Chapter 6

Likelihood Inference

CHAPTER OUTLINE

Section 1 The Likelihood Function
Section 2 Maximum Likelihood Estimation
Section 3 Inferences Based on the MLE
Section 4 Distribution-Free Methods
Section 5 Large Sample Behavior of the MLE (Advanced)

In this chapter, we discuss some of the most basic approaches to inference. In essence, we want our inferences to depend only on the model $\{P_\theta : \theta \in \Omega\}$ and the data s. These methods are very minimal in the sense that they require few assumptions. While successful for certain problems, it seems that the additional structure of Chapter 7 or Chapter 8 is necessary in more involved situations.

The likelihood function is one of the most basic concepts in statistical inference. Entire theories of inference have been constructed based on it. We discuss likelihood methods in Sections 6.1, 6.2, 6.3, and 6.5. In Section 6.4, we introduce some distribution-free methods of inference. These are not really examples of likelihood methods, but they follow the same basic idea of having the inferences depend on as few assumptions as possible.

6.1 The Likelihood Function

Likelihood inferences are based only on the data s and the model $\{P_\theta : \theta \in \Omega\}$ — the set of possible probability measures for the system under investigation. From these ingredients we obtain the basic entity of likelihood inference, namely, the likelihood function.

To motivate the definition of the likelihood function, suppose we have a statistical model in which each P_θ is discrete, given by probability function f_θ. Having observed s, consider the function $L(\cdot \,|\, s)$ defined on the parameter space Ω and taking values in R^1, given by $L(\theta \,|\, s) = f_\theta(s)$. We refer to $L(\cdot \,|\, s)$ as the *likelihood function* determined by the model and the data. The value $L(\theta \,|\, s)$ is called the *likelihood* of θ.

Note that for the likelihood function, we are fixing the data and varying the value of the parameter.

We see that $f_\theta(s)$ is just the probability of obtaining the data s when the true value of the parameter is θ. This imposes a belief ordering on Ω, namely, we believe in θ_1 as the true value of θ over θ_2 whenever $f_{\theta_1}(s) > f_{\theta_2}(s)$. This is because the inequality says that the data are more likely under θ_1 than θ_2. We are indifferent between θ_1 and θ_2 whenever $f_{\theta_1}(s) = f_{\theta_2}(s)$. Likelihood inference about θ is based on this ordering.

It is important to remember the correct interpretation of $L(\theta \,|\, s)$. The value $L(\theta \,|\, s)$ is the probability of s given that θ is the true value — it is *not* the probability of θ given that we have observed s. Also, it is possible that the value of $L(\theta \,|\, s)$ is very small for every value of θ. So it is not the actual value of the likelihood that is telling us how much support to give to a particular θ, but rather its value relative to the likelihoods of other possible parameter values.

EXAMPLE 6.1.1

Suppose $S = \{1, 2, \ldots\}$ and that the statistical model is $\{P_\theta : \theta \in \{1, 2\}\}$, where P_1 is the uniform distribution on the integers $\left\{1, \ldots, 10^3\right\}$ and P_2 is the uniform distribution on $\left\{1, \ldots, 10^6\right\}$. Further suppose that we observe $s = 10$. Then $L(1 \,|\, 10) = 1/10^3$ and $L(2 \,|\, 10) = 1/10^6$. Both values are quite small, but note that the likelihood supports $\theta = 1$ a thousand times more than it supports $\theta = 2$. ▮

Accordingly, we are only interested in *likelihood ratios* $L(\theta_1 \,|\, s)/L(\theta_2 \,|\, s)$ for $\theta_1, \theta_2 \in \Omega$ when it comes to determining inferences for θ based on the likelihood function. This implies that any function that is a positive multiple of $L(\cdot \,|\, s)$, i.e., $L^*(\cdot \,|\, s) = cL(\cdot \,|\, s)$ for some fixed $c > 0$, can serve equally well as a likelihood function. We call two likelihoods equivalent if they are proportional in this way. In general, we refer to any positive multiple of $L(\cdot \,|\, s)$ as a likelihood function.

EXAMPLE 6.1.2

Suppose that a coin is tossed $n = 10$ times and that $s = 4$ heads are observed. With no knowledge whatsoever concerning the probability of getting a head on a single toss, the appropriate statistical model for the data is the Binomial$(10, \theta)$ model with $\theta \in \Omega = [0, 1]$. The likelihood function is given by

$$L(\theta \,|\, 4) = \binom{10}{4} \theta^4 (1-\theta)^6, \tag{6.1.1}$$

which is plotted in Figure 6.1.1.

This likelihood peaks at $\theta = 0.4$ and takes the value 0.2508 there. We will examine uses of the likelihood to estimate the unknown θ and assess the accuracy of the estimate. Roughly speaking, however, this is based on where the likelihood takes its maximum and how much spread there is in the likelihood about its peak. ▮

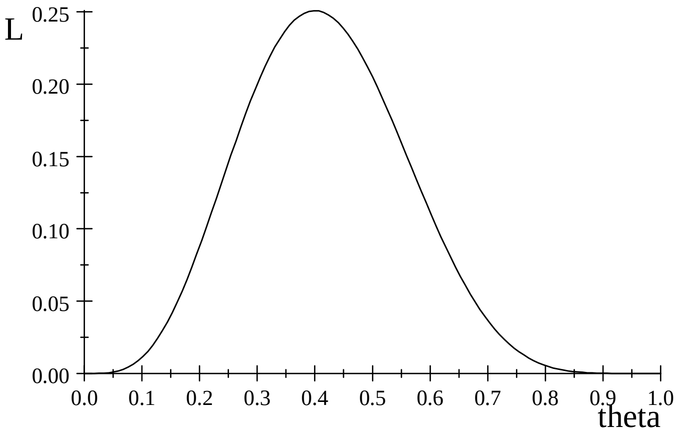

Figure 6.1.1: Likelihood function from the Binomial$(10, \theta)$ model when $s = 4$ is observed.

There is a range of approaches to obtaining inferences via the likelihood function. At one extreme is the likelihood principle.

> *Likelihood Principle*: If two model and data combinations yield equivalent likelihood functions, then inferences about the unknown parameter must be the same.

This principle dictates that anything we want to say about the unknown value of θ must be based only on $L(\cdot \,|\, s)$. For many statisticians, this is viewed as a very severe proscription. Consider the following example.

EXAMPLE 6.1.3

Suppose a coin is tossed in independent tosses until four heads are obtained and the number of tails observed until the fourth head is $s = 6$. Then s is distributed Negative-Binomial$(4, \theta)$, and the likelihood specified by the observed data is

$$L(\theta \,|\, 6) = \binom{9}{6}\theta^4(1-\theta)^6.$$

Note that this likelihood function is a positive multiple of (6.1.1).

So the likelihood principle asserts that these two model and data combinations must yield the same inferences about the unknown θ. In effect, the likelihood principle says we must ignore the fact that the data were obtained in entirely different ways. If, however, we take into account additional model features beyond the likelihood function, then it turns out that we can derive different inferences for the two situations. In particular, assessing a hypothesized value $\theta = \theta_0$ can be carried out in different ways when the sampling method is taken into account. Many statisticians believe this additional information should be used when deriving inferences. ∎

As an example of an inference derived from a likelihood function, consider a set of the form

$$C(s) = \{\theta : L(\theta \,|\, s) \geq c\},$$

for some $c \geq 0$. The set $C(s)$ is referred to as a *likelihood region*. It contains all those θ values for which their likelihood is at least c. A likelihood region, for some c, seems like a sensible set to quote as possibly containing the true value of θ. For, if $\theta^* \notin C(s)$, then $L(\theta^* \,|\, s) < L(\theta \,|\, s)$ for every $\theta \in C(s)$ and so is not as well-supported by the observed data as any value in $C(s)$. The size of $C(s)$ can then be taken as a measure of how uncertain we are about the true value of θ.

We are left with the problem, however, of choosing a suitable value for c and, as Example 6.1.1 seems to indicate, the likelihood itself does not suggest a natural way to do this. In Section 6.3.2, we will discuss a method for choosing c that is based upon additional model properties beyond the likelihood function.

So far in this section, we have assumed that our statistical models are comprised of discrete distributions. The definition of the likelihood is quite natural, as $L(\theta \,|\, s)$ is simply the probability of s occurring when θ is the true value. This interpretation is clearly not directly available, however, when we have a continuous model because every data point has probability 0 of occurring. Imagine, however, that $f_{\theta_1}(s) > f_{\theta_2}(s)$ and that $s \in R^1$. Then, assuming the continuity of every f_θ at s, we have

$$P_{\theta_1}(V) = \int_a^b f_{\theta_1}(s)\,dx > P_{\theta_2}(V) = \int_a^b f_{\theta_2}(s)\,dx$$

for every interval $V = (a, b)$ containing s that is small enough. We interpret this to mean that the probability of s occurring when θ_1 is true is greater than the probability of s occurring when θ_2 is true. So the data s support θ_1 more than θ_2. A similar interpretation applies when $s \in R^n$ for $n > 1$ and V is a region containing s.

Therefore, in the continuous case, we again define the likelihood function by $L(\theta \,|\, s) = f_\theta(s)$ and interpret the ordering this imposes on the values of θ exactly as we do in the discrete case.[1] Again, two likelihoods will be considered equivalent if one is a positive multiple of the other.

Now consider a very important example.

EXAMPLE 6.1.4 *Location Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is an observed independently and identically distributed (i.i.d.) sample from an $N(\theta, \sigma_0^2)$ distribution where $\theta \in \Omega = R^1$ is unknown and $\sigma_0^2 > 0$ is known. The likelihood function is given by

$$L(\theta \,|\, x_1, \ldots, x_n) = \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \left(2\pi\sigma_0^2\right)^{-1/2} \exp\left(-\frac{1}{2\sigma_0^2}(x_i - \theta)^2\right)$$

[1] Note, however, that whenever we have a situation in which $f_{\theta_1}(s) = f_{\theta_2}(s)$, we could still have $P_{\theta_1}(V) > P_{\theta_2}(V)$ for every V containing s, and small enough. This implies that θ_1 is supported more than θ_2 rather than these two values having equal support, as implied by the likelihood. This phenomenon does not occur in the examples we discuss, so we will ignore it here.

and clearly this simplifies to

$$\begin{aligned} L(\theta \,|\, x_1, \ldots, x_n) &= \left(2\pi\sigma_0^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \\ &= \left(2\pi\sigma_0^2\right)^{-n/2} \exp\left(-\frac{n}{2\sigma_0^2}(\bar{x}-\theta)^2\right)\exp\left(-\frac{n-1}{2\sigma_0^2}s^2\right). \end{aligned}$$

An equivalent, simpler version of the likelihood function is then given by

$$L(\theta \,|\, x_1, \ldots, x_n) = \exp\left(-\frac{n}{2\sigma_0^2}(\bar{x}-\theta)^2\right),$$

and we will use this version.

For example, suppose $n = 25$, $\sigma_0^2 = 1$, and we observe $\bar{x} = 3.3$. This function is plotted in Figure 6.1.2.

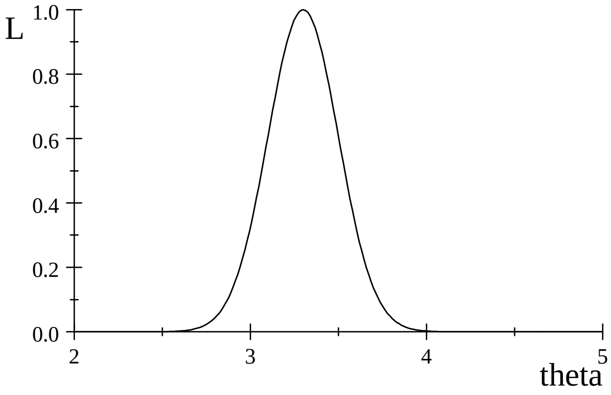

Figure 6.1.2: Likelihood from a location normal model based on a sample of 25 with $\bar{x} = 3.3$.

The likelihood peaks at $\theta = \bar{x} = 3.3$, and the plotted function takes the value 1 there. The likelihood interval

$$C(x) = \{\theta : L(\theta \,|\, x_1, \ldots, x_n) \geq 0.5\} = (3.0645, 3.53548)$$

contains all those θ values whose likelihood is at least 0.5 of the value of the likelihood at its peak.

The location normal model is impractical for many applications, as it assumes that the variance is known, while the mean is unknown. For example, if we are interested in the distribution of heights in a population, it seems unlikely that we will know the population variance but not know the population mean. Still, it is an important statistical model, as it is a context where inference methods can be developed fairly easily.

The methodology developed for this situation is often used as a paradigm for inference methods in much more complicated models. ▮

The parameter θ need not be one-dimensional. The interpretation of the likelihood is still the same, but it is not possible to plot it — at least not when the dimension of θ is greater than 2.

EXAMPLE 6.1.5 *Multinomial Models*
In Example 2.8.5, we introduced multinomial distributions. These arise in applications when we have a categorical response variable s that can take a finite number k of values, say, $\{1, \ldots, k\}$, and $P(s = i) = \theta_i$.

Suppose, then, that $k = 3$ and we do not know the value of $(\theta_1, \theta_2, \theta_3)$. In this case, the parameter space is given by

$$\Omega = \{(\theta_1, \theta_2, \theta_3) : \theta_i \geq 0, \text{ for } i = 1, 2, 3, \text{ and } \theta_1 + \theta_2 + \theta_3 = 1\}.$$

Notice that it is really only two-dimensional, because as soon as we know the value of any two of the θ_i's, say, θ_1 and θ_2, we immediately know the value of the remaining parameter, as $\theta_3 = 1 - \theta_1 - \theta_2$. This fact should always be remembered when we are dealing with multinomial models.

Now suppose we observe a sample of n from this distribution, say, $(s_1, \ldots, s_n)$. The likelihood function for this sample is given by

$$L(\theta_1, \theta_2, \theta_3 \,|\, s_1, \ldots, s_n) = \theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}, \tag{6.1.2}$$

where x_i is the number of i's in the sample.

Using the fact that we can treat positive multiples of the likelihood as being equivalent, we see that the likelihood based on the observed counts (x_1, x_2, x_3) (since they arise from a Multinomial$(n, \theta_1, \theta_2, \theta_3)$ distribution) is given by

$$L(\theta_1, \theta_2, \theta_3 \,|\, x_1, x_2, x_3) = \theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}.$$

This is identical to the likelihood (as functions of θ_1, θ_2, and θ_3) for the original sample. It is certainly simpler to deal with the counts rather than the original sample. This is a very important phenomenon in statistics and is characterized by the concept of sufficiency, discussed in the next section. ▮

6.1.1 Sufficient Statistics

The equivalence for inference of positive multiples of the likelihood function leads to a useful equivalence amongst possible data values coming from the same model. For example, suppose data values s_1 and s_2 are such that $L(\cdot \,|\, s_1) = cL(\cdot \,|\, s_2)$ for some $c > 0$. From the point of view of likelihood, we are indifferent as to whether we obtained the data s_1 or the data s_2, as they lead to the same likelihood ratios.

This leads to the definition of a sufficient statistic.

Definition 6.1.1 A function T defined on the sample space S is called a *sufficient statistic* for the model if, whenever $T(s_1) = T(s_2)$, then

$$L(\cdot \,|\, s_1) = c(s_1, s_2)\, L(\cdot \,|\, s_2)$$

for some constant $c(s_1, s_2) > 0$.

The terminology is motivated by the fact that we need only observe the value t for the function T, as we can pick any value

$$s \in T^{-1}\{t\} = \{s : T(s) = t\}$$

and use the likelihood based on s. All of these choices give the same likelihood ratios. Typically, $T(s)$ will be of lower dimension than s, so we can consider replacing s by $T(s)$ as a *data reduction* which simplifies the analysis somewhat.

We illustrate the computation of a sufficient statistic in a simple context.

EXAMPLE 6.1.6
Suppose that $S = \{1, 2, 3, 4\}, \Omega = \{a, b\}$, and the two probability distributions are given by the following table.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$
$\theta = a$	$1/2$	$1/6$	$1/6$	$1/6$
$\theta = b$	$1/4$	$1/4$	$1/4$	$1/4$

Then $L(\cdot \,|\, 2) = L(\cdot \,|\, 3) = L(\cdot \,|\, 4)$ (e.g., $L(a \,|\, 2) = 1/6$ and $L(b \,|\, 2) = 1/4$), so the data values in $\{2, 3, 4\}$ all give the same likelihood ratios. Therefore, $T : S \longrightarrow \{0, 1\}$ given by $T(1) = 0$ and $T(2) = T(3) = T(4) = 1$ is a sufficient statistic. The model has simplified a bit, as now the sample space for T has only two elements instead of four for the original model. ∎

The following result helps identify sufficient statistics.

Theorem 6.1.1 (*Factorization theorem*) If the density (or probability function) for a model factors as $f_\theta(s) = h(s) g_\theta(T(s))$, where g_θ and h are nonnegative, then T is a sufficient statistic.

PROOF By hypothesis, it is clear that, when $T(s_1) = T(s_2)$, we have

$$\begin{aligned} L(\cdot \,|\, s_1) &= h(s_1) g_\theta(T(s_1)) = \frac{h(s_1) g_\theta(T(s_1))}{h(s_2) g_\theta(T(s_2))} h(s_2) g_\theta(T(s_2)) \\ &= \frac{h(s_1)}{h(s_2)} h(s_2) g_\theta(T(s_2)) = c(s_1, s_2)\, L(\cdot \,|\, s_2) \end{aligned}$$

because $g_\theta(T(s_1)) = g_\theta(T(s_2))$. ∎

Note that the name of this result is motivated by the fact that we have factored f_θ as a product of two functions. The important point about a sufficient statistic T is that we are indifferent, at least when considering inferences about θ, between observing the full data s or the value of $T(s)$. We will see in Chapter 9 that there is information in the data, beyond the value of $T(s)$, that is useful when we want to check assumptions.

Minimal Sufficient Statistics

Given that a sufficient statistic makes a reduction in the data, without losing relevant information in the data for inferences about θ, we look for a sufficient statistic that makes the greatest reduction. Such a statistic is called a minimal sufficient statistic.

Definition 6.1.2 A sufficient statistic T for a model is a *minimal sufficient statistic*, whenever the value of $T(s)$ can be calculated once we know the likelihood function $L(\cdot \mid s)$.

So a relevant likelihood function can always be obtained from the value of any sufficient statistic T, but if T is minimal sufficient as well, then we can also obtain the value of T from any likelihood function. It can be shown that a minimal sufficient statistic gives the greatest reduction of the data in the sense that, if T is minimal sufficient and U is sufficient, then there is a function h such that $T = h(U)$. Note that the definitions of sufficient statistic and minimal sufficient statistic depend on the model, i.e., different models can give rise to different sufficient and minimal sufficient statistics.

While the idea of a minimal sufficient statistic is a bit subtle, it is usually quite simple to find one, as the following examples illustrate.

EXAMPLE 6.1.7 *Location Normal Model*
By the factorization theorem we see immediately, from the discussion in Example 6.1.4, that $\bar{x}$ is a sufficient statistic. Now any likelihood function for this model is a positive multiple of

$$\exp\left(-\frac{n}{2\sigma_0^2}(\bar{x}-\theta)^2\right).$$

Notice that any such function of θ is completely specified by the point where it takes its maximum, namely, at $\theta = \bar{x}$. So we have that $\bar{x}$ can be obtained from any likelihood function for this model, and it is therefore a minimal sufficient statistic. ∎

EXAMPLE 6.1.8 *Location-Scale Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution in which $\mu \in R^1$ and $\sigma > 0$ are unknown. Recall the discussion and application of this model in Examples 5.3.4 and 5.5.6.

The parameter in this model is two-dimensional and is given by $\theta = \left(\mu, \sigma^2\right) \in \Omega = R^1 \times (0, \infty)$. Therefore, the likelihood function is given by

$$\begin{aligned} L(\theta \mid x_1, \ldots, x_n) &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right) \exp\left(-\frac{n-1}{2\sigma^2}s^2\right). \end{aligned}$$

We see immediately, from the factorization theorem, that $(\bar{x}, s^2)$ is a sufficient statistic.

