var}(X) = p(1 - p)$, so the more precise (smaller variance) X is the more information it provides.

In this example, it can be readily verified that the assumptions made in the proof of Theorem 8.8.1 are satisfied. Indeed, because X can take only the two values 0 and 1, the integrals in Eqs. (8.8.5) and (8.8.6) reduce to summations over the two values $x = 0$ and $x = 1$. Since it is always possible to take a derivative "inside a finite summation" and to differentiate the sum term by term, Eqs. (8.8.5) and (8.8.6) must be satisfied. ◀

Example 8.8.3 The Normal Distributions. Suppose that X has the normal distribution with unknown mean μ and known variance σ^2. We shall determine the Fisher information $I(\mu)$ in X.

For $-\infty < x < \infty$,

$$\lambda(x|\mu) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x - \mu)^2}{2\sigma^2}.$$

Hence,

$$\lambda'(x|\mu) = \frac{x-\mu}{\sigma^2} \quad \text{and} \quad \lambda''(x|\mu) = -\frac{1}{\sigma^2}.$$

It now follows from Eq. (8.8.3) that the Fisher information is

$$I(\mu) = \frac{1}{\sigma^2}.$$

Since $\text{Var}(X) = \sigma^2$, we see again that the more precise (smaller variance) X is, the more information it provides.

In this example, it can be verified directly (see Exercise 1 at the end of this section) that Eqs. (8.8.5) and (8.8.6) are satisfied. ◀

It should be emphasized that the concept of Fisher information cannot be applied to a distribution, such as the uniform distribution on the interval $[0, \theta]$, for which the necessary assumptions are not satisfied.

The Fisher Information in a Random Sample When we have a random sample from a distribution, the Fisher information is defined in an analogous manner. Indeed, Definition 8.8.2 subsumes Definition 8.8.1 as the special case in which $n = 1$.

Definition 8.8.2 Fisher Information in a Random Sample. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from a distribution for which the p.f. or p.d.f. is $f(x|\theta)$, where the value of the parameter θ must lie in an open interval Ω of the real line. Let $f_n(\boldsymbol{x}|\theta)$ denote the joint p.f. or joint p.d.f. of $\boldsymbol{X}$. Define

$$\lambda_n(\boldsymbol{x}|\theta) = \log f_n(x|\theta). \tag{8.8.9}$$

Assume that the set of $\boldsymbol{x}$ such that $f_n(\boldsymbol{x}|\theta) > 0$ is the same for all θ and that $\log f_n(\boldsymbol{x}|\theta)$ is twice differentiable with respect to θ. The *Fisher information* $I_n(\theta)$ in the random sample $\boldsymbol{X}$ is defined as

$$I_n(\theta) = E_\theta\{[\lambda_n'(\boldsymbol{X}|\theta)]^2\}.$$

For continuous distributions, the Fisher information $I_n(\theta)$ in the entire sample is given by the following n-dimensional integral:

$$I_n(\theta) = \int_S \cdots \int_S [\lambda_n'(\boldsymbol{x}|\theta)]^2 f_n(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n.$$

For discrete distributions, replace the n-dimensional integral by an n-fold summation.

Furthermore, if we again assume that derivatives can be passed under the integrals, then we may express $I_n(\theta)$ in either of the following two ways:

$$I_n(\theta) = \text{Var}_\theta[\lambda_n'(\boldsymbol{X}|\theta)] \tag{8.8.10}$$

or

$$I_n(\theta) = -E_\theta[\lambda_n''(\boldsymbol{X}|\theta)]. \tag{8.8.11}$$

We shall now show that there is a simple relation between the Fisher information $I_n(\theta)$ in the entire sample and the Fisher information $I(\theta)$ in a single observation X_i.

Theorem 8.8.2 Under the conditions of Definitions 8.8.1 and 8.8.2,

$$I_n(\theta) = nI(\theta). \tag{8.8.12}$$

In words, the Fisher information in a random sample of n observations is simply n times the Fisher information in a single observation.

Proof Since $f_n(\boldsymbol{x}|\theta) = f(x_1|\theta) \ldots f(x_n|\theta)$, it follows that

$$\lambda_n(\boldsymbol{x}|\theta) = \sum_{i=1}^{n} \lambda(x_i|\theta).$$

Hence,

$$\lambda_n''(\boldsymbol{x}|\theta) = \sum_{i=1}^{n} \lambda''(x_i|\theta). \tag{8.8.13}$$

Since each observation X_i has the p.d.f. $f(x|\theta)$, the Fisher information in each X_i is $I(\theta)$. It follows from Eqs. (8.8.3) and (8.8.11) that by taking expectations on both sides of Eq. (8.8.13), we obtain Eq. (8.8.12). ■

Example 8.8.4 Studying Customer Arrivals. Return to the store owner in Example 8.8.1 who is trying to choose between sampling a Poisson random variable, Y, with mean $t\theta$ or sampling a gamma random variable, X, with parameters n and θ. The reader can compute the Fisher information in each random variable in Exercises 3 and 19 in this section. We shall label them $I_Y(\theta)$ and $I_X(\theta)$. They are

$$I_X(\theta) = \frac{n}{\theta^2} \quad \text{and} \quad I_Y(\theta) = \frac{t}{\theta}.$$

Which is larger will clearly depend on the particular values of n, t, and θ. Both n and t can be chosen by the store owner, but θ is unknown. In order for $I_X(\theta) = I_Y(\theta)$, it is necessary and sufficient that $n = t\theta$. This relation actually makes intuitive sense. For example, if the store owner chooses to observe Y, then the total number N of customers observed will be random and $N = Y$. The mean of N is then $E(Y) = t\theta$. Similarly, if the store owner chooses to observe X, then the length of time T that it takes to observe n customers will be random. In fact, $T = X$, and the mean of $T\theta$ is n. So long as the manufacturer is comparing sampling plans that are expected to observe the same numbers of customers or observe for the same length of time, the two sampling plans should provide the same amount of information. ◀

The Information Inequality

Example 8.8.5 Studying Customer Arrivals. Another way that the store owner in Example 8.8.4 could choose between the two sampling plans is to compare the estimators that she will use to make inferential statements about customer arrivals. For example, she may want to estimate θ, the rate of customer arrivals. Alternatively, she may want to estimate $1/\theta$, the mean time between customer arrivals. Each sampling plan lends itself to estimation of both parameters. Indeed, there are unbiased estimators of both parameters available from at least one of these sampling plans. ◀

As one application of the results that have been derived concerning Fisher information, we shall show how the Fisher information can be used to determine a lower bound for the variance of an arbitrary estimator of the parameter θ in a given problem. The following result was independently developed by H. Cramér and C. R. Rao during the 1940s.

Theorem 8.8.3 Cramér-Rao (Information) Inequality. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from a distribution for which the p.d.f. is $f(x|\theta)$. Suppose also that all the

assumptions which have been made about $f(x|\theta)$ thus far in this section continue to hold. Let $T = r(\boldsymbol{X})$ be a statistic with finite variance. Let $m(\theta) = E_\theta(T)$. Assume that $m(\theta)$ is a differentiable function of θ. Then

$$\text{Var}_\theta(T) \geq \frac{[m'(\theta)]^2}{nI(\theta)}. \tag{8.8.14}$$

There will be equality in (8.8.14) if and only if there exist functions $u(\theta)$ and $v(\theta)$ that may depend on θ but do not depend on $\boldsymbol{X}$ and that satisfy the relation

$$T = u(\theta)\lambda_n'(\boldsymbol{X}|\theta) + v(\theta). \tag{8.8.15}$$

Proof The inequality derives from applying Theorem 4.6.3 to the covariance between T and the random variable $\lambda_n'(\boldsymbol{X}|\theta)$ defined in Eq. (8.8.9). Since $\lambda_n'(\boldsymbol{x}|\theta) = f_n'(\boldsymbol{x}|\theta)/f_n(\boldsymbol{x}|\theta)$, it follows just as for a single observation that

$$E_\theta[\lambda_n'(\boldsymbol{X}|\theta)] = \int_S \cdots \int_S f_n'(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n = 0.$$

Therefore,

$$\begin{aligned}\text{Cov}_\theta[T, \lambda_n'(\boldsymbol{X}|\theta)] &= E_\theta[T\lambda_n'(\boldsymbol{X}|\theta)] \\ &= \int_S \cdots \int_S r(\boldsymbol{x})\lambda_n'(\boldsymbol{x}|\theta) f_n(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n \\ &= \int_S \cdots \int_S r(\boldsymbol{x}) f_n'(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n. \end{aligned} \tag{8.8.16}$$

Next, write

$$m(\theta) = \int_S \cdots \int_S r(\boldsymbol{x}) f_n(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n \quad \text{for } \theta \in \Omega. \tag{8.8.17}$$

Finally, suppose that when both sides of Eq. (8.8.17) are differentiated with respect to θ, the derivative can be taken "inside the integrals" on the left side. Then

$$m'(\theta) = \int_S \cdots \int_S r(\boldsymbol{x}) f_n'(\boldsymbol{x}|\theta)\, dx_1 \ldots dx_n \quad \text{for } \theta \in \Omega. \tag{8.8.18}$$

It follows from Eqs. (8.8.16) and (8.8.18) that

$$\text{Cov}_\theta[T, \lambda_n'(\boldsymbol{X}|\theta)] = m'(\theta) \quad \text{for } \theta \in \Omega. \tag{8.8.19}$$

Theorem 4.6.3 says that

$$\{\text{Cov}_\theta[T, \lambda_n'(\boldsymbol{X}|\theta)]\}^2 \leq \text{Var}_\theta(T)\, \text{Var}_\theta[\lambda_n'(\boldsymbol{X}|\theta)]. \tag{8.8.20}$$

Therefore, it follows from Eqs. (8.8.10), (8.8.12), (8.8.19), and (8.8.20) that Eq. (8.8.14) holds.

Finally, notice that (8.8.14) is an equality if and only if (8.8.20) is an equality. This, in turn, is an equality if and only if there exist nonzero constants a and b and a constant c such that $aT + b\lambda_n(\boldsymbol{X}|\theta) = c$. This last claim follows from the similar statement in Theorem 4.6.3. In all of the calculations concerned with Fisher infomration, we have been treating θ as a constant; hence, the constants a, b, and c just mentioned can depend on θ, but must not depend on $\boldsymbol{X}$. Then $u(\theta) = b/a$ and $v(\theta) = c/a$. ■

The following simple corollary to Theorem 8.8.3 gives a lower bound on the variance of an unbiased estimator of θ.

Corollary 8.8.1 Cramér-Rao Lower Bound on the Variance of an Unbiased Estimator. Assume the assumptions of Theorem 8.8.3. Let T be an unbiased estimator of θ. Then

$$\text{Var}_\theta(T) \geq \frac{1}{nI(\theta)}.$$

Proof Because T is an unbiased estimator of θ, $m(\theta) = \theta$ and $m'(\theta) = 1$ for every value of $\theta \in \Omega$. Now apply Eq. (8.8.14). ■

In words, Corollary 8.8.1 says that the variance of an unbiased estimator of θ cannot be smaller than the reciprocal of the Fisher information in the sample.

Example 8.8.6 Unbiased Estimation of the Parameter of an Exponential Distribution. Let $X_1, \ldots, X_n$ be a random sample of size $n > 2$ from the exponential distribution with parameter β. That is, each X_i has p.d.f. $f(x|\beta) = \beta \exp(-\beta x)$ for $x > 0$. Then

$$\lambda(x|\beta) = \log(\beta) - \beta x,$$
$$\lambda'(x|\beta) = \frac{1}{\beta} - x,$$
$$\lambda''(x|\beta) = -\frac{1}{\beta^2}.$$

It can be verified that the conditions required to establish (8.8.3) hold in this example. Then the Fisher information in one observation is

$$I(\beta) = -E_\theta\left[-\frac{1}{\beta^2}\right] = \frac{1}{\beta^2}.$$

The information in the whole sample is then $I_n(\beta) = n/\beta^2$. Consider the estimator $T = (n-1)/\sum_{i=1}^n X_i$. Theorem 5.7.7 says that $\sum_{i=1}^n X_i$ has the gamma distribution with parameters n and β. In Exercise 21 in Sec. 5.7, you proved that the mean and variance of $1/\sum_{i=1}^n X_i$ are $\beta/(n-1)$ and $\beta^2/[(n-1)^2(n-2)]$, respectively. Thus, T is unbiased and its variance is $\beta^2/(n-2)$. The variance is indeed larger than the lower bound, $1/I_n(\beta) = \beta^2/n$. The reason the inequality is strict is that T is not a linear function of $\lambda_n'(\boldsymbol{X}|\theta)$. Indeed, T is 1 over a linear function of $\lambda_n'(\boldsymbol{X}|\theta)$.