Now, fixing σ^2, any positive multiple of $L(\cdot \mid x_1, \ldots, x_n)$ is maximized, as a function of μ, at $\mu = \bar{x}$. This is independent of σ^2. Fixing μ at $\bar{x}$, we have that

$$L\left(\left(\bar{x}, \sigma^2\right) \mid x_1, \ldots, x_n\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{n-1}{2\sigma^2}s^2\right)$$

is maximized, as a function of σ^2, at the same point as $\ln L((\bar{x}, \sigma^2) \mid x_1, \ldots, x_n)$ because ln is a strictly increasing function. Now

$$\begin{aligned} \frac{\partial \ln L\left((\bar{x}, \sigma^2) \mid x\right)}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\ln \sigma^2 - \frac{n-1}{2\sigma^2}s^2\right) \\ &= -\frac{n}{2\sigma^2} + \frac{n-1}{2\sigma^4}s^2. \end{aligned}$$

Setting this equal to 0 yields the solution

$$\hat{\sigma}^2 = \frac{n-1}{n}s^2,$$

which is a 1–1 function of s^2. So, given any likelihood function for this model, we can compute $(\bar{x}, s^2)$, which establishes that $(\bar{x}, s^2)$ is a minimal sufficient statistic for the model. In fact, the likelihood is maximized at $(\bar{x}, \hat{\sigma}^2)$ (Problem 6.1.22). ▮

EXAMPLE 6.1.9 *Multinomial Models*
We saw in Example 6.1.5 that the likelihood function for a sample is given by (6.1.2). This makes clear that if two different samples have the same counts, then they have the same likelihood, so the counts (x_1, x_2, x_3) comprise a sufficient statistic.

Now it turns out that this likelihood function is maximized by taking

$$(\theta_1, \theta_2, \theta_3) = \left(\frac{x_1}{n}, \frac{x_2}{n}, \frac{x_3}{n}\right).$$

So, given the likelihood, we can compute the counts (the sample size n is assumed known). Therefore, (x_1, x_2, x_3) is a minimal sufficient statistic. ▮

Summary of Section 6.1

- The likelihood function for a model and data shows how the data support the various possible values of the parameter. It is not the actual value of the likelihood that is important but the ratios of the likelihood at different values of the parameter.
- A sufficient statistic T for a model is any function of the data s such that once we know the value of $T(s)$, then we can determine the likelihood function $L(\cdot \mid s)$ (up to a positive constant multiple).
- A minimal sufficient statistic T for a model is any sufficient statistic such that once we know a likelihood function $L(\cdot \mid s)$ for the model and data, then we can determine $T(s)$.

EXERCISES

6.1.1 Suppose a sample of n individuals is being tested for the presence of an antibody in their blood and that the number with the antibody present is recorded. Record an appropriate statistical model for this situation when we assume that the responses from individuals are independent. If we have a sample of 10 and record 3 positives, graph a representative likelihood function.

6.1.2 Suppose that suicides occur in a population at a rate p per person year and that p is assumed completely unknown. If we model the number of suicides observed in a population with a total of N person years as Poisson(Np), then record a representative likelihood function for p when we observe 22 suicides with $N = 30,345$.

6.1.3 Suppose that the lifelengths (in thousands of hours) of light bulbs are distributed Exponential(θ), where $\theta > 0$ is unknown. If we observe $\bar{x} = 5.2$ for a sample of 20 light bulbs, record a representative likelihood function. Why is it that we only need to observe the sample average to obtain a representative likelihood?

6.1.4 Suppose we take a sample of $n = 100$ students from a university with over $50,000$ students enrolled. We classify these students as either living on campus, living off campus with their parents, or living off campus independently. Suppose we observe the counts $(x_1, x_2, x_3) = (34, 44, 22)$. Determine the form of the likelihood function for the unknown proportions of students in the population that are in these categories.

6.1.5 Determine the constant that makes the likelihood functions in Examples 6.1.2 and 6.1.3 equal.

6.1.6 Suppose that $(x_1, \ldots, x_n)$ is a sample from the Bernoulli(θ) distribution, where $\theta \in [0, 1]$ is unknown. Determine the likelihood function and a minimal sufficient statistic for this model. (Hint: Use the factorization theorem and maximize the logarithm of the likelihood function.)

6.1.7 Suppose $(x_1, \ldots, x_n)$ is a sample from the Poisson(θ) distribution where $\theta > 0$ is unknown. Determine the likelihood function and a minimal sufficient statistic for this model. (Hint: the Factorization Theorem and maximization of the logarithm of the likelihood function.)

6.1.8 Suppose that a statistical model is comprised of two distributions given by the following table:

	$s = 1$	$s = 2$	$s = 3$
$f_1(s)$	0.3	0.1	0.6
$f_2(s)$	0.1	0.7	0.2

(a) Plot the likelihood function for each possible data value s.

(b) Find a sufficient statistic that makes a reduction in the data.

6.1.9 Suppose a statistical model is given by $\{f_1, f_2\}$, where f_i is an $N(i, 1)$ distribution. Compute the likelihood ratio $L(1 \mid 0)/L(2 \mid 0)$ and explain how you interpret this number.

6.1.10 Explain why a likelihood function can never take negative values. Can a likelihood function be equal to 0 at a parameter value?

6.1.11 Suppose we have a statistical model $\{f_\theta : \theta \in [0, 1]\}$ and we observe x_0. Is it true that $\int_0^1 L(\theta \mid x_0)\, d\theta = 1$? Explain why or why not.

6.1.12 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Geometric(θ) distribution, where $\theta \in [0, 1]$ is unknown. Determine the likelihood function and a minimal sufficient statistic for this model. (Hint: Use the factorization theorem and maximize the logarithm of the likelihood.)

6.1.13 Suppose you are told that the likelihood of a particular parameter value is 10^9. Is it possible to interpret this number in any meaningful way? Explain why or why not.

6.1.14 Suppose one statistician records a likelihood function as θ^2 for $\theta \in [0, 1]$ while another statistician records a likelihood function as $100\theta^2$ for $\theta \in [0, 1]$. Explain why these likelihood functions are effectively the same.

PROBLEMS

6.1.15 Show that T defined in Example 6.1.6 is a minimal sufficient statistic. (Hint: Show that once you know the likelihood function, you can determine which of the two possible values for T has occurred.)

6.1.16 Suppose that $S = \{1, 2, 3, 4\}$, $\Omega = \{a, b, c\}$, where the three probability distributions are given by the following table.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$
$\theta = a$	1/2	1/6	1/6	1/6
$\theta = b$	1/4	1/4	1/4	1/4
$\theta = c$	1/2	1/4	1/4	0

Determine a minimal sufficient statistic for this model. Is the minimal sufficient statistic in Example 6.1.6 sufficient for this model?

6.1.17 Suppose that $(x_1, \ldots, x_n)$ is a sample from the $N(\mu, \sigma_0^2)$ distribution where $\mu \in R^1$ is unknown. Determine the form of likelihood intervals for this model.

6.1.18 Suppose that $(x_1, \ldots, x_n) \in R^n$ is a sample from f_θ, where $\theta \in \Omega$ is unknown. Show that the order statistics $(x_{(1)}, \ldots, x_{(n)})$ is a sufficient statistic for the model.

6.1.19 Determine a minimal sufficient statistic for a sample of n from the rate gamma model, i.e.,

$$f_\theta(x) = \frac{\theta^{\alpha_0}}{\Gamma(\alpha_0)} x^{\alpha_0 - 1} \exp\{-\theta x\}$$

for $x > 0$, $\theta > 0$ and where $\alpha_0 > 0$ is fixed.

6.1.20 Determine the form of a minimal sufficient statistic for a sample of size n from the Uniform$[0, \theta]$ model where $\theta > 0$.

6.1.21 Determine the form of a minimal sufficient statistic for a sample of size n from the Uniform$[\theta_1, \theta_2]$ model where $\theta_1 < \theta_2$.

6.1.22 For the location-scale normal model, establish that the point where the likelihood is maximized is given by $(\bar{x}, \hat{\sigma}^2)$ as defined in Example 6.1.8. (Hint: Show that the second derivative of $\ln L((\bar{x}, \sigma^2) \,|\, x)$, with respect to σ^2, is negative at $\hat{\sigma}^2$ and then argue that $(\bar{x}, \hat{\sigma}^2)$ is the maximum.)

6.1.23 Suppose we have a sample of n from a Bernoulli(θ) distribution where $\theta \in [0, 0.5]$. Determine a minimal sufficient statistic for this model. (Hint: It is easy to establish the sufficiency of $\bar{x}$, but this point will not maximize the likelihood when $\bar{x} > 0.5$, so $\bar{x}$ cannot be obtained from the likelihood by maximization, as in Exercise 6.1.6. In general, consider the second derivative of the log of the likelihood at any point $\theta \in (0, 0.5)$ and note that knowing the likelihood means that we can compute any of its derivatives at any values where these exist.)

6.1.24 Suppose we have a sample of n from the Multinomial$(1, \theta, 2\theta, 1 - 3\theta)$ distribution, where $\theta \in \left[0, 1/3\right]$ is unknown. Determine the form of the likelihood function and show that $x_1 + x_2$ is a minimal sufficient statistic where x_i is the number of sample values corresponding to an observation in the ith category. (Hint: Problem 6.1.23.)

6.1.25 Suppose we observe s from a statistical model with two densities f_1 and f_2. Show that the likelihood ratio $T(s) = f_1(s)/f_2(s)$ is a minimal sufficient statistic. (Hint: Use the definition of sufficiency directly.)

CHALLENGES

6.1.26 Consider the location-scale gamma model, i.e.,

$$f_{(\mu,\sigma)}(x) = \frac{1}{\Gamma(\alpha_0)}\left(\frac{x-\mu}{\sigma}\right)^{\alpha_0 - 1} \exp\left\{-\frac{x-\mu}{\sigma}\right\}\frac{1}{\sigma}$$

for $x > \mu \in R^1, \sigma > 0$ and where $\alpha_0 > 0$ is fixed.

(a) Determine the minimal sufficient statistic for a sample of n when $\alpha_0 = 1$. (Hint: Determine where the likelihood is positive and calculate the partial derivative of the log of the likelihood with respect to μ.)

(b) Determine the minimal sufficient statistic for a sample of n when $\alpha_0 \neq 1$. (Hint: Use Problem 6.1.18, the partial derivative of the log of the likelihood with respect to μ, and determine where it is infinite.)

DISCUSSION TOPICS

6.1.27 How important do you think it is for a statistician to try to quantify how much error there is in an inference drawn? For example, if an estimate is being quoted for some unknown quantity, is it important that the statistician give some indication about how accurate (or inaccurate) this inference is?

6.2 Maximum Likelihood Estimation

In Section 6.1, we introduced the likelihood function $L(\cdot \,|\, s)$ as a basis for making inferences about the unknown true value $\theta \in \Omega$. We now begin to consider the specific types of inferences discussed in Section 5.5.3 and start with estimation.

When we are interested in a point estimate of θ, then a value $\hat{\theta}(s)$ that maximizes $L(\theta \,|\, s)$ is a sensible choice, as this value is the best supported by the data, i.e.,

$$L(\hat{\theta}(s) \,|\, s) \geq L(\theta \,|\, s) \tag{6.2.1}$$

for every $\theta \in \Omega$.

Definition 6.2.1 We call $\hat{\theta} : S \to \Omega$ satisfying (6.2.1) for every $\theta \in \Omega$ a *maximum likelihood estimator*, and the value $\hat{\theta}(s)$ is called a *maximum likelihood estimate,* or *MLE* for short.

Notice that, if we use $cL(\cdot \,|\, s)$ as the likelihood function, for fixed $c > 0$, then $\hat{\theta}(s)$ is also an MLE using this version of the likelihood. So we can use any version of the likelihood to calculate an MLE.

EXAMPLE 6.2.1
Suppose the sample space is $S = \{1, 2, 3\}$, the parameter space is $\Omega = \{1, 2\}$, and the model is given by the following table.

	$s = 1$	$s = 2$	$s = 3$
$f_1(s)$	0.3	0.4	0.3
$f_2(s)$	0.1	0.7	0.2

Further suppose we observe $s = 1$. So, for example, we could be presented with one of two bowls of chips containing these proportions of chips labeled 1, 2, and 3. We draw a chip, observe that it is labelled 1, and now want to make inferences about which bowl we have been presented with.

In this case, the MLE is given by $\hat{\theta}(1) = 1$, since $0.3 = L(1 \,|\, 1) > 0.1 = L(2 \,|\, 1)$. If we had instead observed $s = 2$, then $\hat{\theta}(2) = 2$; if we had observed $s = 3$, then $\hat{\theta}(3) = 1$. ▮

Note that an MLE need not be unique. For example, in Example 6.2.1, if f_2 was defined by $f_2(1) = 0$, $f_2(2) = 0.7$ and $f_2(3) = 0.3$, then an MLE is as given there, but putting $\hat{\theta}(3) = 2$ also gives an MLE.

The MLE has a very important invariance property. Suppose we *reparameterize* a model via a 1–1 function Ψ defined on Ω. By this we mean that, instead of labelling the individual distributions in the model using $\theta \in \Omega$, we use $\psi \in \Upsilon = \{\Psi(\theta) : \theta \in \Omega\}$. For example, in Example 6.2.1, we could take $\Psi(1) = a$ and $\Psi(2) = b$ so that $\Upsilon = \{a, b\}$. So the model is now given by $\left\{g_\psi : \psi \in \Upsilon\right\}$, where $g_\psi = f_\theta$ for the unique value θ such that $\Psi(\theta) = \psi$. We have a new parameter ψ and a new parameter space Υ. Nothing has changed about the probability distributions in the statistical model, only the way they are labelled. We then have the following result.

Theorem 6.2.1 If $\hat{\theta}(s)$ is an MLE for the original parameterization and, if Ψ is a 1–1 function defined on Ω, then $\hat{\psi}(s) = \Psi(\hat{\theta}(s))$ is an MLE in the new parameterization.

PROOF If we select the likelihood function for the new parameterization to be $L^*(\psi \,|\, s) = g_\psi(s)$, and the likelihood for the original parameterization to be $L(\theta \,|\, s) = f_\theta(s)$, then we have

$$L^*(\hat{\psi}(s) \,|\, s) = g_{\Psi(\hat{\theta}(s))}(s) = f_{\hat{\theta}(s)}(s) = L(\hat{\theta}(s) \,|\, s) \geq L(\theta \,|\, s) = L^*(\Psi(\theta) \,|\, s)$$

for every $\theta \in \Omega$. This implies that $L^*(\hat{\psi}(s) \,|\, s) \geq L^*(\psi \,|\, s)$ for every $\psi \in \Upsilon$ and establishes the result. ▮

Theorem 6.2.1 shows that no matter how we parameterize the model, the MLE behaves in a consistent way under the reparameterization. This is an important property, and not all estimation procedures satisfy this.

6.2.1 Computation of the MLE

An important issue is the computation of MLEs. In Example 6.2.1, we were able to do this by simply examining the table giving the distributions. With more complicated models, this approach is not possible. In many situations, however, we can use the methods of calculus to compute $\hat{\theta}(s)$. For this we require that $f_\theta(s)$ be a continuously differentiable function of θ so that we can use optimization methods from calculus.

Rather than using the likelihood function, it is often convenient to use the log-likelihood function.

Definition 6.2.2 For likelihood function $L(\cdot \,|\, s)$, the *log-likelihood function* $l(\cdot \,|\, s)$ defined on Ω, is given by $l(\,|\, s) = \ln L(\cdot \,|\, s)$.

Note that $\ln(x)$ is a 1–1 increasing function of $x > 0$ and this implies that $L(\hat{\theta}(s) \,|\, s) \geq L(\theta \,|\, s)$ for every $\theta \in \Omega$ if and only if $l(\hat{\theta}(s) \,|\, s) \geq l(\theta \,|\, s)$ for every $\theta \in \Omega$. So we can maximize $l(\cdot \,|\, s)$ instead when computing an MLE. The convenience of the log-likelihood arises from the fact that, for a sample $(s_1, \ldots, s_n)$ from $\{f_\theta : \theta \in \Omega\}$, the likelihood function is given by

$$L(\theta \,|\, s_1, \ldots, s_n) = \prod_{i=1}^{n} f_\theta(s_i)$$

whereas the log-likelihood is given by

$$l(\theta \,|\, s_1, \ldots, s_n) = \sum_{i=1}^{n} \ln f_\theta(s_i).$$

It is typically much easier to differentiate a sum than a product.

Because we are going to be differentiating the log-likelihood, it is convenient to give a name to this derivative. We define the *score function* $S(\theta \,|\, s)$ of a model to be the derivative of its log-likelihood function whenever this exists. So when θ is a one-dimensional real-valued parameter, then

$$S(\theta \,|\, s) = \frac{\partial l(\theta \,|\, s)}{\partial \theta},$$

provided this partial derivative exists (see Appendix A.5 for a definition of partial derivative). We restrict our attention now to the situation in which θ is one-dimensional.

To obtain the MLE, we must then solve the *score equation*

$$S(\theta \,|\, s) = 0 \qquad (6.2.2)$$

for θ. Of course, a solution to (6.2.2) is not necessarily an MLE, because such a point may be a local minimum or only a local maximum rather than a global maximum. To guarantee that a solution $\hat{\theta}(s)$ is at least a local maximum, we must also check that

$$\left.\frac{\partial S(\theta \,|\, s)}{\partial \theta}\right|_{\theta=\hat{\theta}(s)} = \left.\frac{\partial^2 l(\theta \,|\, s)}{\partial \theta^2}\right|_{\theta=\hat{\theta}(s)} < 0. \qquad (6.2.3)$$

Then we must evaluate $l\,(\cdot \mid s)$ at each local maximum in order to determine the global maximum.

Let us compute some MLEs using calculus.

EXAMPLE 6.2.2 *Location Normal Model*
Consider the likelihood function

$$L\,(\theta \mid x_1, \ldots, x_n) = \exp\left(-\frac{n}{2\sigma_0^2}\,(\bar{x}-\theta)^2\right),$$

obtained in Example 6.1.4 for a sample $(x_1, \ldots, x_n)$ from the $N(\theta, \sigma_0^2)$ model where $\theta \in R^1$ is unknown and σ_0^2 is known. The log-likelihood function is then

$$l(\theta \mid x_1, \ldots, x_n) = -\frac{n}{2\sigma_0^2}\,(\bar{x}-\theta)^2\,,$$

and the score function is

$$S(\theta \mid x_1, \ldots, x_n) = \frac{n}{\sigma_0^2}\,(\bar{x}-\theta)\,.$$

The score equation is given by

$$\frac{n}{\sigma_0^2}\,(\bar{x}-\theta) = 0.$$

Solving this for θ gives the unique solution $\hat{\theta}\,(x_1, \ldots, x_n) = \bar{x}$. To check that this is a local maximum, we calculate

$$\left.\frac{\partial S(\theta \mid x_1, \ldots, x_n)}{\partial \theta}\right|_{\theta=\bar{x}.} = -\frac{n}{\sigma_0^2},$$

which is negative, and thus indicates that $\bar{x}$ is a local maximum. Because we have only one local maximum, it is also the global maximum and we have indeed obtained the MLE. ▮

EXAMPLE 6.2.3 *Exponential Model*
Suppose that a lifetime is known to be distributed Exponential$(1/\theta)$, where $\theta > 0$ is unknown. Then based on a sample $(x_1, \ldots, x_n)$, the likelihood is given by

$$L\,(\theta \mid x_1, \ldots, x_n) = \frac{1}{\theta^n}\exp\left(-\frac{n\bar{x}}{\theta}\right),$$

the log-likelihood is given by

$$l(\theta \mid x_1, \ldots, x_n) = -n\ln\theta - \frac{n\bar{x}}{\theta},$$

and the score function is given by

$$S(\theta \mid x_1, \ldots, x_n) = -\frac{n}{\theta} + \frac{n\bar{x}}{\theta^2}.$$

Solving the score equation gives $\hat{\theta}\,(x_1, \ldots, x_n) = \bar{x}$, and because $\bar{x} > 0$,

$$\left.\frac{\partial S(\theta \,|\, x_1, \ldots, x_n)}{\partial \theta}\right|_{\theta=\bar{x}} = \left.\frac{n}{\theta^2} - 2\frac{n\bar{x}}{\theta^3}\right|_{\theta=\bar{x}} = -\frac{n}{\bar{x}^2} < 0,$$

so $\bar{x}$ is indeed the MLE. ∎

In both examples just considered, we were able to derive simple formulas for the MLE. This is not always possible. Consider the following example.