On the other hand, if we wish to estimate $m(\beta) = 1/\beta$, $U = \overline{X}_n$ is an unbiased estimator with variance $1/(n\beta^2)$. The information inequality says that the lower bound on the variance of an estimator of $1/\beta$ is

$$\frac{m'(\beta)^2}{n/\beta^2} = \frac{(-1/\beta^2)^2}{n/\beta^2} = \frac{1}{n\beta^2}.$$

In this case, we see that there is equality in (8.8.14). ◀

Example 8.8.7 Studying Customer Arrivals. Return to the store owner in Example 8.8.5 who wants to compare the estimators of θ and $1/\theta$ that she could compute from either the Poisson random variable Y or the gamma random variable X. The case of unbiased estimators based on X was already handled in Example 8.8.6, where our X has the same distribution as $\sum_{i=1}^n X_i$ in that example when $\theta = \beta$. Hence, X/n is an unbiased estimator of $1/\theta$ whose variance equals the Cramér-Rao lower bound, and $(n-1)/X$ is an unbiased estimator of θ whose variance is strictly larger than the lower bound. Since $E_\theta(Y) = t\theta$, we see that Y/t is an unbiased estimator of θ whose variance is also known to be θ/t, which is the Cramér-Rao lower bound. Unfortunately, there is no

unbiased estimator of $1/\theta$ based on Y alone. The estimator $\delta(Y) = t/(Y+1)$ satisfies

$$E_\theta[\delta(Y)] = \frac{1}{\theta}\left[1 - e^{-t\theta}\right].$$

If t is large and θ is not too small, the bias will be small, but it is impossible to find an unbiased estimator. The reason is that the mean of every function of Y is $\exp(-t\theta)$ times a power series in θ. Every such function is differentiable in a neighborhood of $\theta = 0$. The function $1/\theta$ is not differentiable at $\theta = 0$. ◀

Efficient Estimators

Example 8.8.8 Variance of a Poisson Distribution. In Example 8.7.5, we presented a collection of different unbiased estimators of the the variance of a Poisson distribution based on a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from that distribution. After that example, we made the claim that one of the estimators has the smallest variance among the entire collection. The information inequality gives us a way to address comparisons of such collections of estimators without necessarily listing them all or computing their variances. ◀

An estimator whose variance equals the Cramér-Rao lower bound makes the most efficient use of the data $\boldsymbol{X}$ in some sense.

Definition 8.8.3 Efficient Estimator. It is said that an estimator T is an *efficient estimator of its expectation* $m(\theta)$ if there is equality in (8.8.14) for every value of $\theta \in \Omega$.

One difficulty with Definition 8.8.3 is that, in a given problem, there may be no estimator of a particular function $m(\theta)$ whose variance actually attains the Cramér-Rao lower bound. For example, if the random variable X has the normal distribution for which the mean is 0 and the standard deviation σ is unknown ($\sigma > 0$), then it can be shown that the variance of every unbiased estimator of σ based on the single observation X is strictly greater than $1/I(\sigma)$ for every value of $\sigma > 0$ (see Exercise 9). In Example 8.8.6, no efficient estimator of β exists.

On the other hand, in many standard estimation problems there do exist efficient estimators. Of course, the estimator that is identically equal to a constant is an efficient estimator of that constant, since the variance of this estimator is 0. However, as we shall now show, there are often efficient estimators of more interesting functions of θ as well.

According to Theorem 8.8.3, there will be equality in the information inequality (8.8.14) if and only if the estimator T is a linear function of $\lambda_n'(\boldsymbol{X}|\theta)$. It is possible that the only efficient estimators in a given problem will be constants. The reason is as follows: Because T is an estimator, it cannot involve the parameter θ. Therefore, in order for T to be efficient, it must be possible to find functions $u(\theta)$ and $v(\theta)$ such that the parameter θ will actually be canceled from the right side of Eq. (8.8.15), and the value of T will depend only on the observations $\boldsymbol{X}$ and not on θ.

Example 8.8.9 Sampling from a Poisson Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with unknown mean θ ($\theta > 0$). We shall show that $\overline{X}_n$ is an efficient estimator of θ.

The joint p.f. of $X_1, \ldots, X_n$ can be written in the form

$$f_n(\boldsymbol{x}|\theta) = \frac{e^{-n\theta}\theta^{n\overline{x}_n}}{\Pi_{i=I}^n(x_i!)}.$$

Therefore,

$$\lambda_n(\boldsymbol{X}|\theta) = -n\theta + n\overline{X}_n \log\theta - \sum_{i=1}^{n} \log(X_i!)$$

and

$$\lambda_n'(\boldsymbol{X}|\theta) = -n + \frac{n\overline{X}_n}{\theta}. \tag{8.8.21}$$

If we now let $u(\theta) = \theta/n$ and $v(\theta) = \theta$, then it is found from Eq. (8.8.21) that

$$\overline{X}_n = u(\theta)\lambda_n'(\boldsymbol{X}|\theta) + v(\theta).$$

Since the statistic $\overline{X}_n$ has been represented as a linear function of $\lambda_n'(\boldsymbol{X}|\theta)$, it follows that $\overline{X}_n$ is an efficient estimator of its expectation θ. In other words, the variance of $\overline{X}_n$ will attain the lower bound given by the information inequality, which in this example is θ/n (see Exercise 3). This fact can also be verified directly. ◀

Unbiased Estimators with Minimum Variance Suppose that in a given problem a particular estimator T is an efficient estimator of its expectation $m(\theta)$, and let T_1 denote any other unbiased estimator of $m(\theta)$. Then for every value of $\theta \in \Omega$, $\mathrm{Var}_\theta(T)$ will be equal to the lower bound provided by the information inequality, and $\mathrm{Var}_\theta(T_1)$ will be at least as large as that lower bound. Hence, $\mathrm{Var}_\theta(T) \le \mathrm{Var}_\theta(T_1)$ for $\theta \in \Omega$. In other words, if T is an efficient estimator of $m(\theta)$, then among all unbiased estimators of $m(\theta)$, T will have the smallest variance for every possible value of θ.