EXAMPLE 6.2.4

Consider a population in which individuals are classified according to one of three types labelled 1, 2, and 3, respectively. Further suppose that the proportions of individuals falling in these categories are known to follow the law $p_1 = \theta$, $p_2 = \theta^2$, $p_3 = 1 - \theta - \theta^2$, where

$$\theta \in [0, (\sqrt{5} - 1)/2] = [0, 0.618\,03]$$

is unknown. Here, p_i denotes the proportion of individuals in the ith class. Note that the requirement that $0 \le \theta + \theta^2 \le 1$ imposes the upper bound on θ, and the precise bound is obtained by solving $\theta + \theta^2 - 1 = 0$ for θ using the formula for the roots of a quadratic. Relationships like this, amongst the proportions of the distribution of a categorical variable, often arise in genetics. For example, the categorical variable might serve to classify individuals into different genotypes.

For a sample of n (where n is small relative to the size of the population so that we can assume observations are i.i.d.), the likelihood function is given by

$$L(\theta \,|\, x_1, \ldots, x_n) = \theta^{x_1}\theta^{2x_2}\left(1 - \theta - \theta^2\right)^{x_3},$$

where x_i denotes the sample count in the ith class. The log-likelihood function is then

$$l(\theta \,|\, s_1, \ldots, s_n) = (x_1 + 2x_2)\ln\theta + x_3 \ln\left(1 - \theta - \theta^2\right),$$

and the score function is

$$S(\theta \,|\, s_1, \ldots, s_n) = \frac{(x_1 + 2x_2)}{\theta} - \frac{x_3\,(1 + 2\theta)}{1 - \theta - \theta^2}.$$

The score equation then leads to a solution $\hat{\theta}$ being a root of the quadratic

$$\begin{aligned} &(x_1 + 2x_2)\left(1 - \theta - \theta^2\right) - x_3\left(\theta + 2\theta^2\right) \\ &= -\,(x_1 + 2x_2 + 2x_3)\,\theta^2 - (x_1 + 2x_2 + x_3)\,\theta + (x_1 + 2x_2)\,. \end{aligned}$$

Using the formula for the roots of a quadratic, we obtain

$$\begin{aligned} \hat{\theta} &= \frac{1}{2\,(x_1 + 2x_2 + 2x_3)} \\ &\quad \times \left(-x_1 - 2x_2 - x_3 \pm \sqrt{5x_1^2 + 20x_1x_2 + 10x_1x_3 + 20x_2^2 + 20x_2x_3 + x_3^2}\right). \end{aligned}$$

Notice that the formula for the roots does not determine the MLE in a clear way. In fact, we cannot even tell if either of the roots lies in [0, 1]! So there are four possible values for the MLE at this point — either of the roots or the boundary points 0 and 0.61803.

We can resolve this easily in an application by simply numerically evaluating the likelihood at the four points. For example, if $x_1 = 70, x_2 = 5$, and $x_3 = 25$, then the roots are -1.28616 and 0.47847, so it is immediate that the MLE is $\hat{\theta}\,(x_1, \ldots, x_n) = 0.47847$. We can see this graphically in the plot of the log-likelihood provided in Figure 6.2.1. ∎

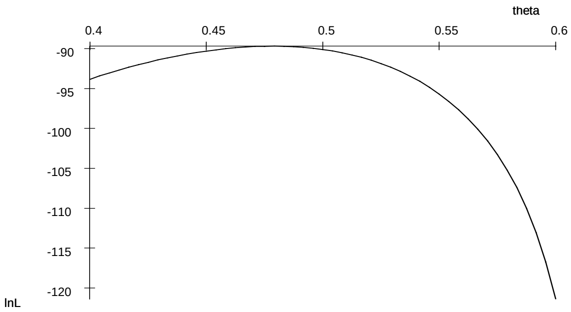

Figure 6.2.1: The log-likelihood function in Example 6.2.4 when $x_1 = 70$, $x_2 = 5$, and $x_3 = 25$.

In general, the score equation (6.2.2) must be solved numerically, using an iterative routine like Newton–Raphson. Example 6.2.4 demonstrates that we must be very careful not to just accept a solution from such a procedure as the MLE, but to check that the fundamental defining property (6.2.1) is satisfied. We also have to be careful that the necessary smoothness conditions are satisfied so that calculus can be used. Consider the following example.

EXAMPLE 6.2.5 *Uniform*$[0, \theta]$ *Model*

Suppose $(x_1, \ldots, x_n)$ is a sample from the Uniform$[0, \theta]$ model where $\theta > 0$ is unknown. Then the likelihood function is given by

$$\begin{aligned} L(\theta \,|\, x_1, \ldots, x_n) &= \begin{cases} \theta^{-n} & x_i \le \theta \text{ for } i = 1, \ldots, n \\ 0 & x_i > \theta \text{ for some } i \end{cases} \\ &= \theta^{-n} I_{[x_{(n)}, \infty)}(\theta), \end{aligned}$$

where $x_{(n)}$ is the largest order statistic from the sample. In Figure 6.2.2, we have graphed this function when $n = 10$ and $x_{(n)} = 1.916$. Notice that the maximum clearly occurs at $x_{(n)}$; we cannot obtain this value via differentiation, as $L(\cdot \,|\, x_1, \ldots, x_n)$ is not differentiable there. ∎

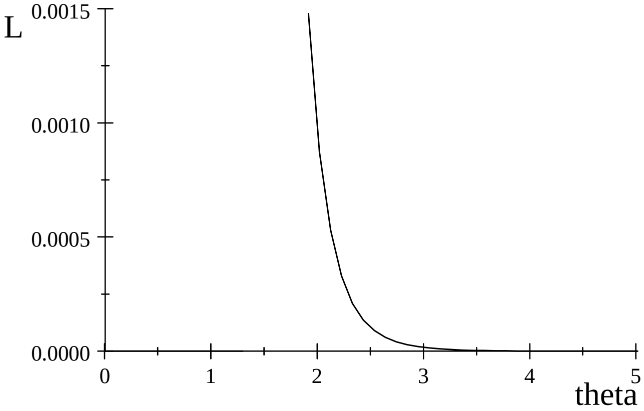

Figure 6.2.2: Plot of the likelihood function in Example 6.2.5 when $n = 10$ and $x_{(10)} = 1.916$.

The lesson of Examples 6.2.4 and 6.2.5 is that we have to be careful when computing MLEs. We now look at an example of a two-dimensional problem in which the MLE can be obtained using one-dimensional methods.

EXAMPLE 6.2.6 *Location-Scale Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $\mu \in R^1$ and $\sigma > 0$ are unknown. The parameter in this model is two-dimensional, given by $\theta = \left(\mu, \sigma^2\right) \in \Omega = R^1 \times (0, \infty)$. The likelihood function is then given by

$$L\left(\mu, \sigma^2 \,|\, x_1, \ldots, x_n\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right) \exp\left(-\frac{n-1}{2\sigma^2}s^2\right),$$

as shown in Example 6.1.8. The log-likelihood function is given by

$$l\left(\mu, \sigma^2 \,|\, x_1, \ldots, x_n\right) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{n}{2\sigma^2}(\bar{x} - \mu)^2 - \frac{n-1}{2\sigma^2}s^2. \quad (6.2.4)$$

As discussed in Example 6.1.8, it is clear that, for fixed σ^2, (6.2.4) is maximized, as a function of μ, by $\hat{\mu} = \bar{x}$. Note that this does not involve σ^2, so this must be the first coordinate of the MLE.

Substituting $\mu = \bar{x}$ into (6.2.4), we obtain

$$-\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{n-1}{2\sigma^2}s^2, \quad (6.2.5)$$

and the second coordinate of the MLE must be the value of σ^2 that maximizes (6.2.5). Differentiating (6.2.5) with respect to σ^2 and setting this equal to 0 gives

$$-\frac{n}{2\sigma^2} + \frac{n-1}{2\left(\sigma^2\right)^2}s^2 = 0. \quad (6.2.6)$$

Solving (6.2.6) for σ^2 leads to the solution

$$\hat{\sigma}^2 = \frac{n-1}{n}s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

Differentiating (6.2.6) with respect to σ^2, and substituting in $\hat{\sigma}^2$, we see that the second derivative is negative, hence $\hat{\sigma}^2$ is a point where the maximum is attained.

Therefore, we have shown that the MLE of (μ, σ^2) is given by

$$\left(\bar{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right).$$

In the following section we will show that this result can also be obtained using multidimensional calculus. ▮

So far we have talked about estimating only the full parameter θ for a model. What about estimating a general characteristic of interest $\psi(\theta)$ for some function ψ defined on the parameter space Ω? Perhaps the obvious answer here is to use the estimate $\hat{\psi}(s) = \psi(\hat{\theta}(s))$ where $\hat{\theta}(s)$ is an MLE of θ. This is sometimes referred to as the *plug-in MLE* of ψ. Notice, however, that the plug-in MLE is not necessarily a true MLE, in the sense that we have a likelihood function for a model indexed by ψ and that takes its maximum value at $\hat{\psi}(s)$. If ψ is a 1–1 function defined on Ω, then Theorem 6.2.1 establishes that $\hat{\psi}(s)$ is a true MLE but not otherwise.

If ψ is not 1–1, then we can often find a *complementing function* λ defined on Ω so that (ψ, λ) is a 1–1 function of θ. Then, by Theorem 6.2.1,

$$\left(\hat{\psi}(s), \hat{\lambda}(s)\right) = \left(\psi(\hat{\theta}(s)), \lambda(\hat{\theta}(s))\right)$$

is the joint MLE, but $\hat{\psi}(s)$ is still not formally an MLE. Sometimes a plug-in MLE can perform badly, as it ignores the information in $\lambda(\hat{\theta}(s))$ about the true value of ψ. An example illustrates this phenomenon.

EXAMPLE 6.2.7 *Sum of Squared Means*
Suppose that $X_i \sim N(\mu_i, 1)$ for $i = 1, \ldots, n$ and that these are independent with the μ_i completely unknown. So here, $\theta = (\mu_1, \ldots, \mu_n)$ and $\Omega = R^n$. Suppose we want to estimate $\psi(\theta) = \mu_1^2 + \cdots + \mu_n^2$.

The log-likelihood function is given by

$$l(\theta \,|\, x_1, \ldots, x_n) = -\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_i)^2.$$

Clearly this is maximized by $\hat{\theta}(x_1, \ldots, x_n) = (x_1, \ldots, x_n)$. So the plug-in MLE of ψ is given by $\hat{\psi} = \sum_{i=1}^{n} x_i^2$.

Now observe that

$$E_\theta\left(\sum_{i=1}^{n} X_i^2\right) = \sum_{i=1}^{n} E_\theta(X_i^2) = \sum_{i=1}^{n}\left(\text{Var}_\theta(X_i) + \mu_i^2\right) = n + \psi(\theta),$$

where $E_\theta(g)$ refers to the expectation of $g(s)$ when $s \sim f_\theta$. So when n is large, it is likely that $\hat{\psi}$ is far from the true value. An immediate improvement in this estimator is to use $\sum_{i=1}^{n} x_i^2 - n$ instead. ∎

There have been various attempts to correct problems such as the one illustrated in Example 6.2.7. Typically, these involve modifying the likelihood in some way. We do not pursue this issue further in this text but we do advise caution when using plug-in MLEs. Sometimes, as in Example 6.2.6, where we estimate μ by $\bar{x}$ and σ^2 by s^2, they seem appropriate; other times, as in Example 6.2.7, they do not.

6.2.2 The Multidimensional Case (Advanced)

We now consider the situation in which $\theta = (\theta_1, \ldots, \theta_k) \in R^k$ is multidimensional, i.e., $k > 1$. The likelihood and log-likelihood are then defined just as before, but the score function is now given by

$$S(\theta \mid s) = \begin{pmatrix} \frac{\partial l(\theta \mid s)}{\partial \theta_1} \\ \frac{\partial l(\theta \mid s)}{\partial \theta_2} \\ \vdots \\ \frac{\partial l(\theta \mid s)}{\partial \theta_k} \end{pmatrix},$$

provided all these partial derivatives exist. For the score equation, we get

$$\begin{pmatrix} \frac{\partial l(\theta \mid s)}{\partial \theta_1} \\ \frac{\partial l(\theta \mid s)}{\partial \theta_2} \\ \vdots \\ \frac{\partial l(\theta \mid s)}{\partial \theta_k} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and we must solve this k-dimensional equation for $(\theta_1, \ldots, \theta_k)$. This is often much more difficult than in the one-dimensional case, and we typically have to resort to numerical methods.

A necessary and sufficient condition for $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ to be a local maximum, when the log-likelihood has continuous second partial derivatives, is that the matrix of second partial derivatives of the log-likelihood, evaluated at $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$, must be negative definite (equivalently, all of its eigenvalues must be negative). We then must evaluate the likelihood at each of the local maxima obtained to determine the global maximum or MLE.

We will not pursue the numerical computation of MLEs in the multidimensional case any further here, but we restrict our attention to a situation in which we carry out the calculations in closed form.

EXAMPLE 6.2.8 *Location-Scale Normal Model*
We determined the log-likelihood function for this model in (6.2.4). The score function is then

$$
\begin{aligned}
S\left(\mu, \sigma^2 \,|\, x_1, \ldots, x_n\right) &= \begin{pmatrix} \frac{\partial S(\theta \,|\, x_1,\ldots,x_n)}{\partial \mu} \\ \frac{\partial S(\theta \,|\, x_1,\ldots,x_n)}{\partial \sigma^2} \end{pmatrix} \\
&= \begin{pmatrix} \frac{n}{\sigma^2}\left(\bar{x} - \mu\right) \\ -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4}\left(\bar{x} - \mu\right)^2 + -\frac{n-1}{2\sigma^4}s^2 \end{pmatrix}.
\end{aligned}
$$

The score equation is

$$
\begin{pmatrix} \frac{n}{\sigma^2}\left(\bar{x} - \mu\right) \\ -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4}\left(\bar{x} - \mu\right)^2 + \frac{n-1}{2\sigma^4}s^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},
$$

and the first of these equations immediately implies that $\hat{\mu} = \bar{x}$. Substituting this value for μ into the second equation and solving for σ^2 leads to the solution

$$
\hat{\sigma}^2 = \frac{n-1}{n}s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2.
$$

From Example 6.2.6, we know that this solution does indeed give the MLE. ▮

Summary of Section 6.2

- An MLE (maximum likelihood estimator) is a value of the parameter θ that maximizes the likelihood function. It is the value of θ that is best supported by the model and data.
- We can often compute an MLE by using the methods of calculus. When applicable, this leads to solving the score equation for θ either explicitly or using numerical algorithms. Always be careful to check that these methods are applicable to the specific problem at hand. Furthermore, always check that any solution to the score equation is a maximum and indeed an absolute maximum.

EXERCISES

6.2.1 Suppose that $S = \{1, 2, 3, 4\}, \Omega = \{a, b\}$, where the two probability distributions are given by the following table.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$
$\theta = a$	1/2	1/6	1/6	1/6
$\theta = b$	1/3	1/3	1/3	0

Determine the MLE of θ for each possible data value.

6.2.2 If $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) distribution, where $\theta \in [0, 1]$ is unknown, then determine the MLE of θ.

6.2.3 If $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) distribution, where $\theta \in [0, 1]$ is unknown, then determine the MLE of θ^2.

6.2.4 If $(x_1, \ldots, x_n)$ is a sample from a Poisson(θ) distribution, where $\theta \in (0, \infty)$ is unknown, then determine the MLE of θ.

6.2.5 If $(x_1, \ldots, x_n)$ is a sample from a Gamma(α_0, θ) distribution, where $\alpha_0 > 0$ and $\theta \in (0, \infty)$ is unknown, then determine the MLE of θ.

6.2.6 Suppose that $(x_1, \ldots, x_n)$ is the result of independent tosses of a coin where we toss until the first head occurs and where the probability of a head on a single toss is $\theta \in (0, 1]$. Determine the MLE of θ.

6.2.7 If $(x_1, \ldots, x_n)$ is a sample from a Beta$(\alpha, 1)$ distribution (see Problem 2.4.24) where $\alpha > 0$ is unknown, then determine the MLE of α. (Hint: Assume $\Gamma(\alpha)$ is a differentiable function of α.)

6.2.8 If $(x_1, \ldots, x_n)$ is a sample from a Weibull(β) distribution (see Problem 2.4.19), where $\beta > 0$ is unknown, then determine the score equation for the MLE of β.

6.2.9 If $(x_1, \ldots, x_n)$ is a sample from a Pareto(α) distribution (see Problem 2.4.20), where $\alpha > 0$ is unknown, then determine the MLE of α.

6.2.10 If $(x_1, \ldots, x_n)$ is a sample from a Log-normal(τ) distribution (see Problem 2.6.12), where $\tau > 0$ is unknown, then determine the MLE of τ.

6.2.11 Suppose you are measuring the volume of a cubic box in centimeters by taking repeated independent measurements of one of the sides. Suppose it is reasonable to assume that a single measurement follows an $N(\mu, \sigma_0^2)$ distribution, where μ is unknown and σ_0^2 is known. Based on a sample of measurements, you obtain the MLE of μ as 3.2 cm. What is your estimate of the volume of the box? How do you justify this in terms of the likelihood function?

6.2.12 If $(x_1, \ldots, x_n)$ is a sample from an $N(\mu_0, \sigma^2)$ distribution, where $\sigma^2 > 0$ is unknown and μ_0 is known, then determine the MLE of σ^2. How does this MLE differ from the plug-in MLE of σ^2 computed using the location-scale normal model?

6.2.13 Explain why it is not possible that the function $\theta^3 \exp(-(\theta - 5.3)^2)$ for $\theta \in R^1$ is a likelihood function.

6.2.14 Suppose you are told that a likelihood function has local maxima at the points $-2.2, 4.6$ and 9.2, as determined using calculus. Explain how you would determine the MLE.

6.2.15 If two functions of θ are equivalent versions of the likelihood when one is a positive multiple of the other, then when are two log-likelihood functions equivalent?

6.2.16 Suppose you are told that the likelihood of θ at $\theta = 2$ is given by $1/4$. Is this the probability that $\theta = 2$? Explain why or why not.

COMPUTER EXERCISES

6.2.17 A likelihood function is given by $\exp(-(\theta - 1)^2/2) + 3\exp(-(\theta - 2)^2/2)$ for $\theta \in R^1$. Numerically approximate the MLE by evaluating this function at 1000 equispaced points in $(-10, 10]$. Also plot the likelihood function.

6.2.18 A likelihood function is given by $\exp(-(\theta - 1)^2/2) + 3\exp(-(\theta - 5)^2/2)$ for $\theta \in R^1$. Numerically approximate the MLE by evaluating this function at 1000 equispaced points in $(-10, 10]$. Also plot the likelihood function. Comment on the form of likelihood intervals.

PROBLEMS

6.2.19 (*Hardy–Weinberg law*) The Hardy–Weinberg law in genetics says that the proportions of genotypes AA, Aa, and aa are θ^2, $2\theta(1-\theta)$, and $(1-\theta)^2$, respectively, where $\theta \in [0,1]$. Suppose that in a sample of n from the population (small relative to the size of the population), we observe x_1 individuals of type AA, x_2 individuals of type Aa, and x_3 individuals of type aa.

(a) What distribution do the counts (X_1, X_2, X_3) follow?

(b) Record the likelihood function, the log-likelihood function, and the score function for θ.

(c) Record the form of the MLE for θ.