Example 8.8.10

Variance of a Poisson Distribution. In Example 8.8.9, we saw that $\overline{X}_n$ is an efficient estimator of the mean θ of a Poisson distribution. Therefore, for every value of $\theta > 0$, $\overline{X}_n$ has the smallest variance among all unbiased estimators of θ. Since θ is also the variance of the Poisson distribution with mean θ, we know that $\overline{X}_n$ has the smallest variance among all unbiased estimators of the variance. This establishes the claim that was made without proof after Example 8.7.5. In particular, the estimator $\hat{\sigma}_1^2$ in Example 8.7.5 is not a linear function of $\lambda_n'(\boldsymbol{X}|\theta)$, and hence its variance must be strictly larger than Cramér-Rao lower bound. Similarly, the other estimators in Eq. (8.7.5) must each have variance larger than the Cramér-Rao lower bound. ◀

Properties of Maximum Likelihood Estimators for Large Samples

Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. or the p.f. is $f(x|\theta)$, and suppose also that $f(x|\theta)$ satisfies conditions similar to those which were needed to derive the information inequality. For each sample size n, let $\hat{\theta}_n$ denote the M.L.E. of θ. We shall show that if n is large, then the distribution of $\hat{\theta}_n$ is approximately the normal distribution with mean θ and variance $1/[nI(\theta)]$.

Theorem 8.8.4

Asymptotic Distribution of an Efficient Estimator. Assume the assumptions of Theorem 8.8.3. Let T be an efficient estimator of its mean $m(\theta)$. Assume that $m'(\theta)$ is never 0. Then the asymptotic distribution of

$$\frac{[nI(\theta)]^{1/2}}{m'(\theta)}\,[T - m(\theta)]$$

is the standard normal distribution.

Proof Consider first the random variable $\lambda_n'(\boldsymbol{X}|\theta)$. Since $\lambda_n(\boldsymbol{X}|\theta) = \sum_{i=1}^n \lambda(X_i|\theta)$, then

$$\lambda_n'(\boldsymbol{X}|\theta) = \sum_{i=1}^n \lambda'(X_i|\theta).$$

Furthermore, since the n random variables $X_1, \ldots, X_n$ are i.i.d., the n random variables $\lambda'(X_1|\theta), \ldots, \lambda'(X_n|\theta)$ will also be i.i.d. We know from Eqs. (8.8.7) and (8.8.4) that the mean of each of these variables is 0, and the variance of each is $I(\theta)$. Hence, it follows from the central limit theorem of Lindeberg and Lévy (Theorem 6.3.1) that the asymptotic distribution of the random variable $\lambda_n'(\boldsymbol{X}|\theta)/[nI(\theta)]^{1/2}$ is the standard normal distribution.

Since T is an efficient estimator of $m(\theta)$, we have

$$E_\theta(T) = m(\theta) \quad \text{and} \quad \text{Var}_\theta(T) = \frac{[m'(\theta)]^2}{nI(\theta)}. \tag{8.8.22}$$

Furthermore, there must exist functions $u(\theta)$ and $v(\theta)$ that satisfy Eq. (8.8.15). Because the random variable $\lambda_n'(\boldsymbol{X}|\theta)$ has mean 0 and variance $nI(\theta)$, it follows from Eq. (8.8.15) that

$$E_\theta(T) = v(\theta) \quad \text{and} \quad \text{Var}_\theta(T) = [u(\theta)]^2 nI(\theta).$$

When these values for the mean and the variance of T are compared with the values in Eq. (8.8.22), we find that $v(\theta) = m(\theta)$ and $|u(\theta)| = |m'(\theta)|/[nI(\theta)]$. To be specific, we shall assume that $u(\theta) = m'(\theta)/[nI(\theta)]$, although the same conclusions would be obtained if $u(\theta) = -m'(\theta)/[nI(\theta)]$.

Next, substitute the values $u(\theta) = m'(\theta)/[nI(\theta)]$ and $v(\theta) = m(\theta)$ into Eq. (8.8.15) to obtain

$$T = \frac{m'(\theta)}{nI(\theta)}\lambda_n'(\boldsymbol{X}|\theta) + m(\theta).$$

Rearranging this equation slightly yields

$$\frac{[nI(\theta)]^{1/2}}{m'(\theta)}\,[T - m(\theta)] = \frac{\lambda_n'(\boldsymbol{X}|\theta)}{[nI(\theta)]^{1/2}}. \tag{8.8.23}$$

We have already shown that the asymptotic distribution of the random variable on the right side of Eq. (8.8.23) is the standard normal distribution. Therefore, the asymptotic distribution of the random variable on the left side of Eq. (8.8.23) is also the standard normal distribution. ■

Asymptotic Distribution of an M.L.E It follows from Theorem 8.8.4 that if the M.L.E. $\hat{\theta}_n$ is an efficient estimator of θ for each value of n, then the asymptotic distribution of $[nI(\theta)]^{1/2}(\hat{\theta}_n - \theta)$ is the standard normal distribution. However, it can be shown that even in an arbitrary problem in which $\hat{\theta}_n$ is not an efficient estimator, $[nI(\theta)]^{1/2}(\hat{\theta}_n - \theta)$ has this same asymptotic distribution under certain conditions. Without presenting all the required conditions in full detail, we can state the following result. The proof of this result can be found in Schervish (1995, chapter 7).

Theorem 8.8.5

Asymptotic Distribution of M.L.E. Suppose that in an arbitrary problem the M.L.E. $\hat{\theta}_n$ is determined by solving the equation $\lambda_n'(\boldsymbol{x}|\theta) = 0$, and in addition both the second and third derivatives $\lambda_n''(\boldsymbol{x}|\theta)$ and $\lambda_n'''(\boldsymbol{x}|\theta)$ exist and satisfy certain regularity conditions. Then the asymptotic distribution of $[nI(\theta)]^{1/2}(\hat{\theta}_n - \theta)$ is the standard normal distribution. ■

In practical terms, Theorem 8.8.5 states that in most problems in which the sample size n is large, and the M.L.E. $\hat{\theta}_n$ is found by differentiating the likelihood function $f_n(\boldsymbol{x}|\theta)$ or its logarithm, the distribution of $[nI(\theta)]^{1/2}(\hat{\theta}_n - \theta)$ will be approximately the standard normal distribution. Equivalently, the distribution of $\hat{\theta}_n$ will be approximately the normal distribution with mean θ and variance $1/[nI(\theta)]$. Under these conditions, it is said that $\hat{\theta}_n$ is an *asymptotically efficient estimator*.

Example 8.8.11

Estimating the Standard Deviation of a Normal Distribution. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with known mean 0 and unknown standard deviation σ ($\sigma > 0$). It can be shown that the M.L.E. of σ is

$$\hat{\sigma} = \left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right]^{1/2}.$$

Also, it can be shown (see Exercise 4) that the Fisher information in a single observation is $I(\sigma) = 2/\sigma^2$. Therefore, if the sample size n is large, the distribution of $\hat{\sigma}$ will be approximately the normal distribution with mean σ and variance $\sigma^2/(2n)$. ◀

For cases in which it is difficult to compute the M.L.E., there is a result similar to Theorem 8.8.5. The proof of Theorem 8.8.6 can also be found as a special case of theorem 7.75 in Schervish (1995).

Theorem 8.8.6

Efficient Estimation. Assume the same smoothness conditions on the likelihood function as in Theorem 8.8.5. Assume that $\tilde{\theta}_n$ is a sequence of estimators of θ such that $\sqrt{n}(\tilde{\theta}_n - \theta)$ converges in distribution to some distribution (it doesn't matter what distribution). Use $\tilde{\theta}_n$ as the starting value, and perform one step of Newton's method (Definition 7.6.2) toward finding the M.L.E. of θ. Let the result of this one step be called θ_n^*. Then the asymptotic distribution of $[nI(\theta)]^{1/2}(\theta_n^* - \theta)$ is the standard normal distribution. ■

A typical choice of $\tilde{\theta}_n$ in Theorem 8.8.6 is a method of moments estimator (Definition 7.6.3). Example 7.6.6 illustrates such an application of Theorem 8.8.6 when sampling from a gamma distribution.

The Bayesian Point of View Another general property of the M.L.E. $\hat{\theta}_n$ pertains to making inferences about a parameter θ from the Bayesian point of view. Suppose that the prior distribution of θ is represented by a positive and differentiable p.d.f. over the interval Ω, and the sample size n is large. Then under conditions similar to the regularity conditions that are needed to assure the asymptotic normality of the distribution of $\hat{\theta}_n$, it can be shown that the posterior distribution of θ, after the values of $X_1, \ldots, X_n$ have been observed, will be approximately the normal distribution with mean $\hat{\theta}_n$ and variance $1/[nI(\hat{\theta}_n)]$.

Example 8.8.12

The Posterior Distribution of the Standard Deviation. Suppose again that $X_1, \ldots, X_n$ form a random sample from the normal distribution with known mean 0 and unknown standard deviation σ. Suppose also that the prior p.d.f. of σ is a positive and differentiable function for $\sigma > 0$, and the sample size n is large. Since $I(\sigma) = 2/\sigma^2$, it follows that the posterior distribution of σ will be approximately the normal distribution with mean $\hat{\sigma}$ and variance $\hat{\sigma}^2/(2n)$, where $\hat{\sigma}$ is the M.L.E. of σ calculated from the observed values in the sample. Figure 8.9 illustrates this approximation based on a

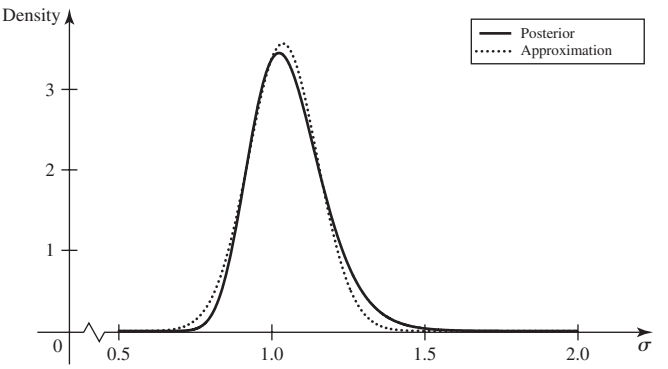

Figure 8.9 Posterior p.d.f. of σ and approximation based on Fisher information in Example 8.8.12.

sample of $n = 40$ i.i.d. simulated normal random variables with mean 0 and variance 1. In this sample, the M.L.E. was $\hat{\sigma} = 1.061$. Figure 8.9 shows the actual posterior p.d.f. based on an improper prior with "p.d.f." $1/\sigma$ together with the approximate normal posterior p.d.f. with mean 1.061 and variance $1.061^2/80 = 0.0141$. ◀

Fisher Information for Multiple Parameters

Example 8.8.13 Sample from a Normal Distribution. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the normal distribution with mean μ and variance σ^2. Is there an analog to Fisher information for the vector parameter $\theta = (\mu, \sigma^2)$? ◀

In the spirit of Definition 8.8.1 and Theorem 8.8.1, we define Fisher information in terms of derivatives of the logarithm of the likelihood function. We shall define the Fisher information in a random sample of size n with the understanding that the Fisher information in a single random variable corresponds to a sample size of $n = 1$.