6.2.20 If $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, 1)$ distribution where $\mu \in R^1$ is unknown, determine the MLE of the probability content of the interval $(-\infty, 1)$. Justify your answer.

6.2.21 If $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, 1)$ distribution where $\mu \geq 0$ is unknown, determine the MLE of μ.

6.2.22 Prove that, if $\hat{\theta}(s)$ is the MLE for a model for response s and if T is a sufficient statistic for the model, then $\hat{\theta}(s)$ is also the MLE for the model for $T(s)$.

6.2.23 Suppose that $(X_1, X_2, X_3) \sim$ Multinomial$(n, \theta_1, \theta_2, \theta_3)$ (see Example 6.1.5), where

$$\Omega = \{(\theta_1, \theta_2, \theta_3) : 0 \leq \theta_i \leq 1, \theta_1 + \theta_2 + \theta_3 = 1\}$$

and we observe $(X_1, X_2, X_3) = (x_1, x_2, x_3)$.

(a) Determine the MLE of $(\theta_1, \theta_2, \theta_3)$.

(b) What is the plug-in MLE of $\theta_1 + \theta_2^2 - \theta_3^2$?

6.2.24 If $(x_1, \ldots, x_n)$ is a sample from a Uniform$[\theta_1, \theta_2]$ distribution with

$$\Omega = \left\{(\theta_1, \theta_2) \in R^2 : \theta_1 < \theta_2\right\},$$

determine the MLE of (θ_1, θ_2). (Hint: You cannot use calculus. Instead, directly determine the maximum over θ_1 when θ_2 is fixed, and then vary θ_2.)

COMPUTER PROBLEMS

6.2.25 Suppose the proportion of left-handed individuals in a population is θ. Based on a simple random sample of 20, you observe four left-handed individuals.

(a) Assuming the sample size is small relative to the population size, plot the log-likelihood function and determine the MLE.

(b) If instead the population size is only 50, then plot the log-likelihood function and determine the MLE. (Hint: Remember that the number of left-handed individuals follows a hypergeometric distribution. This forces θ to be of the form $i/50$ for some integer i between 4 and 34. From a tabulation of the log-likelihood, you can obtain the MLE.)

CHALLENGES

6.2.26 If $(x_1, \ldots, x_n)$ is a sample from a distribution with density

$$f_\theta(x) = (1/2) \exp(-|x - \theta|)$$

for $x \in R^1$ and where $\theta \in R^1$ is unknown, then determine the MLE of θ. (Hint: You cannot use calculus. Instead, maximize the log-likelihood in each of the intervals $(-\infty, x_{(1)})$, $[x_{(1)} \le \theta < x_{(2)})$, etc.)

DISCUSSION TOPICS

6.2.27 One approach to quantifying the uncertainty in an MLE $\hat{\theta}(s)$ is to report the MLE together with a liklihood interval $\{\theta : L(\theta \,|\, s) \ge cL(\hat{\theta}(s) \,|\, s)\}$ for some constant $c \in (0, 1)$. What problems do you see with this approach? In particular, how would you choose c?

6.3 Inferences Based on the MLE

In Table 6.3.1. we have recorded $n = 66$ measurements of the speed of light (passage time recorded as deviations from 24,800 nanoseconds between two mirrors 7400 meters apart) made by A. A. Michelson and S. Newcomb in 1882.

28	26	33	24	34	−44	27	16	40	−2	29
22	24	21	25	30	23	29	31	19	24	20
36	32	36	28	25	21	28	29	37	25	28
26	30	32	36	26	30	22	36	23	27	27
28	27	31	27	26	33	26	32	32	24	39
28	24	25	32	25	29	27	28	29	16	23

Table 6.3.1: Speed of light measurements.

Figure 6.3.1 is a boxplot of these data with the variable labeled as x. Notice there are two outliers at $x = -2$ and $x = -44$. We will presume there is something very special about these observations and discard them for the remainder of our discussion.

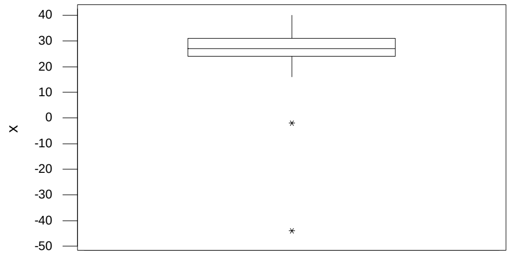

Figure 6.3.1: Boxplot of the data values in Table 6.3.1.

Figure 6.3.2 presents a histogram of these data minus the two data values identified as outliers. Notice that the histogram looks reasonably symmetrical, so it seems plausible to assume that these data are from an $N(\mu, \sigma^2)$ distribution for some values of μ and σ^2. Accordingly, a reasonable statistical model for these data would appear to be the location-scale normal model. In Chapter 9, we will discuss further how to assess the validity of the normality assumption.

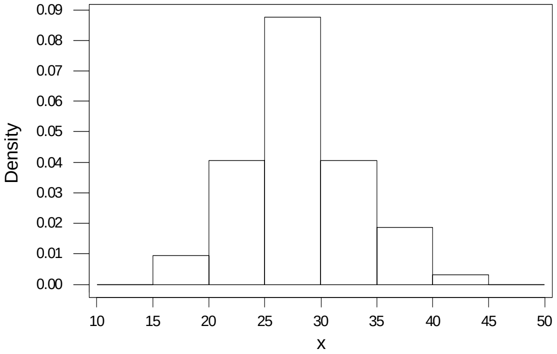

Figure 6.3.2: Density histogram of the data in Table 6.3.1 with the outliers removed.

If we accept that the location-scale normal model makes sense, the question arises concerning how to make inferences about the unknown parameters μ and σ^2. The purpose of this section is to develop methods for handling problems like this. The methods developed in this section depend on special features of the MLE in a given context. In Section 6.5, we develop a more general approach based on the MLE.

6.3.1 Standard Errors, Bias, and Consistency

Based on the justification for the likelihood, the MLE $\hat{\theta}(s)$ seems like a natural estimate of the true value of θ. Let us suppose that we will then use the plug-in MLE estimate $\hat{\psi}(s) = \psi(\hat{\theta}(s))$ for a characteristic of interest $\psi(\theta)$ (e.g., $\psi(\theta)$ might be the first quartile or the variance).

In an application, we want to know how reliable the estimate $\hat{\psi}(s)$ is. In other words, can we expect $\hat{\psi}(s)$ to be close to the true value of $\psi(\theta)$, or is there a reasonable chance that $\hat{\psi}(s)$ is far from the true value? This leads us to consider the sampling distribution of $\hat{\psi}(s)$, as this tells us how much variability there will be in $\hat{\psi}(s)$ under repeated sampling from the true distribution f_θ. Because we do not know what the true value of θ is, we have to look at the sampling distribution of $\hat{\psi}(s)$ for every $\theta \in \Omega$.

To simplify this, we substitute a numerical measure of how concentrated these sampling distributions are about $\psi(\theta)$. Perhaps the most commonly used measure of the accuracy of a general estimator $T(s)$ of $\psi(\theta)$, i.e., we are not restricting ourselves to plug-in MLEs, is the mean-squared error.

Definition 6.3.1 The *mean-squared error (MSE)* of the estimator T of $\psi(\theta) \in R^1$, is given by $\text{MSE}_\theta(T) = E_\theta((T - \psi(\theta))^2)$ for each $\theta \in \Omega$.

Clearly, the smaller $\text{MSE}_\theta(T)$ is, the more concentrated the sampling distribution of $T(s)$ is about the value $\psi(\theta)$.

Looking at $\text{MSE}_\theta(T)$ as a function of θ gives us some idea of how reliable $T(s)$ is as an estimate of the true value of $\psi(\theta)$. Because we do not know the true value of θ, and thus the true value of $\text{MSE}_\theta(T)$, statisticians record an estimate of the mean-squared error at the true value. Often

$$\text{MSE}_{\hat{\theta}(s)}(T)$$

is used for this. In other words, we evaluate $\text{MSE}_\theta(T)$ at $\theta = \hat{\theta}(s)$ as a measure of the accuracy of the estimate $T(s)$.

The following result gives an important identity for the MSE.

Theorem 6.3.1 If $\psi(\theta) \in R^1$and T is a real-valued function defined on S such that $E_\theta(T)$ exists, then

$$\text{MSE}_\theta(T) = \text{Var}_\theta(T) + (E_\theta(T) - \psi(\theta))^2 . \tag{6.3.1}$$

PROOF We have

$$\begin{aligned} E_\theta((T - \psi(\theta)))^2 &= E_\theta((T - E_\theta(T) + E_\theta(T) - \psi(\theta))^2) \\ &= E_\theta((T - E_\theta(T))^2) \\ &\quad + 2E_\theta((T - E_\theta(T))(E_\theta(T) - \psi(\theta))) + (E_\theta(T) - \psi(\theta))^2 \\ &= \text{Var}_\theta(T) + (E_\theta(T) - \psi(\theta))^2 \end{aligned}$$

because

$$\begin{aligned} E_\theta((T - E_\theta(T))(E_\theta(T) - \psi(\theta))) &= (E_\theta(T - E_\theta(T)))(E_\theta(T) - \psi(\theta)) \\ &= 0. \blacksquare \end{aligned}$$

The second term in (6.3.1) is the square of the bias in the estimator T.

Definition 6.3.2 The *bias* in the estimator T of $\psi(\theta)$ is given by $E_\theta(T) - \psi(\theta)$ whenever $E_\theta(T)$ exists. When the bias in an estimator T is 0 for every θ, we call T an *unbiased estimator* of ψ, i.e., T is unbiased whenever $E_\theta(T) = \psi(\theta)$ for every $\theta \in \Omega$.

Note that when the bias in an estimator is 0, then the MSE is just the variance.

Unbiasedness tells us that, in a sense, the sampling distribution of the estimator is centered on the true value. For unbiased estimators,

$$\text{MSE}_{\hat{\theta}(s)}(T) = \text{Var}_{\hat{\theta}(s)}(T)$$

and

$$\mathrm{Sd}_{\hat{\theta}(s)}(T) = \sqrt{\mathrm{Var}_{\hat{\theta}(s)}(T)}$$

is an estimate of the standard deviation of T and is referred to as the *standard error of the estimate* $T(s)$. As a principle of good statistical practice, whenever we quote an estimate of a quantity, we should also provide its standard error — at least when we have an unbiased estimator, as this tells us something about the accuracy of the estimate.

We consider some examples.

EXAMPLE 6.3.1 *Location Normal Model*
Consider the likelihood function

$$L(\mu \mid x_1, \ldots, x_n) = \exp\left(-\frac{n}{2\sigma_0^2}(\bar{x} - \mu)^2\right),$$

obtained in Example 6.1.4 for a sample $(x_1, \ldots, x_n)$ from the $N(\mu, \sigma_0^2)$ model, where $\mu \in R^1$ is unknown and $\sigma_0^2 > 0$ is known. Suppose we want to estimate μ. The MLE of μ was computed in Example 6.2.2 to be $\bar{x}$.

In this case, we can determine the sampling distribution of the MLE exactly from the results in Section 4.6. We have that $\bar{X} \sim N(\mu, \sigma_0^2/n)$ and so $\bar{X}$ is unbiased, and

$$\mathrm{MSE}_\mu(\bar{X}) = \mathrm{Var}_\mu(\bar{X}) = \frac{\sigma_0^2}{n},$$

which is independent of μ. So we do not need to estimate the MSE in this case. The standard error of the estimate is given by

$$\mathrm{Sd}_\mu(\bar{X}) = \frac{\sigma_0}{\sqrt{n}}.$$

Note that the standard error decreases as the population variance σ_0^2 decreases and as the sample size n increases. ∎

EXAMPLE 6.3.2 *Bernoulli Model*
Suppose $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) distribution where $\theta \in [0, 1]$ is unknown. Suppose we wish to estimate θ. The likelihood function is given by

$$L(\theta \mid x_1, \ldots, x_n) = \theta^{n\bar{x}}(1 - \theta)^{n(1-\bar{x})},$$

and the MLE of θ is $\bar{x}$ (Exercise 6.2.2), the proportion of successes in the n performances. We have $E_\theta(\bar{X}) = \theta$ for every $\theta \in [0, 1]$, so the MLE is an unbiased estimator of θ.

Therefore,

$$\mathrm{MSE}_\theta(\bar{X}) = \mathrm{Var}_\theta(\bar{X}) = \frac{\theta(1 - \theta)}{n},$$

and the estimated MSE is

$$\mathrm{MSE}_{\hat{\theta}}(\bar{X}) = \frac{\bar{x}(1 - \bar{x})}{n}.$$

The standard error of the estimate $\bar{x}$ is then given by

$$\mathrm{Sd}_{\hat{\theta}}(\bar{X}) = \sqrt{\frac{\bar{x}\,(1-\bar{x})}{n}}.$$

Note how this standard error is quite different from the standard error of $\bar{x}$ in Example 6.3.1. ▮

EXAMPLE 6.3.3 *Application of the Bernoulli Model*
A polling organization is asked to estimate the proportion of households in the population in a specific district who will participate in a proposed recycling program by separating their garbage into various components. The pollsters decided to take a sample of $n = 1000$ from the population of approximately 1.5 million households (we will say more on how to choose this number later).

Each respondent will indicate either yes or no to a question concerning their participation. Given that the sample size is small relative to the population size, we can assume that we are sampling from a Bernoulli(θ) model where $\theta \in [0, 1]$ is the proportion of individuals in the population who will respond yes.

After conducting the sample, there were 790 respondents who replied yes and 210 who responded no. Therefore, the MLE of θ is

$$\hat{\theta} = \bar{x} = \frac{790}{1000} = 0.79$$

and the standard error of the estimate is

$$\sqrt{\frac{\bar{x}\,(1-\bar{x})}{1000}} = \sqrt{\frac{0.79\,(1-0.79)}{1000}} = 0.01288.$$

Notice that it is not entirely clear how we should interpret the value 0.01288. Does it mean our estimate 0.79 is highly accurate, modestly accurate, or not accurate at all? We will discuss this further in Section 6.3.2. ▮

EXAMPLE 6.3.4 *Location-Scale Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution where $\mu \in R^1$ and $\sigma^2 > 0$ are unknown. The parameter in this model is given by $\theta = \left(\mu, \sigma^2\right) \in \Omega = R^1 \times (0, \infty)$. Suppose that we want to estimate $\mu = \psi(\mu, \sigma^2)$, i.e., just the first coordinate of the full model parameter.

In Example 6.1.8, we determined that the likelihood function is given by

$$L\left(\mu, \sigma^2 \,|\, x_1, \ldots, x_n\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}\,(\bar{x}-\mu)^2\right) \exp\left(-\frac{n-1}{2\sigma^2}s^2\right).$$

In Example 6.2.6 we showed that the MLE of θ is

$$\left(\bar{x}, \frac{n-1}{n}s^2\right).$$

Furthermore, from Theorem 4.6.6, the sampling distribution of the MLE is given by $\bar{X} \sim N(\mu, \sigma^2/n)$ independent of $(n-1)\,S^2/\sigma^2 \sim \chi^2(n-1)$.

The plug-in MLE of μ is $\bar{x}$. This estimator is unbiased and has

$$\text{MSE}_\theta(\bar{X}) = \text{Var}_\theta(\bar{X}) = \frac{\sigma^2}{n}.$$

Since σ^2 is unknown we estimate $\text{MSE}_\theta(\bar{X})$ by

$$\text{MSE}_{\hat{\theta}}(\bar{X}) = \frac{\frac{n-1}{n}s^2}{n} = \frac{n-1}{n^2}s^2 \approx \frac{s^2}{n}.$$

The value s^2/n is commonly used instead of $\text{MSE}_{\hat{\theta}}(\bar{X})$, because (Corollary 4.6.2)

$$E_\theta(S^2) = \sigma^2,$$

i.e., S^2 is an unbiased estimator of σ^2. The quantity $s/\sqrt{n}$ is referred to as the *standard error* of the estimate $\bar{x}$. ▮

EXAMPLE 6.3.5 *Application of the Location-Scale Normal Model*
In Example 5.5.6, we have a sample of $n = 30$ heights (in inches) of students. We calculated $\bar{x} = 64.517$ as our estimate of the mean population height μ. In addition, we obtained the estimate $s = 2.379$ of σ. Therefore, the standard error of the estimate $\bar{x} = 64.517$ is $s/\sqrt{30} = 2.379/\sqrt{30} = 0.43434$. As in Example 6.3.3, we are faced with interpreting exactly what this number means in terms of the accuracy of the estimate. ▮

Consistency of Estimators

Perhaps the most important property that any estimator T of a characteristic $\psi(\theta)$ can have is that it be consistent. Broadly speaking, this means that as we increase the amount of data we collect, then the sequence of estimates should converge to the true value of $\psi(\theta)$. To see why this is a necessary property of any estimation procedure, consider the finite population sampling context discussed in Section 5.4.1. When the sample size is equal to the population size, then of course we have the full information and can compute exactly every characteristic of the distribution of any measurement defined on the population. So it would be an error to use an estimation procedure for a characteristic of interest that did not converge to the true value of the characteristic as we increase the sample size.

Fortunately, we have already developed the necessary mathematics in Chapter 4 to define precisely what we mean by consistency.

Definition 6.3.3 A sequence of of estimates $T_1, T_2, \ldots$ is said to be *consistent* (in probability) for $\psi(\theta)$ if $T_n \xrightarrow{P_\theta} \psi(\theta)$ as $n \to \infty$ for every $\theta \in \Omega$. A sequence of of estimates $T_1, T_2, \ldots$ is said to be *consistent* (almost surely) for $\psi(\theta)$ if $T_n \xrightarrow{a.s.} \psi(\theta)$ as $n \to \infty$ for every $\theta \in \Omega$.

Notice that Theorem 4.3.1 says that if the sequence is consistent almost surely, then it is also consistent in probability.

Consider now a sample $(x_1, \ldots, x_n)$ from a model $\{f_\theta : \theta \in \Omega\}$ and let $T_n = n^{-1}\sum_{i=1}^n x_i$ be the nth sample average as an estimator of $\psi(\theta) = E_\theta(X)$, which we presume exists. The weak and strong laws of large numbers immediately give us the consistency of the sequence $T_1, T_2, \ldots$ for $\psi(\theta)$. We see immediately that this gives the consistency of some of the estimators discussed in this section. In fact, Theorem 6.5.2 gives the consistency of the MLE in very general circumstances. Furthermore, the plug-in MLE will also be consistent under weak restrictions on ψ. Accordingly, we can think of maximum likelihood estimation as doing the right thing in a problem at least from the point of view of consistency.

More generally, we should always restrict our attention to statistical procedures that perform correctly as the amount of data increases. Increasing the amount of data means that we are acquiring more information and thus reducing our uncertainty so that in the limit we know everything. A statistical procedure that was inconsistent would be potentially misleading.

6.3.2 Confidence Intervals

While the standard error seems like a reasonable quantity for measuring the accuracy of an estimate of $\psi(\theta)$, its interpretation is not entirely clear at this point. It turns out that this is intrinsically tied up with the idea of a *confidence interval*.

Consider the construction of an interval

$$C(s) = (l(s), u(s)),$$

based on the data s, that we believe is likely to contain the true value of $\psi(\theta)$. To do this, we have to specify the lower endpoint $l(s)$ and upper endpoint $u(s)$ for each data value s. How should we do this?

One approach is to specify a probability $\gamma \in [0, 1]$ and then require that random interval C have the *confidence property*, as specified in the following definition.

Definition 6.3.4 An interval $C(s) = (l(s), u(s))$ is a γ*-confidence interval* for $\psi(\theta)$ if $P_\theta(\psi(\theta) \in C(s)) = P_\theta(l(s) \le \psi(\theta) \le u(s)) \ge \gamma$ for every $\theta \in \Omega$. We refer to γ as the *confidence level* of the interval.