Definition 8.8.4 Fisher Information for a Vector Parameter. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ form a random sample from a distribution for which the p.d.f. is $f(x|\theta)$, where the value of the parameter $\theta = (\theta_1, \ldots, \theta_k)$ must lie in an open subset Ω of a k-dimensional real space. Let $f_n(\boldsymbol{x}|\theta)$ denote the joint p.d.f. or joint p.f. of $\boldsymbol{X}$. Define

$$\lambda_n(\boldsymbol{x}|\theta) = \log f_n(x|\theta).$$

Assume that the set of $\boldsymbol{x}$ such that $f_n(\boldsymbol{x}|\theta) > 0$ is the same for all θ and that $\log f_n(\boldsymbol{x}|\theta)$ is twice differentiable with respect to θ. The *Fisher information matrix* $I_n(\theta)$ in the random sample $\boldsymbol{X}$ is defined as the $k \times k$ matrix with (i, j) element equal to

$$I_{n,i,j}(\theta) = \mathrm{Cov}_\theta \left[\frac{\partial}{\partial \theta_i} \lambda_n'(\boldsymbol{X}|\theta), \frac{\partial}{\partial \theta_j} \lambda_n'(\boldsymbol{X}|\theta) \right].$$

Example 8.8.14 Sample from a Normal Distribution. In Example 8.8.13, let $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. As in Eq. (7.5.3), we obtain

$$\lambda_n(\boldsymbol{X}|\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^{n}(X_i - \theta_1)^2.$$

The first partial derivatives are

$$\frac{\partial}{\partial\theta_1}\lambda_n(\boldsymbol{x}|\theta) = \frac{1}{\theta_2}\sum_{i=1}^{n}(X_i - \theta_1), \tag{8.8.24}$$

$$\frac{\partial}{\partial\theta_2}\lambda_n(\boldsymbol{x}|\theta) = \frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(X_i - \theta_1)^2. \tag{8.8.25}$$

Since the means of the two random variables above are both 0, their covariances are the means of the products. The distribution of $\sum_{i=1}^{n}(X_i - \theta_1)$ is the normal distribution with mean 0 and variance $n\theta_2$. The distribution of $\sum_{i=1}^{n}(X_i - \theta_1)^2/\theta_2$ is the χ^2 distribution with n degrees of freedom. So the variance of (8.8.24) is n/θ_2, and the variance of (8.8.25) is $2n/\theta_2^2$. The mean of the product of (8.8.24) and (8.8.25) is 0 because the third central moment of a normal distribution is 0. This makes

$$I_n(\theta) = \begin{pmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{\theta_2^2} \end{pmatrix}. \qquad \blacktriangleleft$$

The results for one-dimensional parameters all have versions for k-dimensional parameters. For example, in Eq. (8.8.3), $\lambda''(X|\theta)$ is replaced by the $k \times k$ matrix of second partial derivatives. In the Cramér-Rao inequality, we need the inverse of the matrix $I_n(\theta)$, and $m'(\theta)$ must be replaced by the vector of partial derivatives. Specifically, if T is a statistic with finite variance and mean $m(\theta)$, then

$$\mathrm{Var}_\theta(T) \geq \left(\frac{\partial}{\partial\theta_1}m(\theta), \ldots, \frac{\partial}{\partial\theta_k}m(\theta)\right) I_n(\theta)^{-1} \begin{pmatrix} \frac{\partial}{\partial\theta_1}m(\theta) \\ \vdots \\ \frac{\partial}{\partial\theta_k}m(\theta) \end{pmatrix}. \tag{8.8.26}$$

Also, the inequality in (8.8.26) is equality if and only if T is a linear function of the vector

$$\left(\frac{\partial}{\partial\theta_1}\lambda_n(\boldsymbol{x}|\theta), \ldots, \frac{\partial}{\partial\theta_k}\lambda_n(\boldsymbol{x}|\theta)\right). \tag{8.8.27}$$

Example 8.8.15 Sample from a Normal Distribution. In Example 8.8.14, the coordinates of the vector in (8.8.27) are linear functions of the two random variables $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$. So, the only statistics whose variances equal the lower bound in (8.8.26) are of the form $T = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 + c$. The mean of such a statistic T is

$$E_\theta(T) = an\theta_1 + bn(\theta_2 + \theta_1^2) + c. \tag{8.8.28}$$

In particular, it is impossible to obtain θ_2 as a special case of (8.8.28). There is no efficient unbiased estimator of $\theta_2 = \sigma^2$. It can be proven that $(\sigma')^2$, which was defined in Eq. (8.4.3), is an unbiased estimator that has minimum variance among all unbiased estimators. The proof of this fact is beyond the scope of this text. The variance of $(\sigma')^2$ is $2\theta_2^2/(n-1)$, while the Cramér-Rao lower bound is $2\theta_2^2/n$. ◀

Example 8.8.16 Multinomial Distributions. Let $\boldsymbol{X} = (X_1, \ldots, X_k)$ have the multinomial distribution with parameters n and $\boldsymbol{p} = (p_1, \ldots, p_k)$ as defined in Definition 5.9.1. Finding the Fisher information in this example involves a subtle point. The parameter vector $\boldsymbol{p}$ takes values in the set

$$\{\boldsymbol{p} : p_1 + \cdots + p_k = 1, \text{ all } p_i \geq 0\}.$$

No subset of this set is open. Hence, no matter what set we choose for the parameter space, Definition 8.8.4 does not apply to this parameter. However, there is an

equivalent paramter $\boldsymbol{p}^* = (p_1, \ldots, p_{k-1})$ that takes values in the set

$$\{\boldsymbol{p}^* : p_1 + \cdots + p_{k-1} \le 1, \text{ all } p_i \ge 0\},$$

which has nonempty interior. With this version of the parameter, and assuming that the parameter space is the interior of the set above, it is straightforward to calculate the Fisher information, as in Exercise 20. ◀

Summary

Fisher information attempts to measure the amount of information about a parameter that a random variable or sample contains. Fisher information from independent random variables adds together to form the Fisher information in the sample. The information inequality (Cramér-Rao lower bound) provides lower bounds on the variances of all estimators. An estimator is efficient if its variance equals the lower bound. The asymptotic distribution of a maximum likelihood estimator of θ is (under regularity conditions) normal with mean θ and variance equal to 1 over the Fisher information in the sample. Also, for large sample sizes, the posterior distribution of θ is approximately normal with mean equal to the M.L.E. and variance equal to 1 over the Fisher information in the sample evaluated at the M.L.E.