So C is a γ-confidence interval for $\psi(\theta)$ if, whenever we are sampling from P_θ, the probability that $\psi(\theta)$ is in the interval is at least equal to γ. For a given data set, such an interval either covers $\psi(\theta)$ or it does not. So note that it is not correct to say that a particular instance of a γ-confidence region has probability γ of containing the true value of $\psi(\theta)$.

If we choose γ to be a value close to 1, then we are highly confident that the true value of $\psi(\theta)$ is in $C(s)$. Of course, we can always take $C(s) = R^1$ (a very big interval!), and we are then 100% confident that the interval contains the true value. But this tells us nothing we did not already know. So the idea is to try to make use of the information in the data to construct an interval such that we have a high confidence, say, $\gamma = 0.95$ or $\gamma = 0.99$, that it contains the true value and is not any longer than necessary. We then interpret the length of the interval as a measure of how accurately the data allow us to know the true value of $\psi(\theta)$.

z-Confidence Intervals

Consider the following example, which provides one approach to the construction of confidence intervals.

EXAMPLE 6.3.6 *Location Normal Model and z-Confidence Intervals*
Suppose we have a sample $(x_1, \ldots, x_n)$ from the $N(\mu, \sigma_0^2)$ model, where $\mu \in R^1$ is unknown and $\sigma_0^2 > 0$ is known. The likelihood function is as specified in Example 6.3.1. Suppose we want a confidence interval for μ.

The reasoning that underlies the likelihood function leads naturally to the following restriction for such a region: If $\mu_1 \in C(x_1, \ldots, x_n)$ and

$$L(\mu_2 \,|\, x_1, \ldots, x_n) \geq L(\mu_1 \,|\, x_1, \ldots, x_n),$$

then we should also have $\mu_2 \in C(x_1, \ldots, x_n)$. This restriction is implied by the likelihood because the model and the data support μ_2 at least as well as μ_1. Thus, if we conclude that μ_1 is a plausible value, so is μ_2.

Therefore, $C(x_1, \ldots, x_n)$ is of the form

$$C(x_1, \ldots, x_n) = \{\mu : L\,(\mu \,|\, x_1, \ldots, x_n) \geq k\,(x_1, \ldots, x_n)\}$$

for some $k\,(x_1, \ldots, x_n)$, i.e., $C(x_1, \ldots, x_n)$ is a likelihood interval for μ. Then

$$\begin{aligned}
C(x_1, \ldots, x_n) &= \left\{\mu : \exp\left(-\frac{n}{2\sigma_0^2}\,(\bar{x}-\mu)^2\right) \geq k(x_1, \ldots, x_n)\right\} \\
&= \left\{\mu : -\frac{n}{2\sigma_0^2}\,(\bar{x}-\mu)^2 \geq \ln k(x_1, \ldots, x_n)\right\} \\
&= \left\{\mu : (\bar{x}-\mu)^2 \leq -\frac{2\sigma_0^2}{n}\ln k(x_1, \ldots, x_n)\right\} \\
&= \left[\bar{x} - k^*(x_1, \ldots, x_n)\frac{\sigma_0}{\sqrt{n}}, \bar{x} + k^*(x_1, \ldots, x_n)\,\frac{\sigma_0}{\sqrt{n}}\right]
\end{aligned}$$

where $k^*(x_1, \ldots, x_n) = \sqrt{-2\ln k(x_1, \ldots, x_n)}$.

We are now left to choose k, or equivalently k^*, so that the interval C is a γ-confidence interval for μ. Perhaps the simplest choice is to try to choose k^* so that $k^*(x_1, \ldots, x_n)$ is constant and is such that the interval as short as possible. Because

$$Z = \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \sim N(0,1), \tag{6.3.2}$$

we have

$$\begin{aligned}
\gamma &\leq P_\mu\,(\mu \in C(x_1, \ldots, x_n)) = P_\mu\left(\bar{X} - k^*\,\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + k^*\,\frac{\sigma_0}{\sqrt{n}}\right) \\
&= P_\mu\left(-k^* \leq \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \leq c\right) = P_\mu\left(\left|\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}}\right| \leq k^*\right) \\
&= 1 - 2\left(1 - \Phi\left(k^*\right)\right) \qquad (6.3.3)
\end{aligned}$$

for every $\mu \in R^1$, where Φ is the $N(0,1)$ cumulative distribution function. We have equality in (6.3.3) whenever

$$\Phi\left(k^*\right) = \frac{1+\gamma}{2},$$

and so $k^* = z_{(1+\gamma)/2}$, where z_α denotes the αth quantile of the $N(0,1)$ distribution. This is the smallest constant k^* satisfying (6.3.3).

We have shown that the likelihood interval given by

$$\left[\bar{x} - z_{(1+\gamma)/2}\frac{\sigma_0}{\sqrt{n}}, \bar{x} + z_{(1+\gamma)/2}\frac{\sigma_0}{\sqrt{n}}\right] \tag{6.3.4}$$

is an exact γ-confidence interval for μ. As these intervals are based on the *z-statistic*, given by (6.3.2), they are called *z-confidence intervals*. For example, if we take $\gamma = 0.95$, then $(1+\gamma)/2 = 0.975$, and, from a statistical package (or Table D.2 in Appendix D), we obtain $z_{0.975} = 1.96$. Therefore, in repeated sampling, 95% of the intervals of the form

$$\left[\bar{x} - 1.96\frac{\sigma_0}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma_0}{\sqrt{n}}\right]$$

will contain the true value of μ.

This is illustrated in Figure 6.3.3. Here we have plotted the upper and lower endpoints of the 0.95-confidence intervals for μ for each of $N = 25$ samples of size $n = 10$ generated from an $N(0,1)$ distribution. The theory says that when N is large, approximately 95% of these intervals will contain the true value $\mu = 0$. In the plot, coverage means that the lower endpoint (denoted by •) must be below the horizontal line at 0 and that the upper endpoint (denoted by ○) must be above this horizontal line. We see that only the fourth and twenty-third confidence intervals do not contain 0, so $23/25 = 92\%$ of the intervals contain 0. As $N \to \infty$, this proportion will converge to 0.95.

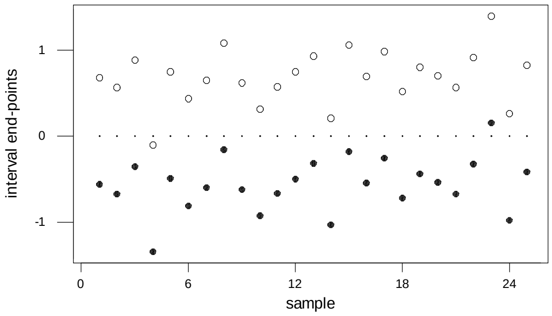

Figure 6.3.3: Plot of 0.95-confidence intervals for $\mu = 0$ (lower endpoint = •, upper endpoint = ○) for $N = 25$ samples of size $n = 10$ from an $N(0,1)$ distribution.

Notice that interval (6.3.4) is symmetrical about $\bar{x}$. Accordingly, the half-length of this interval,

$$z_{(1+\gamma)/2}\frac{\sigma_0}{\sqrt{n}},$$

is a measure of the accuracy of the estimate $\bar{x}$. The half-length is often referred to as the *margin of error*.

From the margin of error, we now see how to interpret the standard error; the standard error controls the lengths of the confidence intervals for the unknown μ. For example, we know that with probability approximately equal to 1 (actually $\gamma = 0.9974$), the interval $\left[\bar{x} \pm 3\sigma_0/\sqrt{n}\right]$ contains the true value of μ. ▮

Example 6.3.6 serves as a standard example for how confidence intervals are often constructed in statistics. Basically, the idea is that we take an estimate and then look at the intervals formed by taking symmetrical intervals around the estimate via multiples of its standard error. We illustrate this via some further examples.

EXAMPLE 6.3.7 *Bernoulli Model*

Suppose that $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) distribution where $\theta \in [0, 1]$ is unknown and we want a γ-confidence interval for θ. Following Example 6.3.2, we have that the MLE is $\bar{x}$ (see Exercise 6.2.2) and the standard error of this estimate is

$$\sqrt{\frac{\bar{x}\,(1-\bar{x})}{n}}.$$

For this model, likelihood intervals take the form

$$C(x_1, \ldots, x_n) = \{\theta : \theta^{n\bar{x}}\,(1-\theta)^{n(1-\bar{x})} \geq k(x_1, \ldots, x_n)\}$$

for some $k(x_1, \ldots, x_n)$. Again restricting to constant k, we see that to determine these intervals, we have to find the roots of equations of the form

$$\theta^{n\bar{x}}\,(1-\theta)^{n(1-\bar{x})} = k(x_1, \ldots, x_n).$$

While numerical root-finding methods can handle this quite easily, this approach is not very tractable when we want to find the appropriate value of $k(x_1, \ldots, x_n)$ to give a γ-confidence interval.

To avoid these computational complexities, it is common to use an approximate likelihood and confidence interval based on the central limit theorem. The central limit theorem (see Example 4.4.9) implies that

$$\frac{\sqrt{n}\left(\bar{X}-\theta\right)}{\sqrt{\theta\,(1-\theta)}} \xrightarrow{D} N(0,1)$$

as $n \to \infty$. Furthermore, a generalization of the central limit theorem (see Section 4.4.2), shows that

$$Z = \frac{\sqrt{n}\left(\bar{X}-\theta\right)}{\sqrt{\bar{X}\left(1-\bar{X}\right)}} \xrightarrow{D} N(0,1).$$

Therefore, we have

$$\begin{aligned}
\gamma &= \lim_{n\to\infty} P_\theta\left(-z_{(1+\gamma)/2} \leq \frac{\sqrt{n}\left(\bar{X}-\theta\right)}{\sqrt{\bar{X}\left(1-\bar{X}\right)}} \leq z_{(1+\gamma)/2}\right) \\
&= \lim_{n\to\infty} P_\theta\left(\bar{X} - z_{(1+\gamma)/2}\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n}} \leq \theta \leq \bar{X} + z_{(1+\gamma)/2}\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n}}\right),
\end{aligned}$$

and

$$\left[\bar{x} - z_{(1+\gamma)/2}\sqrt{\frac{\bar{x}\,(1-\bar{x})}{n}}, \bar{x} + z_{(1+\gamma)/2}\sqrt{\frac{\bar{x}\,(1-\bar{x})}{n}}\right] \tag{6.3.5}$$

is an approximate γ-confidence interval for θ. Notice that this takes the same form as the interval in Example 6.3.6, except that the standard error has changed.

For example, if we want an approximate 0.95-confidence interval for θ in Example 6.3.3, then based on the observed $\bar{x} = 0.79$, we obtain

$$\left[0.79 \pm 1.96\sqrt{\frac{0.79\,(1-0.79)}{1000}}\right] = [0.76475, 0.81525]\,.$$

The margin of error in this case equals 0.025245, so we can conclude that we know the true proportion with reasonable accuracy based on our sample. Actually, it may be that this accuracy is not good enough or is even too good. We will discuss methods for ensuring that we achieve appropriate accuracy in Section 6.3.5.

The γ-confidence interval derived here for θ is one of many that you will see recommended in the literature. Recall that (6.3.5) is only an approximate γ-confidence interval for θ, and n may need to be large for the approximation to be accurate. In other words, the true confidence level for (6.3.5) will not equal γ and could be far from that value if n is too small. In particular, if the true θ is near 0 or 1, then n may need to be very large. In an actual application, we usually have some idea of a small range of possible values a population proportion θ can take. Accordingly, it is advisable to carry out some simulation studies to assess whether or not (6.3.5) is going to provide an acceptable approximation for θ in that range (see Computer Exercise 6.3.21). ▮

t-Confidence Intervals

Now we consider confidence intervals for μ in an $N(\mu, \sigma^2)$ model when we drop the unrealistic assumption that we know the population variance.

EXAMPLE 6.3.8 *Location-Scale Normal Model and t-Confidence Intervals*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $\mu \in R^1$ and $\sigma > 0$ are unknown. The parameter in this model is given by $\theta = (\mu, \sigma^2) \in \Omega = R^1 \times (0, \infty)$. Suppose we want to form confidence intervals for $\mu = \psi(\mu, \sigma^2)$.

The likelihood function in this case is a function of two variables, μ and σ^2, and so the reasoning we employed in Example 6.3.6 to determine the form of the confidence interval is not directly applicable. In Example 6.3.4, we developed $s/\sqrt{n}$ as the standard error of the estimate $\bar{x}$ of μ. Accordingly, we restrict our attention to confidence intervals of the form

$$C(x_1, \ldots, x_n) = \left[\bar{x} - k\frac{s}{\sqrt{n}}, \bar{x} + k\frac{s}{\sqrt{n}}\right]$$

for some constant k.

We then have

$$P_{(\mu,\sigma^2)}\left(\bar{X}-k\,\frac{S}{\sqrt{n}}\le\mu\le\bar{X}+k\,\frac{S}{\sqrt{n}}\right)=P_{(\mu,\sigma^2)}\left(-k\le\frac{\bar{X}-\mu}{S/\sqrt{n}}\le k\right)$$
$$=P_{(\mu,\sigma^2)}\left(\left|\frac{\bar{X}-\mu}{S/\sqrt{n}}\right|\le k\right)=1-2\,(1-G\,(k\,;\,n-1))\,,$$

where $G\,(\cdot\,;\,n-1)$ is the distribution function of

$$T=\frac{\bar{X}-\mu}{S/\sqrt{n}}. \tag{6.3.6}$$

Now, by Theorem 4.6.6,

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

independent of $(n-1)\,S^2/\sigma^2\sim\chi^2\,(n-1)$. Therefore, by Definition 4.6.2,

$$T=\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)/\sqrt{\frac{(n-1)\,S^2}{\sigma^2}}=\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1).$$

So if we take

$$k=t_{(1+\gamma)/2}\,(n-1)\,,$$

where $t_\alpha\,(\lambda)$ is the αth quantile of the $t\,(\lambda)$ distribution,

$$\left[\bar{x}-t_{(1+\gamma)/2}\,(n-1)\,\frac{s}{\sqrt{n}},\,\bar{x}+t_{(1+\gamma)/2}\,(n-1)\,\frac{s}{\sqrt{n}}\right]$$

is an exact γ-confidence interval for μ. The quantiles of the t distributions are available from a statistical package (or Table D.4 in Appendix D). As these intervals are based on the *t-statistic*, given by (6.3.6), they are called *t-confidence intervals*.

These confidence intervals for μ tend to be longer than those obtained in Example 6.3.6, and this reflects the greater uncertainty due to σ being unknown. When $n=5$, then it can be shown that $\bar{x}\pm 3s/\sqrt{n}$ is a 0.97-confidence interval. When we replace s by the true value of σ, then $\bar{x}\pm 3\sigma/\sqrt{n}$ is a 0.9974-confidence interval.

As already noted, the intervals $\bar{x}\pm ks/\sqrt{n}$ are not likelihood intervals for μ. So the justification for using these must be a little different from that given in Example 6.3.6. In fact, the likelihood is defined for the full parameter $\theta=(\mu,\sigma^2)$, and it is not entirely clear how to extract inferences from it when our interest is in a marginal parameter like μ. There are a number of different attempts at resolving this issue. Here, however, we rely on the intuitive reasonableness of these intervals. In Chapter 7, we will see that these intervals also arise from another approach to inference, which reinforces our belief that the use of these intervals is appropriate.

In Example 6.3.5, we have a sample of $n=30$ heights (in inches) of students. We calculated $\bar{x}=64.517$ as our estimate of μ with standard error $s/\sqrt{30}=0.43434$. Using software (or Table D.4), we obtain $t_{0.975}\,(29)=2.0452$. So a 0.95-confidence interval for μ is given by

$$[64.517\pm 2.0452\,(0.43434)]=[63.629,65.405]\,.$$

The margin of error is 0.888, so we are very confident that the estimate $\bar{x} = 64.517$ is within an inch of the true mean height. ∎

6.3.3 Testing Hypotheses and P-Values

As discussed in Section 5.5.3, another class of inference procedures is concerned with what we call *hypothesis assessment*. Suppose there is a theory, conjecture, or hypothesis, that specifies a value for a characteristic of interest $\psi(\theta)$, say $\psi(\theta) = \psi_0$. Often this hypothesis is written $H_0 : \psi(\theta) = \psi_0$ and is referred to as the *null hypothesis*.

The word *null* is used because, as we will see in Chapter 10, the value specified in H_0 is often associated with a treatment having no effect. For example, if we want to assess whether or not a proposed new drug does a better job of treating a particular condition than a standard treatment does, the null hypothesis will often be equivalent to the new drug providing no improvement. Of course, we have to show how this can be expressed in terms of some characteristic $\psi(\theta)$ of an unknown distribution, and we will do so in Chapter 10.

The statistician is then charged with assessing whether or not the observed s is in accord with this hypothesis. So we wish to assess the evidence in s for $\psi(\theta) = \psi_0$ being true. A statistical procedure that does this can be referred to as a hypothesis assessment, a *test of significance*, or a *test of hypothesis*. Such a procedure involves measuring how surprising the observed s is when we assume H_0 to be true. It is clear that s is surprising whenever s lies in a region of low probability for each of the distributions specified by the null hypothesis, i.e., for each of the distributions in the model for which $\psi(\theta) = \psi_0$ is true. If we decide that the data are surprising under H_0, then this is evidence against H_0. This assessment is carried out by calculating a probability, called a *P-value*, so that small values of the P-value indicate that s is surprising.

It is important to always remember that while a P-value is a probability, this probability is a measure of surprise. Small values of the P-value indicate to us that a surprising event has occurred *if* the null hypothesis H_0 was true. A large P-value is not evidence that the null hypothesis is true. Moreover, a P-value is not the probability that the null hypothesis is true. The power of a hypothesis assessment method (see Section 6.3.6) also has a bearing on how we interpret a P-value.

z-Tests

We now illustrate the computation and use of P-values via several examples.

EXAMPLE 6.3.9 *Location Normal Model and the z-Test*
Suppose we have a sample $(x_1, \ldots, x_n)$ from the $N(\mu, \sigma_0^2)$ model, where $\mu \in R^1$ is unknown and $\sigma_0^2 > 0$ is known, and we have a theory that specifies a value for the unknown mean, say, $H_0 : \mu = \mu_0$. Note that, by Corollary 4.6.1, when H_0 is true, the sampling distribution of the MLE is given by $\bar{X} \sim N(\mu_0, \sigma_0^2/n)$.

So one method of assessing whether or not the hypothesis H_0 makes sense is to compare the observed value $\bar{x}$ with this distribution. If $\bar{x}$ is in a region of low probability for the $N(\mu_0, \sigma_0^2/n)$ distribution, then this is evidence that H_0 is false. Because the density of the $N(\mu_0, \sigma_0^2/n)$ distribution is unimodal, the regions of low probability for this distribution occur in its tails. The farther out in the tails $\bar{x}$ lies, the more surprising this will be when H_0 is true, and thus the more evidence we will have against H_0.

In Figure 6.3.4, we have plotted a density of the MLE together with an observed value $\bar{x}$ that lies far in the right tail of the distribution. This would clearly be a surprising value from this distribution.

So we want to measure how far out in the tails of the $N(\mu_0, \sigma_0^2/n)$ distribution the value $\bar{x}$ is. We can do this by computing the probability of observing a value of $\bar{x}$ as far, or farther, away from the center of the distribution under H_0 as $\bar{x}$. The center of this distribution is given by μ_0. Because

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0, 1) \tag{6.3.7}$$

under H_0, the P-value is then given by

$$\begin{aligned} P_{\mu_0}\left(\left|\bar{X} - \mu_0\right| \geq \left|\bar{x} - \mu_0\right|\right) &= P_{\mu_0}\left(\left|\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}\right| \geq \left|\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right|\right) \\ &= 2\left[1 - \Phi\left(\left|\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right|\right)\right], \end{aligned}$$

where Φ denotes the $N(0, 1)$ distribution function. If the P-value is small, then we have evidence that $\bar{x}$ is a surprising value because this tells us that $\bar{x}$ is out in a tail of the $N(\mu_0, \sigma_0^2/n)$ distribution. Because this P-value is based on the statistic Z defined in (6.3.7), this is referred to as the *z-test* procedure.