Exercises

1. Suppose that a random variable X has a normal distribution for which the mean μ is unknown ($-\infty < \mu < \infty$) and the variance σ^2 is known. Let $f(x|\mu)$ denote the p.d.f. of X, and let $f'(x|\mu)$ and $f''(x|\mu)$ denote the first and second partial derivatives with respect to μ. Show that

$$\int_{-\infty}^{\infty} f'(x|\mu)\, dx = 0 \quad \text{and} \quad \int_{-\infty}^{\infty} f''(x|\mu)\, dx = 0.$$

2. Suppose that X has the geometric distribution with parameter p. (See Sec. 5.5.) Find the Fisher information $I(p)$ in X.

3. Suppose that a random variable X has the Poisson distribution with unknown mean $\theta > 0$. Find the Fisher information $I(\theta)$ in X.

4. Suppose that a random variable has the normal distribution with mean 0 and unknown standard deviation $\sigma > 0$. Find the Fisher information $I(\sigma)$ in X.

5. Suppose that a random variable X has the normal distribution with mean 0 and unknown variance $\sigma^2 > 0$. Find the Fisher information $I(\sigma^2)$ in X. Note that in this exercise the variance σ^2 is regarded as the parameter, whereas in Exercise 4 the standard deviation σ is regarded as the parameter.

6. Suppose that X is a random variable for which the p.d.f. or the p.f. is $f(x|\theta)$, where the value of the parameter θ is unknown but must lie in an open interval Ω. Let $I_0(\theta)$ denote the Fisher information in X. Suppose now that the parameter θ is replaced by a new parameter μ, where $\theta = \psi(\mu)$, and ψ is a differentiable function. Let $I_1(\mu)$ denote the Fisher information in X when the parameter is regarded as μ. Show that

$$I_1(\mu) = [\psi'(\mu)]^2 I_0[\psi(\mu)].$$

7. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p. Show that $\overline{X}_n$ is an efficient estimator of p.

8. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and known variance $\sigma^2 > 0$. Show that $\overline{X}_n$ is an efficient estimator of μ.

9. Suppose that a single observation X is taken from the normal distribution with mean 0 and unknown standard deviation $\sigma > 0$. Find an unbiased estimator of σ, determine its variance, and show that this variance is greater than $1/I(\sigma)$ for every value of $\sigma > 0$. Note that the value of $I(\sigma)$ was found in Exercise 4.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean 0 and unknown standard deviation $\sigma > 0$. Find the lower bound specified by the information inequality for the variance of any unbiased estimator of $\log \sigma$.

11. Suppose that $X_1, \ldots, X_n$ form a random sample from an exponential family for which the p.d.f. or the p.f. $f(x|\theta)$ is as specified in Exercise 23 of Sec. 7.3. Suppose also that the unknown value of θ must belong to an open interval Ω of the real line. Show that the estimator $T = \sum_{i=1}^{n} d(X_i)$ is an efficient estimator. *Hint:* Show that T can be represented in the form given in Eq. (8.8.15).

12. Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the mean is known and the variance is unknown. Construct an efficient estimator that is not identically equal to a constant, and determine the expectation and the variance of this estimator.

13. Determine what is wrong with the following argument: Suppose that the random variable X has the uniform distribution on the interval $[0, \theta]$, where the value of θ is unknown ($\theta > 0$). Then $f(x|\theta) = 1/\theta, \lambda(x|\theta) = -\log\theta$ and $\lambda'(x|\theta) = -(1/\theta)$. Therefore,

$$I(\theta) = E_\theta\{[\lambda'(X|\theta)]^2\} = \frac{1}{\theta^2}.$$

Since $2X$ is an unbiased estimator of θ, the information inequality states that

$$\text{Var}(2X) \geq \frac{1}{I(\theta)} = \theta^2.$$

But

$$\text{Var}(2X) = 4\,\text{Var}(X) = 4 \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3} < \theta^2.$$

Hence, the information inequality is not correct.

14. Suppose that $X_1, \ldots, X_n$ form a random sample from the gamma distribution with parameters α and β, where α is unknown and β is known. Show that if n is large, the distribution of the M.L.E. of α will be approximately a normal distribution with mean α and variance

$$\frac{[\Gamma(\alpha)]^2}{n\{\Gamma(\alpha)\Gamma''(\alpha) - [\Gamma'(\alpha)]^2\}}.$$

15. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with uknown mean μ and known variance σ^2, and the prior p.d.f. of μ is a positive and differentiable function over the entire real line. Show that if n is large, the posterior distribution of μ given that $X_i = x_i$ $(i = 1, \ldots, n)$ will be approximately a normal distribution with mean $\overline{x}_n$ and variance σ^2/n.

16. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p, and the prior p.d.f. of p is a positive and differentiable function over the interval $0 < p < 1$. Suppose, furthermore, that n is large, the observed values of $X_1, \ldots, X_n$ are $x_1, \ldots, x_n$, and $0 < \overline{x}_n < 1$. Show that the posterior distribution of p will be approximately a normal distribution with mean $\overline{x}_n$ and variance $\overline{x}_n(1 - \overline{x}_n)/n$.

17. Let X have the binomial distribution with parameters n and p. Assume that n is known. Show that the Fisher information in X is $I(p) = n/[p(1-p)]$.

18. Let X have the negative binomial distribution with parameters r and p. Assume that r is known. Show that the Fisher information in X is $I(p) = r/[p^2(1-p)]$.

19. Let X have the gamma distribution with parameters n and θ with θ unknown. Show that the Fisher information in X is $I(\theta) = n/\theta^2$.

20. Find the Fisher information matrix about $\boldsymbol{p}^*$ in Example 8.8.16.

8.9 Supplementary Exercises

1. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with known mean 0 and unknown variance σ^2. Show that $\sum_{i=1}^n X_i^2/n$ is the unbiased estimator of σ^2 that has the smallest possible variance for all possible values of σ^2.

2. Prove that if X has the t distribution with one degree of freedom, then $1/X$ also has the t distribution with one degree of freedom.

3. Suppose that U and V are independent random variables, and that each has the standard normal distribution. Show that U/V, $U/|V|$, and $|U|/V$ each has the t distribution with one degree of freedom.