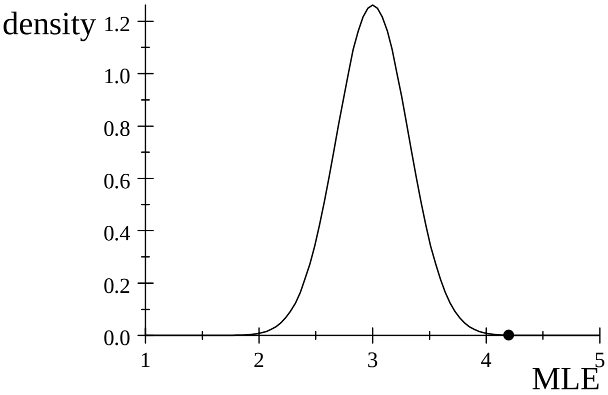

Figure 6.3.4: Plot of the density of the MLE in Example 6.3.9 when $\mu_0 = 3$, $\sigma_0^2 = 1$, and $n = 10$ together with the observed value $\bar{x} = 4.2$ (●).

▮

EXAMPLE 6.3.10 *Application of the z-Test*
We generated the following sample of $n = 10$ from an $N(26, 4)$ distribution.

29.0651	27.3980	23.4346	26.3665	23.4994
28.6592	25.5546	29.4477	28.0979	25.2850

Even though we know the true value of μ, let us suppose we do not and test the hypothesis $H_0 : \mu = 25$. To assess this, we compute (using a statistical package to evaluate Φ) the P-value

$$\begin{aligned} 2\left[1-\Phi\left(\frac{|\bar{x}-\mu_0|}{\sigma_0/\sqrt{n}}\right)\right] &= 2\left[1-\Phi\left(\frac{|26.6808-25|}{2/\sqrt{10}}\right)\right] \\ &= 2\left(1-\Phi\left(2.6576\right)\right) = 0.0078, \end{aligned}$$

which is quite small. For example, if the hypothesis H_0 is correct, then, in repeated sampling, we would see data giving a value of $\bar{x}$ at least as surprising as what we have observed only 0.78% of the time. So we conclude that we have evidence against H_0 being true, which, of course, is appropriate in this case.

If you do not use a statistical package for the evaluation of $\Phi(2.6576)$, then you will have to use Table D.2 of Appendix D to get an approximation. For example, rounding 2.6576 to 2.66, Table D.2 gives $\Phi(2.66) = 0.9961$ and the approximate P-value is $2(1-0.9961) = 0.0078$. In this case, the approximation is exact to four decimal places. ∎

EXAMPLE 6.3.11 *Bernoulli Model*

Suppose that $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) distribution, where $\theta \in [0, 1]$ is unknown, and we want to test $H_0 : \theta = \theta_0$. As in Example 6.3.7, when H_0 is true, we have

$$Z = \frac{\sqrt{n}\left(\bar{X}-\theta_0\right)}{\sqrt{\theta_0\left(1-\theta_0\right)}} \xrightarrow{D} N(0,1)$$

as $n \to \infty$. So we can test this hypothesis by computing the approximate P-value

$$P\left(|Z| \geq \left|\frac{\sqrt{n}\left(\bar{x}-\theta_0\right)}{\sqrt{\theta_0\left(1-\theta_0\right)}}\right|\right) \approx 2\left[1-\Phi\left(\left|\frac{\sqrt{n}\left(\bar{x}-\theta_0\right)}{\sqrt{\theta_0\left(1-\theta_0\right)}}\right|\right)\right]$$

when n is large.

As a specific example, suppose that a psychic claims the ability to predict the value of a randomly tossed fair coin. To test this, a coin was tossed 100 times and the psychic's guesses were recorded as successes or failures. A total of 54 successes were observed.

If the psychic has no predictive ability, then we would expect the successes to occur randomly, just like heads occur when we toss the coin. Therefore, we want to test the null hypothesis that the probability θ of a success occurring is equal to $\theta_0 = 1/2$. This is equivalent to saying that the psychic has no predictive ability. The MLE is 0.54 and the approximate P-value is given by

$$2\left[1-\Phi\left(\left|\frac{\sqrt{100}\left(0.54-0.5\right)}{\sqrt{0.5\left(1-0.5\right)}}\right|\right)\right] = 2\left(1-\Phi\left(0.8\right)\right) = 2\left(1-0.7881\right) = 0.4238,$$

and we would appear to have no evidence that H_0 is false, i.e., no reason to doubt that the psychic has no predictive ability. ∎

Often cutoff values like 0.05 or 0.01 are used to determine whether the results of a test are significant or not. For example, if the P-value is less than 0.05, then the results are said to be *statistically significant* at the 5% level. There is nothing sacrosanct about the 0.05 level, however, and different values can be used depending on the application. For example, if the result of concluding that we have evidence against H_0 is that something very expensive or important will take place, then naturally we might demand that the cutoff value be much smaller than 0.05.

When is Statistical Significance Practically Significant?

It is also important to point out here the difference between statistical significance and *practical significance*. Consider the situation in Example 6.3.9, when the true value of μ is $\mu_1 \neq \mu_0$, but μ_1 is so close to μ_0 that, practically speaking, they are indistinguishable. By the strong law of large numbers, we have that $\bar{X} \overset{a.s.}{\to} \mu_1$ as $n \to \infty$ and therefore

$$\left|\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}\right| \overset{a.s.}{\to} \infty.$$

This implies that

$$2\left[1 - \Phi\left(\left|\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}\right|\right)\right] \overset{a.s.}{\to} 0.$$

We conclude that, if we take a large enough sample size n, we will inevitably conclude that $\mu \neq \mu_0$ because the P-value of the z-test goes to 0. Of course, this is correct because the hypothesis is false.

In spite of this, we do not want to conclude that just because we have statistical significance, the difference between the true value and μ_0 is of any practical importance. If we examine the observed absolute difference $\left|\bar{x} - \mu_0\right|$ as an estimate of $\left|\mu - \mu_0\right|$, however, we will not make this mistake. If this absolute difference is smaller than some threshold δ that we consider represents a practically significant difference, then even if the P-value leads us to conclude that difference exists, we might conclude that no difference of any importance exists. Of course, the value of δ is application dependent. For example, in coin tossing, where we are testing $\theta = 1/2$, we might not care if the coin is slightly unfair, say, $|\theta - \theta_0| \leq 0.01$. In testing the abilities of a psychic, as in Example 6.3.11, however, we might take δ much lower, as any evidence of psychic powers would be an astounding finding. The issue of practical significance is something we should always be aware of when conducting a test of significance.

Hypothesis Assessment via Confidence Intervals

Another approach to testing hypotheses is via confidence intervals. For example, if we have a γ-confidence interval $C(s)$ for $\psi(\theta)$ and $\psi_0 \notin C(s)$, then this seems like clear evidence against $H_0 : \psi(\theta) = \psi_0$, at least when γ is close to 1. It turns out that in many problems, the approach to testing via confidence intervals is equivalent to using P-values with a specific cutoff for the P-value to determine statistical significance. We illustrate this equivalence using the z-test and z-confidence intervals.

EXAMPLE 6.3.12 *An Equivalence Between z-Tests and z-Confidence Intervals*
We develop this equivalence by showing that obtaining a P-value less than $1-\gamma$ for $H_0 : \mu = \mu_0$ is equivalent to μ_0 not being in a γ-confidence interval for μ. Observe that

$$1-\gamma \le 2\left[1-\Phi\left(\left|\frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}}\right|\right)\right]$$

if and only if

$$\Phi\left(\left|\frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}}\right|\right) \le \frac{1+\gamma}{2}.$$

This is true if and only if

$$\frac{\left|\bar{x}-\mu_0\right|}{\sigma_0/\sqrt{n}} \le z_{(1+\gamma)/2},$$

which holds if and only if

$$\mu_0 \in \left[\bar{x}-z_{(1+\gamma)/2}\frac{\sigma_0}{\sqrt{n}}, \bar{x}+z_{(1+\gamma)/2}\frac{\sigma_0}{\sqrt{n}}\right].$$

This implies that the γ-confidence interval for μ comprises those values μ_0 for which the P-value for the hypothesis $H_0 : \mu = \mu_0$ is greater than $1-\gamma$.

Therefore, the P-value, based on the z-statistic, for the null hypothesis $H_0 : \mu = \mu_0$, will be smaller than $1-\gamma$ if and only if μ_0 is not in the γ-confidence interval for μ derived in Example 6.3.6. For example, if we decide that for any P-values less than $1-\gamma = 0.05$, we will declare the results statistically significant, then we know the results will be significant whenever the 0.95-confidence interval for μ does not contain μ_0. For the data of Example 6.3.10, a 0.95-confidence interval is given by $[25.441, 27.920]$. As this interval does not contain $\mu_0 = 25$, we have evidence against the null hypothesis at the 0.05 level.

We can apply the same reasoning for tests about θ when we are sampling from a Bernoulli(θ) model. For the data in Example 6.3.11, we obtain the 0.95-confidence interval

$$\bar{x} \pm z_{0.975}\sqrt{\frac{\bar{x}\,(1-\bar{x})}{n}} = 0.54 \pm 1.96\sqrt{\frac{0.54\,(1-0.54)}{100}} = [0.44231, 0.63769],$$

which includes the value $\theta_0 = 0.5$. So we have no evidence against the null hypothesis of no predictive ability for the psychic at the 0.05 level. ∎

t-Tests

We now consider an example pertaining to the important location-scale normal model.

EXAMPLE 6.3.13 *Location-Scale Normal Model and t-Tests*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $\mu \in R^1$ and $\sigma > 0$ are unknown, and suppose we want to test the null hypothesis $H_0 : \mu = \mu_0$. In Example 6.3.8, we obtained a γ-confidence interval for μ. This was based on the

t-statistic given by (6.3.6). So we base our test on this statistic also. In fact, it can be shown that the test we derive here is equivalent to using the confidence intervals to assess the hypothesis as described in Example 6.3.12.

As in Example 6.3.8, we can prove that when the null hypothesis is true, then

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \tag{6.3.8}$$

is distributed $t(n-1)$. The t distributions are unimodal, with the mode at 0, and the regions of low probability are given by the tails. So we test, or assess, this hypothesis by computing the probability of observing a value as far or farther away from 0 as (6.3.8). Therefore, the P-value is given by

$$P_{(\mu_0,\sigma^2)}\left(|T| \geq \left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right|\right) = 2\left[1 - G\left(\left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right|\, ;\, n-1\right)\right],$$

where $G(\cdot\,;\,n-1)$ is the distribution function of the $t(n-1)$ distribution. We then have evidence against H_0 whenever this probability is small. This procedure is called the *t-test*. Again, it is a good idea to look at the difference $\left|\bar{x} - \mu_0\right|$, when we conclude that H_0 is false, to determine whether or not the detected difference is of practical importance.

Consider now the data in Example 6.3.10 and let us pretend that we do not know μ or σ^2. Then we have $\bar{x} = 26.6808$ and $s = \sqrt{4.8620} = 2.2050$, so to test $H_0 : \mu = 25$, the value of the t-statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{26.6808 - 25}{2.2050/\sqrt{10}} = 2.4105.$$

From a statistics package (or Table D.4) we obtain $t_{0.975}(9) = 2.2622$, so we have a statistically significant result at the 5% level and conclude that we have evidence against $H_0 : \mu = 25$. Using a statistical package, we can determine the precise value of the P-value to be 0.039 in this case. ∎

One-Sided Tests

All the tests we have discussed so far in this section for a characteristic of interest $\psi(\theta)$ have been *two-sided tests*. This means that the null hypothesis specified the value of $\psi(\theta)$ to be a single value ψ_0. Sometimes, however, we want to test a null hypothesis of the form $H_0 : \psi(\theta) \leq \psi_0$ or $H_0 : \psi(\theta) \geq \psi_0$. To carry out such tests, we use the same test statistics as we have developed in the various examples here but compute the P-value in a way that reflects the one-sided nature of the null. These are known as *one-sided tests*. We illustrate a one-sided test using the location normal model.

EXAMPLE 6.3.14 *One-Sided Tests*

Suppose we have a sample $(x_1, \ldots, x_n)$ from the $N(\mu, \sigma_0^2)$ model, where $\mu \in R^1$ is unknown and $\sigma_0^2 > 0$ is known. Suppose further that it is hypothesized that $H_0 : \mu \leq \mu_0$ is true, and we wish to assess this after observing the data.

We will base our test on the z-statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\bar{X} - \mu + \mu - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} + \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}.$$

So Z is the sum of a random variable having an $N(0, 1)$ distribution and the constant $\sqrt{n}\left(\mu - \mu_0\right)/\sigma_0$, which implies that

$$Z \sim N\left(\frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}, 1\right).$$

Note that

$$\frac{\mu - \mu_0}{\sigma_0/\sqrt{n}} \le 0$$

if and only if H_0 is true.

This implies that, when the null hypothesis is false, we will tend to see values of Z in the right tail of the $N(0, 1)$ distribution; when the null hypothesis is true, we will tend to see values of Z that are reasonable for the $N(0, 1)$ distribution, or in the left tail of this distribution. Accordingly, to test H_0, we compute the P-value

$$P\left(Z \ge \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right) = 1 - \Phi\left(\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right),$$

with $Z \sim N(0, 1)$ and conclude that we have evidence against H_0 when this is small. Using the same reasoning, the P-value for the null hypothesis $H_0 : \mu \ge \mu_0$ equals

$$P\left(Z \le \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right) = \Phi\left(\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right).$$

For more discussion of one-sided tests and confidence intervals, see Problems 6.3.25 through 6.3.32. ▮

6.3.4 Inferences for the Variance

In Sections 6.3.1, 6.3.2, and 6.3.3, we focused on inferences for the unknown mean of a distribution, e.g., when we are sampling from an $N(\mu, \sigma^2)$ distribution or a Bernoulli(θ) distribution and our interest is in μ or θ, respectively. In general, location parameters tend to play a much more important role in a statistical analysis than other characteristics of a distribution. There are logical reasons for this, discussed in Chapter 10, when we consider regression models. Sometimes we refer to a parameter such as σ^2 as a *nuisance parameter* because our interest is in μ. Note that the variance of a Bernoulli(θ) distribution is $\theta(1 - \theta)$ so that inferences about θ are logically inferences about the variance too, i.e., there are no nuisance parameters.

But sometimes we are primarily interested in making inferences about σ^2 in the $N(\mu, \sigma^2)$ distribution when it is unknown. For example, suppose that previous experience with a system under study indicates that the true value of the variance is well-approximated by σ_0^2, i.e., the true value does not differ from σ_0^2 by an amount having any practical significance. Now based on the new sample,3 we may want to assess the hypothesis $H_0 : \sigma^2 = \sigma_0^2$, i.e., we wonder whether or not the basic variability in the process has changed.

The discussion in Section 6.3.1 led to consideration of the standard error $s/\sqrt{n}$ as an estimate of the standard deviation $\sigma/\sqrt{n}$ of $\bar{x}$. In many ways s^2 seems like a very natural estimator of σ^2, even when we aren't sampling from a normal distribution.

The following example develops confidence intervals and P-values for σ^2.

EXAMPLE 6.3.15 *Location-Scale Normal Model and Inferences for the Variance*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $\mu \in R^1$ and $\sigma > 0$ are unknown, and we want to make inferences about the population variance σ^2. The plug-in MLE is given by $(n-1)\,s^2/n$, which is the average of the squared deviations of the data values from $\bar{x}$. Often s^2 is recommended as the estimate because it has the unbiasedness property, and we will use this here. An expression can be determined for the standard error of this estimate, but, as it is somewhat complicated, we will not pursue this further here.

We can form a γ-confidence interval for σ^2 using $(n-1)\,S^2/\sigma^2 \sim \chi^2(n-1)$ (Theorem 4.6.6). There are a number of possibilities for this interval, but one is to note that, letting $\chi^2_\alpha(\lambda)$ denote the αth quantile for the $\chi^2(\lambda)$ distribution, then

$$\begin{aligned}\gamma &= P_{(\mu,\sigma^2)}\left(\chi^2_{(1-\gamma)/2}(n-1) \le \frac{(n-1)\,S^2}{\sigma^2} \le \chi^2_{(1+\gamma)/2}(n-1)\right)\\ &= P_{(\mu,\sigma^2)}\left(\frac{(n-1)\,S^2}{\chi^2_{(1+\gamma)/2}(n-1)} \le \sigma^2 \le \frac{(n-1)\,S^2}{\chi^2_{(1-\gamma)/2}(n-1)}\right)\end{aligned}$$

for every $(\mu, \sigma^2) \in R^1 \times (0, \infty)$. So

$$\left[\frac{(n-1)\,s^2}{\chi^2_{(1+\gamma)/2}(n-1)}, \frac{(n-1)\,s^2}{\chi^2_{(1-\gamma)/2}(n-1)}\right]$$

is an exact γ-confidence interval for σ^2. To test a hypothesis such as $H_0 : \sigma = \sigma_0$ at the $1-\gamma$ level, we need only see whether or not σ_0^2 is in the interval. The smallest value of γ such that σ_0^2 is in the interval is the P-value for this hypothesis assessment procedure.

For the data in Example 6.3.10, let us pretend that we do not know that $\sigma^2 = 4$. Here, $n = 10$ and $s^2 = 4.8620$. From a statistics package (or Table D.3 in Appendix D) we obtain $\chi^2_{0.025}(9) = 2.700$, $\chi^2_{0.975}(9) = 19.023$. So a 0.95-confidence interval for σ^2 is given by

$$\begin{aligned}\left[\frac{(n-1)\,s^2}{\chi^2_{(1+\gamma)/2}(n-1)}, \frac{(n-1)\,s^2}{\chi^2_{(1-\gamma)/2}(n-1)}\right] &= \left[\frac{9\,(4.8620)}{19.023}, \frac{9\,(4.8620)}{2.700}\right]\\ &= [2.300\,3, 16.207]\,.\end{aligned}$$

The length of the interval indicates that there is a reasonable degree of uncertainty concerning the true value of σ^2. We see, however, that a test of $H_0 : \sigma^2 = 4$ would not reject this hypothesis at the 5% level because the value 4 is in the 0.95-confidence interval. ∎

6.3.5 Sample-Size Calculations: Confidence Intervals

Quite often a statistician is asked to determine the sample size n to ensure that with a very high probability the results of a statistical analysis will yield definitive results. For example, suppose we are going to take a sample of size n from a population Π and want to estimate the population mean μ so that the estimate is within 0.5 of the true mean with probability at least 0.95. This means that we want the half-length, or margin of error, of the 0.95-confidence interval for the mean to be guaranteed to be less than 0.5.

We consider such problems in the following examples. Note that in general, *sample-size calculations* are the domain of experimental design, which we will discuss more extensively in Chapter 10.

First, we consider the problem of selecting the sample size to ensure that a confidence interval is shorter than some prescribed value.

EXAMPLE 6.3.16 *The Length of a Confidence Interval for a Mean*
Suppose we are in the situation described in Example 6.3.6, in which we have a sample $(x_1, \ldots, x_n)$ from the $N(\mu, \sigma_0^2)$ model, with $\mu \in R^1$ unknown and $\sigma_0^2 > 0$ known. Further suppose that the statistician is asked to determine n so that the margin of error for a γ-confidence interval for the population mean μ is no greater than a prescribed value $\delta > 0$. This entails that n be chosen so that

$$z_{(1+\gamma)/2} \frac{\sigma_0}{\sqrt{n}} \le \delta$$

or, equivalently, so that

$$n \ge \sigma_0^2 \left(\frac{z_{(1+\gamma)/2}}{\delta} \right)^2 .$$

For example, if $\sigma_0^2 = 10$, $\gamma = 0.95$, and $\delta = 0.5$, then the smallest possible value for n is 154.

Now consider the situation described in Example 6.3.8, in which we have a sample $(x_1, \ldots, x_n)$ from the $N(\mu, \sigma^2)$ model with $\mu \in R^1$ and $\sigma^2 > 0$ both unknown. In this case, we want n so that

$$t_{(1+\gamma)/2}\,(n-1)\,\frac{s}{\sqrt{n}} \le \delta,$$

which entails

$$n \ge s^2 \left(\frac{t_{(1+\gamma)/2}\,(n-1)}{\delta} \right)^2 .$$

But note this also depends on the unobserved value of s, so we cannot determine an appropriate value of n.

Often, however, we can determine an upper bound on the population standard deviation, say, $\sigma \le b$. For example, suppose we are measuring human heights in centimeters. Then we have a pretty good idea of upper and lower bounds on the possible heights we will actually obtain. Therefore, with the normality assumption, the interval given by the population mean, plus or minus three standard deviations, must be contained within the interval given by the upper and lower bounds. So dividing the length of this interval by 6 gives a plausible upper bound b for the value of σ. In any case, when we have such an upper bound, we can expect that $s \le b$, at least if we choose b conservatively. Therefore, we take n to satisfy

$$n \ge b^2 \left(\frac{t_{(1+\gamma)/2}\,(n-1)}{\delta} \right)^2 .$$

Note that we need to evaluate $t_{(1+\gamma)/2}\,(n-1)$ for each n as well. It is wise to be fairly conservative in our choice of n in this case, i.e., do not choose the smallest possible value. ▮

EXAMPLE 6.3.17 *The Length of a Confidence Interval for a Proportion*
Suppose we are in the situation described in Example 6.3.2, in which we have a sample $(x_1, \ldots, x_n)$ from the Bernoulli(θ) model and $\theta \in [0, 1]$ is unknown. The statistician is required to specify the sample size n so that the margin of error of a γ-confidence interval for θ is no greater than a prescribed value δ. So, from Example 6.3.7, we want n to satisfy

$$z_{(1+\gamma)/2} \sqrt{\frac{\bar{x}\,(1-\bar{x})}{n}} \le \delta, \qquad (6.3.9)$$

and this entails

$$n \ge \bar{x}\,(1-\bar{x}) \left(\frac{z_{(1+\gamma)/2}}{\delta} \right)^2 .$$

Because this also depends on the unobserved $\bar{x}$, we cannot determine n. Note, however, that $0 \le \bar{x}\,(1-\bar{x}) \le 1/4$ for every $\bar{x}$ (plot this function) and that this upper bound is achieved when $\bar{x} = 1/2$. Therefore, if we determine n so that

$$n \ge \frac{1}{4} \left(\frac{z_{(1+\gamma)/2}}{\delta} \right)^2 ,$$

then we know that (6.3.9) is satisfied. For example, if $\gamma = 0.95, \delta = 0.1$, the smallest possible value of n is 97; if $\gamma = 0.95, \delta = 0.01$, the smallest possible value of n is 9604. ▮

6.3.6 Sample-Size Calculations: Power

Suppose the purpose of a study is to assess a specific hypothesis $H_0 : \psi(\theta) = \psi_0$ and it is has been decided that the results will be declared statistically significant whenever the P-value is less than α. Suppose that the statistician is asked to choose n, so that the P-value obtained is smaller than α, with probability at least β_0, at some specific θ_1 such that $\psi(\theta_1) \ne \psi_0$. The probability that the P-value is less than α for a specific value of θ is called the *power* of the test at θ. We will denote this by $\beta(\theta)$ and call β the *power function* of the test. The notation β is not really complete, as it suppresses the dependence of β on ψ, ψ_0, α, n, and the test procedure, but we will assume that these are clear in a particular context. The problem the statistician is presented with can then be stated as: Find n so that $\beta(\theta_1) \ge \beta_0$.

The power function of a test is a measure of the sensitivity of the test to detect departures from the null hypothesis. We choose α small ($\alpha = 0.05, 0.01$, etc.) so that we do not erroneously declare that we have evidence against the null hypothesis when the null hypothesis is in fact true. When $\psi(\theta) \neq \psi_0$, then $\beta(\theta)$ is the probability that the test does the right thing and detects that H_0 is false.

For any test procedure, it is a good idea to examine its power function, perhaps for several choices of α, to see how good the test is at detecting departures. For it can happen that we do not find any evidence against a null hypothesis when it is false because the sample size is too small. In such a case, the power will be small at θ values that represent practically significant departures from H_0. To avoid this problem, we should always choose a value ψ_1 that represents a practically significant departure from ψ_0 and then determine n so that we reject H_0 with high probability when $\psi(\theta) = \psi_1$.

We consider the computation and use of the power function in several examples.

EXAMPLE 6.3.18 *The Power Function in the Location Normal Model*
For the two-sided z-test in Example 6.3.9, we have

$$\begin{aligned}
\beta(\mu) &= P_\mu\left(2\left[1-\Phi\left(\left|\frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}}\right|\right)\right] < \alpha\right) \\
&= P_\mu\left(\Phi\left(\left|\frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}}\right|\right) > 1-\frac{\alpha}{2}\right) = P_\mu\left(\left|\frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}}\right| > z_{(1-\alpha/2)}\right) \\
&= P_\mu\left(\frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} > z_{(1-\alpha/2)}\right) + P_\mu\left(\frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} < -z_{(1-\alpha/2)}\right) \\
&= P_\mu\left(\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} > \frac{\mu_0-\mu}{\sigma_0/\sqrt{n}} + z_{(1-\alpha/2)}\right) + P_\mu\left(\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} < \frac{\mu_0-\mu}{\sigma_0/\sqrt{n}} - z_{(1-\alpha/2)}\right) \\
&= 1-\Phi\left(\frac{\mu_0-\mu}{\sigma_0/\sqrt{n}} + z_{(1-\alpha/2)}\right) + \Phi\left(\frac{\mu_0-\mu}{\sigma_0/\sqrt{n}} - z_{(1-\alpha/2)}\right). \qquad (6.3.10)
\end{aligned}$$

Notice that

$$\beta(\mu) = \beta(\mu_0 + (\mu-\mu_0)) = \beta(\mu_0 - (\mu-\mu_0)),$$

so β is symmetric about μ_0 (put $\delta = \mu - \mu_0$ and $\mu = \mu_0 + \delta$ in the expression for $\beta(\mu)$ and we get the same value).

Differentiating (6.3.10) with respect to $\sqrt{n}$, we obtain

$$\left[\varphi\left(\frac{\mu_0-\mu}{\sigma_0/\sqrt{n}} - z_{(1-\alpha/2)}\right) - \varphi\left(\frac{\mu_0-\mu}{\sigma_0/\sqrt{n}} + z_{(1-\alpha/2)}\right)\right]\frac{\mu_0-\mu}{\sigma_0} \qquad (6.3.11)$$

where φ is the density of the $N(0,1)$ distribution. We can establish that (6.3.11) is always nonnegative (see Challenge 6.3.34). This implies that $\beta(\mu)$ is increasing in n, so we need only solve $\beta(\mu_1) = \beta_0$ for n (the solution may not be an integer) to determine a suitable sample size (all larger values of n will give a larger power).

For example, when $\sigma_0 = 1, \alpha = 0.05, \beta_0 = 0.99$, and $\mu_1 = \mu_0 \pm 0.1$, we must find n satisfying

$$1-\Phi\left(\sqrt{n}(0.1)+1.96\right) + \Phi\left(\sqrt{n}(0.1)-1.96\right) = 0.99. \qquad (6.3.12)$$

(Note the symmetry of β about μ_0 means we will get the same answer if we use $\mu_0 - 0.1$ here instead of $\mu_0 + 0.1$.) Tabulating (6.3.12) as a function of n using a statistical package determines that $n = 785$ is the smallest value achieving the required bound.

Also observe that the derivative of (6.3.10) with respect to μ is given by

$$\left[\varphi\left(\frac{\mu_0-\mu}{\sigma_0/\sqrt{n}}+z_{(1-\alpha/2)}\right)-\varphi\left(\frac{\mu_0-\mu}{\sigma_0/\sqrt{n}}-z_{(1-\alpha/2)}\right)\right]\frac{\sqrt{n}}{\sigma_0}. \tag{6.3.13}$$

This is positive when $\mu > \mu_0$, negative when $\mu < \mu_0$, and takes the value 0 when $\mu = \mu_0$ (see Challenge 6.3.35). From (6.3.10) we have that $\beta(\mu) \to 1$ as $\mu \to \pm\infty$. These facts establish that β takes its minimum value at μ_0 and that it is increasing as we move away from μ_0. Therefore, once we have determined n so that the power is at least β_0 at some μ_1, we know that the power is at least β_0 for all values of μ satisfying $\left|\mu_0-\mu\right| \geq \left|\mu_0-\mu_1\right|$.

As an example of this, consider Figure 6.3.5, where we have plotted the power function when $n = 10$, $\mu_0 = 0$, $\sigma_0 = 1$, and $\alpha = 0.05$ so that

$$\beta(\mu) = 1-\Phi\left(\sqrt{10}\mu+1.96\right)+\Phi\left(\sqrt{10}\mu-1.96\right).$$

Notice the symmetry about $\mu_0 = 0$ and the fact that $\beta(\mu)$ increases as μ moves away from 0. We obtain $\beta(1.2) = 0.967$ so that when $\mu = \pm 1.2$, the probability that the P-value for testing $H_0 : \mu = 0$ will be less than 0.05 is 0.967. Of course, as we increase n, this graph will rise even more steeply to 1 as we move away from 0.

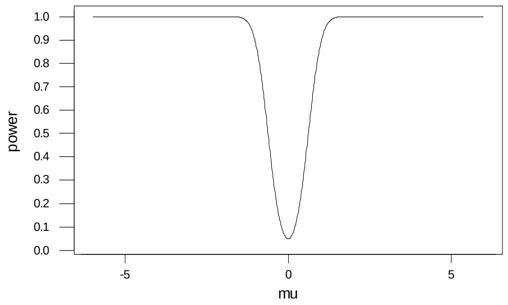

Figure 6.3.5: Plot of the power function $\beta(\mu)$ for Example 6.3.18 when $\alpha = 0.05$, $\mu_0 = 0$, and $\sigma_0 = 1$ is assumed known.

Many statistical packages contain the power function as a built-in function for various tests. This is very convenient for examining the sensitivity of the test and determining sample sizes. ∎

EXAMPLE 6.3.19 *The Power Function for θ in the Bernoulli Model*
For the two-sided test in Example 6.3.11, we have that the power function is given by

$$\beta(\theta) = P_\theta\left(2\left[1-\Phi\left(\left|\frac{\sqrt{n}\left(\bar{X}-\theta_0\right)}{\sqrt{\theta_0\left(1-\theta_0\right)}}\right|\right)<\alpha\right]\right).$$

Under the assumption that we choose n large enough so that $\bar{X}$ is approximately distributed $N(\theta, \theta(1-\theta)/n)$, the approximate calculation of this power function can be approached as in Example 6.3.18, when we put $\sigma_0 = \theta(1-\theta)$. We do not pursue this calculation further here but note that many statistical packages will evaluate β as a built-in function. ▮

EXAMPLE 6.3.20 *The Power Function in the Location-Scale Normal Model*
For the two-sided t-test in Example 6.3.13, we have

$$\begin{aligned}\beta_n(\mu,\sigma^2) &= P_{(\mu,\sigma^2)}\left(2\left[1-G\left(\left|\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right| ;\, n-1\right)\right] < \alpha\right)\\ &= P_{(\mu,\sigma^2)}\left(\left|\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right| > t_{(1-\alpha/2)}\,(n-1)\right),\end{aligned}$$

where $G(\cdot\,;\, n-1)$ is the cumulative distribution function of the $t(n-1)$ distribution. Notice that it is a function of both μ and σ^2. In particular, we have to specify both μ and σ^2 and then determine n so that $\beta_n\left(\mu,\sigma^2\right) \ge \beta_0$. Many statistical packages will have the calculation of this power function built-in so that an appropriate n can be determined using this. Alternatively, we can use Monte Carlo methods to approximate the distribution function of

$$\left|\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right|$$

when sampling from the $N(\mu,\sigma^2)$, for a variety of values of n, to determine an appropriate value. ▮

Summary of Section 6.3

- The MLE $\hat{\theta}$ is the best supported value of the parameter θ by the model and data. As such, it makes sense to base the derivation of inferences about some characteristic $\psi(\theta)$ on the MLE. These inferences include estimates and their standard errors, confidence intervals, and the assessment of hypotheses via P-values.
- An important aspect of the design of a sampling study is to decide on the size n of the sample to ensure that the results of the study produce sufficiently accurate results. Prescribing the half-lengths of confidence intervals (margins of error) or the power of a test are two techniques for doing this.

EXERCISES

6.3.1 Suppose measurements (in centimeters) are taken using an instrument. There is error in the measuring process and a measurement is assumed to be distributed $N(\mu,\sigma_0^2)$, where μ is the exact measurement and $\sigma_0^2 = 0.5$. If the $(n = 10)$ measurements 4.7, 5.5, 4.4, 3.3, 4.6, 5.3, 5.2, 4.8, 5.7, 5.3 were obtained, assess the hypothesis $H_0 : \mu = 5$ by computing the relevant P-value. Also compute a 0.95-confidence interval for the unknown μ.

6.3.2 Suppose in Exercise 6.3.1, we drop the assumption that $\sigma_0^2 = 0.5$. Then assess the hypothesis $H_0 : \mu = 5$ and compute a 0.95-confidence interval for μ.

6.3.3 Marks on an exam in a statistics course are assumed to be normally distributed with unknown mean but with variance equal to 5. A sample of four students is selected, and their marks are 52, 63, 64, 84. Assess the hypothesis $H_0 : \mu = 60$ by computing the relevant P-value and compute a 0.95-confidence interval for the unknown μ.

6.3.4 Suppose in Exercise 6.3.3 that we drop the assumption that the population variance is 5. Assess the hypothesis $H_0 : \mu = 60$ by computing the relevant P-value and compute a 0.95-confidence interval for the unknown μ.

6.3.5 Suppose that in Exercise 6.3.3 we had only observed one mark and that this was 52. Assess the hypothesis $H_0 : \mu = 60$ by computing the relevant P-value and compute a 0.95-confidence interval for the unknown μ. Is it possible to compute a P-value and construct a 0.95-confidence interval for μ without the assumption that we know the population variance? Explain your answer and, if your answer is no, determine the minimum sample size n for which inference is possible without the assumption that the population variance is known.

6.3.6 Assume that the speed of light data in Table 6.3.1 is a sample from an $N(\mu, \sigma^2)$ distribution for some unknown values of μ and σ^2. Determine a 0.99-confidence interval for μ. Assess the null hypothesis $H_0 : \mu = 24$.

6.3.7 A manufacturer wants to assess whether or not rods are being constructed appropriately, where the diameter of the rods is supposed to be 1.0 cm and the variation in the diameters is known to be distributed $N(\mu, 0.1)$. The manufacturer is willing to tolerate a deviation of the population mean from this value of no more than 0.1 cm, i.e., if the population mean is within the interval 1.0 ±0.1 cm, then the manufacturing process is performing correctly. A sample of $n = 500$ rods is taken, and the average diameter of these rods is found to be $\bar{x} = 1.05$ cm, with $s^2 = 0.083$ cm^2. Are these results statistically significant? Are the results practically significant? Justify your answers.

6.3.8 A polling firm conducts a poll to determine what proportion θ of voters in a given population will vote in an upcoming election. A random sample of $n = 250$ was taken from the population, and the proportion answering yes was 0.62. Assess the hypothesis $H_0 : \theta = 0.65$ and construct an approximate 0.90-confidence interval for θ.

6.3.9 A coin was tossed $n = 1000$ times, and the proportion of heads observed was 0.51. Do we have evidence to conclude that the coin is unfair?

6.3.10 How many times must we toss a coin to ensure that a 0.95-confidence interval for the probability of heads on a single toss has length less than 0.1, 0.05, and 0 .01, respectively?

6.3.11 Suppose a possibly biased die is rolled 30 times and that the face containing two pips comes up 10 times. Do we have evidence to conclude that the die is biased?

6.3.12 Suppose a measurement on a population is assumed to be distributed $N(\mu, 2)$ where $\mu \in R^1$ is unknown and that the size of the population is very large. A researcher wants to determine a 0.95-confidence interval for μ that is no longer than 1. What is the minimum sample size that will guarantee this?

6.3.13 Suppose $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) with $\theta \in [0, 1]$ unknown.
(a) Show that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = n\bar{x}(1 - \bar{x})$. (Hint: $x_i^2 = x_i$.)

(b) If $X \sim$ Bernoulli(θ), then $\sigma^2 = \text{Var}(X) = \theta(1-\theta)$. Record the relationship between the plug-in estimate of σ^2 and that given by s^2 in (5.5.5).

(c) Since s^2 is an unbiased estimator of σ^2 (see Problem 6.3.23), use the results in part (b) to determine the bias in the plug-in estimate. What happens to this bias as $n \to \infty$?

6.3.14 Suppose you are told that, based on some data, a 0.95-confidence interval for a characteristic $\psi(\theta)$ is given by $(1.23, 2.45)$. You are than asked if there is any evidence against the hypothesis $H_0 : \psi(\theta) = 2$. State your conclusion and justify your reasoning.

6.3.15 Suppose that x_1 is a value from a Bernoulli(θ) with $\theta \in [0, 1]$ unknown.

(a) Is x_1 an unbiased estimator of θ?

(a) Is x_1^2 an unbiased estimator of θ^2?

6.3.16 Suppose a plug-in MLE of a characteristic $\psi(\theta)$ is given by 5.3. Also a P-value was computed to assess the hypothesis $H_0 : \psi(\theta) = 5$ and the value was 0.000132. If you are told that differences among values of $\psi(\theta)$ less than 0.5 are of no importance as far as the application is concerned, then what do you conclude from these results? Suppose instead you were told that differences among values of $\psi(\theta)$ less than 0.25 are of no importance as far as the application is concerned, then what do you conclude from these results?

6.3.17 A P-value was computed to assess the hypothesis $H_0 : \psi(\theta) = 0$ and the value 0.22 was obtained. The investigator says this is strong evidence that the hypothesis is correct. How do you respond?

6.3.18 A P-value was computed to assess the hypothesis $H_0 : \psi(\theta) = 1$ and the value 0.55 was obtained. You are told that differences in $\psi(\theta)$ greater than 0.5 are considered to be practically significant but not otherwise. The investigator wants to know if enough data were collected to reliably detect a difference of this size or greater. How would you respond?

COMPUTER EXERCISES

6.3.19 Suppose a measurement on a population can be assumed to follow the $N(\mu, \sigma^2)$ distribution, where $(\mu, \sigma^2) \in R^1 \times (0, \infty)$ is unknown and the size of the population is very large. A very conservative upper bound on σ is given by 5. A researcher wants to determine a 0.95-confidence interval for μ that is no longer than 1. Determine a sample size that will guarantee this. (Hint: Start with a large sample approximation.)

6.3.20 Suppose a measurement on a population is assumed to be distributed $N(\mu, 2)$, where $\mu \in R^1$ is unknown and the size of the population is very large. A researcher wants to assess a null hypothesis $H_0 : \mu = \mu_0$ and ensure that the probability is at least 0.80 that the P-value is less than 0.05 when $\mu = \mu_0 \pm 0.5$. What is the minimum sample size that will guarantee this? (Hint: Tabulate the power as a function of the sample size n.)

6.3.21 Generate 10^3 samples of size $n = 5$ from the Bernoulli(0.5) distribution. For each of these samples, calculate (6.3.5) with $\gamma = 0.95$ and record the proportion of intervals that contain the true value. What do you notice? Repeat this simulation with $n = 20$. What do you notice?

6.3.22 Generate 10^4 samples of size $n = 5$ from the $N(0, 1)$ distribution. For each of these samples, calculate the interval $(\bar{x} - s/\sqrt{5}, \bar{x} + s/\sqrt{5})$, where s is the sample standard deviation, and compute the proportion of times this interval contains μ. Repeat this simulation with $n = 10$ and 100 and compare your results.

PROBLEMS

6.3.23 Suppose that $(x_1, \ldots, x_n)$ is a sample from a distribution with mean μ and variance σ^2.

(a) Prove that s^2 given by (5.5.5) is an unbiased estimator of σ^2.