4. Suppose that X_1 and X_2 are independent random variables, and that each has the normal distribution with mean 0 and variance σ^2. Show that $(X_1 + X_2)/(X_1 - X_2)$ has the t distribution with one degree of freedom.

5. Suppose that $X_1, \ldots, X_n$ form a random sample from the exponential distribution with parameter β. Show that $2\beta\sum_{i=1}^n X_i$ has the χ^2 distribution with $2n$ degrees of freedom.

6. Suppose that $X_1, \ldots, X_n$ form a random sample from an unknown probability distribution P on the real line. Let A be a given subset of the real line, and let $\theta = P(A)$. Construct an unbiased estimator of θ, and specify its variance.

7. Suppose that $X_1, \ldots, X_m$ form a random sample from the normal distribution with mean μ_1 and variance σ^2, and $Y_1, \ldots, Y_n$ form an independent random sample from the normal distribution with mean μ_2 and variance $2\sigma^2$. Let $S_X^2 = \sum_{i=1}^m (X_i - \overline{X}_m)^2$ and $S_Y^2 = \sum_{i=1}^n (Y_i - \overline{Y}_n)^2$.

a. For what pairs of values of α and β is $\alpha S_X^2 + \beta S_Y^2$ an unbiased estimator of σ^2?

b. Determine the values of α and β for which $\alpha S_X^2 + \beta S_Y^2$ will be an unbiased estimator with minimum variance.

8. Suppose that $X_1, \ldots, X_{n+1}$ form a random sample from a normal distribution, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $T_n = \left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right]^{1/2}$. Determine the value of a constant k such that the random variable $k(X_{n+1} - \overline{X}_n)/T_n$ will have a t distribution.

9. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2, and Y is an independent random variable having the normal distribution with mean 0 and variance $4\sigma^2$. Determine a function of $X_1, \ldots, X_n$ and Y that does not involve μ or σ^2 but has the t distribution with $n - 1$ degrees of freedom.

10. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with mean μ and variance σ^2, where both μ and σ^2 are unknown. A confidence interval for μ is to be constructed with confidence coefficient 0.90. Determine the smallest value of n such that the expected squared length of this interval will be less than $\sigma^2/2$.

11. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean μ and unknown variance σ^2. Construct a lower confidence limit $L(X_1, \ldots, X_n)$ for μ such that

$$\Pr[\mu > L(X_1, \ldots, X_n)] = 0.99.$$

12. Consider again the conditions of Exercise 11. Construct an upper confidence limit $U(X_1, \ldots, X_n)$ for σ^2 such that

$$\Pr[\sigma^2 < U(X_1, \ldots, X_n)] = 0.99.$$

13. Suppose that $X_1, \ldots, X_n$ form a random sample from the normal distribution with unknown mean θ and known variance σ^2. Suppose also that the prior distribution of θ is normal with mean μ and variance ν^2.

a. Determine the shortest interval I such that $\Pr(\theta \in I | x_1, \ldots, x_n) = 0.95$, where the probability is calculated with respect to the posterior distribution of θ, as indicated.

b. Show that as $\nu^2 \to \infty$, the interval I converges to an interval I^* that is a confidence interval for θ with confidence coefficient 0.95.

14. Suppose that $X_1, \ldots, X_n$ form a random sample from the Poisson distribution with unknown mean θ, and let $Y = \sum_{i=1}^{n} X_i$.

a. Determine the value of a constant c such that the estimator e^{-cY} is an unbiased estimator of $e^{-\theta}$.

b. Use the information inequality to obtain a lower bound for the variance of the unbiased estimator found in part (a).

15. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution for which the p.d.f. is as follows:

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

where the value of θ is unknown ($\theta > 0$). Determine the asymptotic distribution of the M.L.E. of θ. (*Note*: The M.L.E. was found in Exercise 9 of Sec. 7.5.)

16. Suppose that a random variable X has the exponential distribution with mean θ, which is unknown ($\theta > 0$). Find the Fisher information $I(\theta)$ in X.

17. Suppose that $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with unknown parameter p. Show that the variance of every unbiased estimator of $(1 - p)^2$ must be at least $4p(1 - p)^3/n$.

18. Suppose that $X_1, \ldots, X_n$ form a random sample from the exponential distribution with unknown parameter β. Construct an efficient estimator that is not identically equal to a constant, and determine the expectation and the variance of this estimator.

19. Suppose that $X_1, \ldots, X_n$ form a random sample from the exponential distribution with unknown parameter β. Show that if n is large, the distribution of the M.L.E. of β will be approximately a normal distribution with mean β and variance β^2/n.

20. Consider again the conditions of Exercise 19, and let $\hat{\beta}_n$ denote the M.L.E. of β.

a. Use the delta method to determine the asymptotic distribution of $1/\hat{\beta}_n$.

b. Show that $1/\hat{\beta}_n = \overline{X}_n$, and use the central limit theorem to determine the asymptotic distribution of $1/\hat{\beta}_n$.

21. Let $X_1, \ldots, X_n$ be a random sample from the Poisson distribution with mean θ. Let $Y = \sum_{i=1}^{n} X_i$.

a. Prove that there is no unbiased estimator of $1/\theta$. (*Hint:* Write the equation that is equivalent to $E_\theta(r(X)) = 1/\theta$. Simplify it, and then use what you know from calculus of infinite series to show that no function r can satisfy the equation.)

b. Suppose that we wish to estimate $1/\theta$. Consider $r(Y) = n/(Y + 1)$ as an estimator of θ. Find the bias of $r(Y)$, and show that the bias goes to 0 as $n \to \infty$.

c. Use the delta method to find the asymptotic (as $n \to \infty$) distribution of $n/(Y + 1)$.

22. Let $X_1, \ldots, X_n$ be conditionally i.i.d. with the uniform distribution on the interval $[0, \theta]$. Let $Y_n = \max\{X_1, \ldots, X_n\}$.

a. Find the p.d.f. and the quantile function of Y_n/θ.

b. Y_n is often used as an estimator of θ even though it has bias. Compute the bias of Y_n as an estimator of θ.

c. Prove that Y_n/θ is a pivotal.

d. Find a confidence interval for θ with coefficient γ.