(b) If instead we estimate σ^2 by $(n - 1)s^2/n$, then determine the bias in this estimate and what happens to it as $n \to \infty$.

6.3.24 Suppose we have two unbiased estimators T_1 and T_2 of $\psi(\theta) \in R^1$.

(a) Show that $\alpha T_1 + (1 - \alpha)T_2$ is also an unbiased estimator of $\psi(\theta)$ whenever $\alpha \in [0, 1]$.

(b) If T_1 and T_2 are also independent, e.g., determined from independent samples, then calculate $\text{Var}_\theta(\alpha T_1 + (1 - \alpha)T_2)$ in terms of $\text{Var}_\theta(T_1)$ and $\text{Var}_\theta(T_2)$.

(c) For the situation in part (b), determine the best choice of α in the sense that for this choice $\text{Var}_\theta(\alpha T_1 + (1 - \alpha)T_2)$ is smallest. What is the effect on this combined estimator of T_1 having a very large variance relative to T_2?

(d) Repeat part (b) and (c), but now do not assume that T_1 and T_2 are independent so $\text{Var}_\theta(\alpha T_1 + (1 - \alpha)T_2)$ will also involve $\text{Cov}_\theta(T_1, T_2)$.

6.3.25 (*One-sided confidence intervals for means*) Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where $\mu \in R^1$ is unknown and σ_0^2 is known. Suppose we want to make inferences about the interval $\psi(\mu) = (-\infty, \mu)$. Consider the problem of finding an interval $C(x_1, \ldots, x_n) = (-\infty, u(x_1, \ldots, x_n))$ that covers the interval $(-\infty, \mu)$ with probability at least γ. So we want u such that for every μ,

$$P_\mu(\mu \le u(X_1, \ldots, X_n)) \ge \gamma.$$

Note that $(-\infty, \mu) \subset (-\infty, u(x_1, \ldots, x_n))$ if and only if $\mu \le u(x_1, \ldots, x_n)$, so $C(x_1, \ldots, x_n)$ is called a left-sided γ-confidence interval for μ. Obtain an exact left-sided γ-confidence interval for μ using $u(x_1, \ldots, x_n) = \bar{x} + k\left(\sigma_0/\sqrt{n}\right)$, i.e., find the k that gives this property.

6.3.26 (*One-sided hypotheses for means*) Suppose that $(x_1, \ldots, x_n)$ is a sample from a $N(\mu, \sigma_0^2)$ distribution, where μ is unknown and σ_0^2 is known. Suppose we want to assess the hypothesis $H_0 : \mu \le \mu_0$. Under these circumstances, we say that the observed value $\bar{x}$ is surprising if $\bar{x}$ occurs in a region of low probability for every distribution in H_0. Therefore, a sensible P-value for this problem is $\max_{\mu \in H_0} P_\mu(\bar{X} > \bar{x})$. Show that this leads to the P-value $1 - \Phi((\bar{x} - \mu_0)/(\sigma_0/\sqrt{n}))$.

6.3.27 Determine the form of the power function associated with the hypothesis assessment procedure of Problem 6.3.26, when we declare a test result as being statistically significant whenever the P-value is less than α.

6.3.28 Repeat Problems 6.3.25 and 6.3.26, but this time obtain a right-sided γ-confidence interval for μ and assess the hypothesis $H_0 : \mu \ge \mu_0$.

6.3.29 Repeat Problems 6.3.25 and 6.3.26, but this time do not assume the population variance is known. In particular, determine k so that $u\left(x_1, \ldots, x_n\right) = \bar{x} + k\left(s/\sqrt{n}\right)$ gives an exact left-sided γ-confidence interval for μ and show that the P-value for testing $H_0 : \mu \leq \mu_0$ is given by

$$1 - G\left(\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\,;\; n-1\right).$$

6.3.30 (*One-sided confidence intervals for variances*) Suppose that $(x_1, \ldots, x_n)$ is a sample from the $N(\mu, \sigma^2)$ distribution, where $\left(\mu, \sigma^2\right) \in R^1 \times (0, \infty)$ is unknown, and we want a γ-confidence interval of the form

$$C(x_1, \ldots, x_n) = (0, u\left(x_1, \ldots, x_n\right))$$

for σ^2. If $u\left(x_1, \ldots, x_n\right) = ks^2$, then determine k so that this interval is an exact γ-confidence interval.

6.3.31 (*One-sided hypotheses for variances*) Suppose that $(x_1, \ldots, x_n)$ is a sample from the $N(\mu, \sigma^2)$ distribution, where $\left(\mu, \sigma^2\right) \in R^1 \times (0, \infty)$ is unknown, and we want to assess the hypothesis $H_0 : \sigma^2 \leq \sigma_0^2$. Argue that the sample variance s^2 is surprising if s^2 is large and that, therefore, a sensible P-value for this problem is to compute $\max_{(\mu,\sigma^2)\in H_0} P_\mu\left(S^2 > s^2\right)$. Show that this leads to the P-value

$$1 - H\left(\frac{(n-1)\,s^2}{\sigma_0^2}\,;\; n-1\right),$$

where $H\left(\cdot\,;\; n-1\right)$ is the distribution function of the $\chi^2\left(n-1\right)$ distribution.

6.3.32 Determine the form of the power function associated with the hypothesis assessment procedure of Problem 6.3.31, for computing the probability that the P-value is less than α.

6.3.33 Repeat Exercise 6.3.7, but this time do not assume that the population variance is known. In this case, the manufacturer deems the process to be under control if the population standard deviation is less than or equal to 0.1 and the population mean is in the interval 1.0 ±0.1 cm. Use Problem 6.3.31 for the test concerning the population variance.

CHALLENGES

6.3.34 Prove that (6.3.11) is always nonnegative. (Hint: Use the facts that φ is symmetric about 0, increases to the left of 0, and decreases to the right of 0.)

6.3.35 Establish that (6.3.13) is positive when $\mu > \mu_0$, negative when $\mu < \mu_0$, and takes the value 0 when $\mu = \mu_0$.

DISCUSSION TOPICS

6.3.36 Discuss the following statement: The accuracy of the results of a statistical analysis are so important that we should always take the largest possible sample size.

6.3.37 Suppose we have a sequence of estimators $T_1, T_2, \ldots$ for $\psi(\theta)$ and $T_n \xrightarrow{P} \psi(\theta) + \epsilon(\theta)$ as $n \to \infty$ for each $\theta \in \Omega$. Discuss under what circumstances you might consider T_n a useful estimator of $\psi(\theta)$.

6.4 Distribution-Free Methods

The likelihood methods we have been discussing all depend on the assumption that the true distribution lies in $\{P_\theta : \theta \in \Omega\}$. There is typically nothing that guarantees that the assumption $\{P_\theta : \theta \in \Omega\}$ is correct. If the distribution we are sampling from is far different from any of the distributions in $\{P_\theta : \theta \in \Omega\}$, then methods of inference that depend on this assumption, such as likelihood methods, can be very misleading. So it is important in any application to check that our assumptions make sense. We will discuss the topic of model checking in Chapter 9.

Another approach to this problem is to take the model $\{P_\theta : \theta \in \Omega\}$ as large as possible, reflecting the fact that we may have very little information about what the true distribution is like. For example, inferences based on the Bernoulli(θ) model with $\theta \in \Omega = [0, 1]$ really specify no information about the true distribution because this model includes all the possible distributions on the sample space $S = \{0, 1\}$. Inference methods that are suitable when $\{P_\theta : \theta \in \Omega\}$ is very large are sometimes called *distribution-free*, to reflect the fact that very little information is specified in the model about the true distribution.

For finite sample spaces, it is straightforward to adopt the distribution-free approach, as with the just cited Bernoulli model, but when the sample space is infinite, things are more complicated. In fact, sometimes it is very difficult to determine inferences about characteristics of interest when the model is very big. Furthermore, if we have

$$\{P_\theta : \theta \in \Omega_1\} \subset \{P_\theta : \theta \in \Omega\},$$

then, when the smaller model contains the true distribution, methods based on the smaller model will make better use of the information in the data about the true value in Ω_1 than will methods using the bigger model $\{P_\theta : \theta \in \Omega\}$. So there is a trade-off between taking too big a model and taking too precise a model. This is an issue that a statistician must always address.

We now consider some examples of distribution-free inferences. In some cases, the inferences have approximate sampling properties, while in other cases the inferences have exact sampling properties for very large models.

6.4.1 Method of Moments

Suppose we take $\{P_\theta : \theta \in \Omega\}$ to be the set of all distributions on R^1 that have their first l moments, and we want to make inferences about the moments

$$\mu_i = E_\theta(X^i),$$

for $i = 1, \ldots, l$ based on a sample $(x_1, \ldots, x_n)$. The natural sample analog of the population moment μ_i is the ith *sample moment*

$$m_i = \frac{1}{n}\sum_{j=1}^{n} x_j^i,$$

which would seem to be a sensible estimator.

In particular, we have that $E_\theta\,(M_i) = \mu_i$ for every $\theta \in \Omega$, so m_i is unbiased, and the weak and strong laws of large numbers establish that m_i converges to μ_i as n increases. Furthermore, the central limit theorem establishes that

$$\frac{M_i - \mu_i}{\sqrt{\mathrm{Var}_\theta\,(M_i)}} \xrightarrow{D} N(0,1)$$

as $n \to \infty$, provided that $\mathrm{Var}_\theta(M_i) < \infty$. Now, because $X_1, \ldots, X_n$ are i.i.d., we have that

$$\begin{aligned}\mathrm{Var}_\theta(M_i) &= \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{Var}_\theta(X_j^i) = \frac{1}{n}\,\mathrm{Var}_\theta\left(X_1^i\right) = \frac{1}{n}E_\theta((\left(X_1^i - \mu_i\right)^2) \\ &= \frac{1}{n}E_\theta(X_1^{2i} - 2\mu_i X_1^i + \mu_i^2) = \frac{1}{n}(\mu_{2i} - \mu_i^2),\end{aligned}$$

so we have that $\mathrm{Var}_\theta(M_i) < \infty$, provided that $i \le l/2$. In this case, we can estimate $\mu_{2i} - \mu_i^2$ by

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j^i - m_i\right)^2,$$

as we can simply treat $\left(x_1^i, \ldots, x_n^i\right)$ as a sample from a distribution with mean μ_i and variance $\mu_{2i} - \mu_i^2$. Problem 6.3.23 establishes that s_i^2 is an unbiased estimate of $\mathrm{Var}_\theta(M_i)$. So, as with inferences for the population mean based on the z-statistic, we have that

$$m_i \pm z_{(1+\gamma)/2}\frac{s_i}{\sqrt{n}}$$

is an approximate γ-confidence interval for μ_i whenever $i \le l/2$ and n is large. Also, we can test hypotheses $H_0 : \mu_i = \mu_{i0}$ in exactly the same fashion, as we did this for the population mean using the z-statistic.

Notice that the model $\{P_\theta : \theta \in \Omega\}$ is very large (all distributions on R^1 having their first $l/2$ moments finite), and these approximate inferences are appropriate for every distribution in the model. A cautionary note is that estimation of moments becomes more difficult as the order of the moments rises. Very large sample sizes are required for the accurate estimation of high-order moments.

The general *method of moments principle* allows us to make inference about characteristics that are functions of moments. This takes the following form:

> *Method of moments principle:* A function $\psi(\mu_1, \ldots, \mu_k)$ of the first $k \le l$ moments is estimated by $\psi(m_1, \ldots, m_k)$.

When ψ is continuously differentiable and nonzero at $\left(\mu_1, \ldots, \mu_k\right)$, and $k \le l/2$, then it can be proved that $\psi(M_1, \ldots, M_k)$ converges in distribution to a normal with mean given by $\psi(\mu_1, \ldots, \mu_k)$ and variance given by an expression involving the variances and covariances of $M_1, \ldots, M_k$ and the partial derivatives of ψ. We do not pursue this topic further here but note that, in the case $k = 1$ and $l = 2$, these conditions lead to the so-called *delta theorem*, which says that

$$\frac{\sqrt{n}\left(\psi(\bar{X}) - \psi(\mu_1)\right)}{\left|\psi'(\bar{X})\right| s} \overset{D}{\to} N(0, 1) \tag{6.4.1}$$

as $n \to \infty$, provided that ψ is continuously differentiable at μ_1 and $\psi'(\mu_1) \ne 0$; see *Approximation Theorems of Mathematical Statistics*, by R. J. Serfling (John Wiley & Sons, New York, 1980), for a proof of this result. This result provides approximate confidence intervals and tests for $\psi(\mu_1)$.

EXAMPLE 6.4.1 *Inference about a Characteristic Using the Method of Moments*
Suppose $(x_1, \ldots, x_n)$ is a sample from a distribution with unknown mean μ and variance σ^2, and we want to construct a γ-confidence interval for $\psi(\mu) = 1/\mu^2$. Then $\psi'(\mu) = -2/\mu^3$, so the delta theorem says that

$$\frac{\sqrt{n}\left(1/\bar{X}^2 - 1/\mu^2\right)}{2s/\bar{X}^3} \overset{D}{\to} N(0, 1)$$

as $n \to \infty$. Therefore,

$$\left(\frac{1}{\bar{x}}\right)^2 \pm 2\frac{s}{\sqrt{n}\bar{x}^3} z_{(1+\gamma)/2}$$

is an approximate γ-confidence interval for $\psi(\mu) = 1/\mu^2$.

Notice that if $\mu = 0$, then this confidence interval is not valid because ψ is not continuously differentiable at 0. So if you think the population mean could be 0, or even close to 0, this would not be an appropriate choice of confidence interval for ψ. ∎

6.4.2 Bootstrapping

Suppose that $\{P_\theta : \theta \in \Omega\}$ is the set of all distributions on R^1 and that $(x_1, \ldots, x_n)$ is a sample from some unknown distribution with cdf F_θ. Then the empirical distribution function

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, x]}(x_i),$$

introduced in Section 5.4.1, is a natural estimator of the cdf $F(x)$.

We have

$$E_\theta(\hat{F}(x)) = \frac{1}{n}\sum_{i=1}^{n} E_\theta(I_{(-\infty,x]}(X_i)) = \frac{1}{n}\sum_{i=1}^{n} F_\theta(x) = F_\theta(x)$$

for every $\theta \in \Omega$ so that $\hat{F}$ is unbiased for F_θ. The weak and strong laws of large numbers then establish the consistency of $\hat{F}(x)$ for $F_\theta(x)$ as $n \to \infty$. Observing that the $I_{(-\infty,x]}(x_i)$ constitute a sample from the Bernoulli$(F_\theta(x))$ distribution, we have that the standard error of $\hat{F}(x)$ is given by

$$\sqrt{\frac{\hat{F}(x)(1-\hat{F}(x))}{n}}.$$

These facts can be used to form approximate confidence intervals and test hypotheses for $F_\theta(x)$, just as in Examples 6.3.7 and 6.3.11.

Observe that $\hat{F}(x)$ prescribes a distribution on the set $\{x_1, \ldots, x_n\}$, e.g., if the sample values are distinct, this probability distribution puts mass $1/n$ on each x_i. Note that it is easy to sample a value from $\hat{F}$, as we just select a value from $\{x_1, \ldots, x_n\}$ where each point has probability $1/n$ of occurring. When the x_i are not distinct, then this is changed in an obvious way, namely, x_i has probability f_i/n, where f_i is the number of times x_i occurs in $x_1, \ldots, x_n$.

Suppose we are interested in estimating $\psi(\theta) = T(F_\theta)$, where T is a function of the distribution F_θ. We use this notation to emphasize that $\psi(\theta)$ corresponds to some characteristic of the distribution rather than just being an arbitrary mathematical function of θ. For example, $T(F_\theta)$ could be a moment of F_θ, a quantile of F_θ, etc.

Now suppose we have an estimator $\hat{\psi}(x_1, \ldots, x_n)$ that is being proposed for inferences about $\psi(\theta)$. Naturally, we are interested in the accuracy of $\hat{\psi}$, and we could choose to measure this by

$$\text{MSE}_\theta(\hat{\psi}) = \left(E_\theta(\hat{\psi}) - \psi(\theta)\right)^2 + \text{Var}_\theta(\hat{\psi}). \tag{6.4.2}$$

Then, to assess the accuracy of our estimate $\hat{\psi}(x_1, \ldots, x_n)$, we need to estimate (6.4.2).

When n is large, we expect $\hat{F}$ to be close to F_θ, so a natural estimate of $\psi(\theta)$ is $T(\hat{F})$, i.e., simply compute the same characteristic of the empirical distribution. This is the approach adopted in Chapter 5 when we discussed descriptive statistics. Then we estimate the square of the bias in $\hat{\psi}$ by

$$(\hat{\psi} - T(\hat{F}))^2. \tag{6.4.3}$$

To estimate the variance of $\hat{\psi}$, we use

$$\begin{aligned}\text{Var}_{\hat{F}}(\hat{\psi}) &= E_{\hat{F}}(\hat{\psi}^2) - E^2_{\hat{F}}(\hat{\psi}) \\ &= \frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n}\hat{\psi}^2(x_{i_1}, \ldots, x_{i_n}) - \left(\frac{1}{n^n}\sum_{i_1=1}^{n}\cdots\sum_{i_n=1}^{n}\hat{\psi}(x_{i_1}, \ldots, x_{i_n})\right)^2,\end{aligned} \tag{6.4.4}$$

i.e., we treat $x_1, \ldots, x_n$ as i.i.d. random values with cdf given by $\hat{F}$. So to calculate an estimate of (6.4.2), we simply have to calculate $\text{Var}_{\hat{F}}(\hat{\psi})$. This is rarely feasible, however, because the sums in (6.4.4) involve n^n terms. For even very modest sample sizes, like $n = 10$, this cannot be carried out, even on a computer.

The solution to this problem is to approximate (6.4.4) by drawing m independent samples of size n from $\hat{F}$, evaluating $\hat{\psi}$ for each of these samples to obtain

$\hat{\psi}_1, \ldots, \hat{\psi}_m$, and then using the sample variance

$$\widehat{\text{Var}}_{\hat{F}}(\hat{\psi}) = \frac{1}{m-1}\left\{\sum_{i=1}^{m}\hat{\psi}_i^2 - \left(\frac{1}{m}\sum_{i=1}^{m}\hat{\psi}_i\right)^2\right\} \tag{6.4.5}$$

as the estimate. The m samples from $\hat{F}$ are referred to as *bootstrap samples* or *resamples*, and this technique is referred to as *bootstrapping* or *resampling*. Combining (6.4.3) and (6.4.5) gives an estimate of $\text{MSE}_\theta(\hat{\psi})$. Furthermore, $m^{-1}\sum_{i=1}^{m}\hat{\psi}_i$ is called the *bootstrap mean*, and

$$\sqrt{\widehat{\text{Var}}_{\hat{F}}(\hat{\psi})}$$

is the *bootstrap standard error*. Note that the bootstrap standard error is a valid estimate of the error in $\hat{\psi}$ whenever $\hat{\psi}$ has little or no bias.

Consider the following example.

EXAMPLE 6.4.2 *The Sample Median as an Estimator of the Population Mean*
Suppose we want to estimate the location of a unimodal, symmetric distribution. While the sample mean might seem like the obvious choice for this, it turns out that for some distributions there are better estimators. This is because the distribution we are sampling may have long tails, i.e., may produce extreme values that are far from the center of the distribution. This implies that the sample average itself could be highly influenced by a few extreme observations and would thus be a poor estimate of the true mean.

Not all estimators suffer from this defect. For example, if we are sampling from a symmetric distribution, then either the sample mean or the sample median could serve as an estimator of the population mean. But, as we have previously discussed, the sample median is not influenced by extreme values, i.e., it does not change as we move the smallest (or largest) values away from the rest of the data, and this is not the case for the sample mean.

A problem with working with the sample median $\hat{x}_{0.5}$, rather than the sample mean $\bar{x}$, is that the sampling distribution for $\hat{x}_{0.5}$ is typically more difficult to study than that of $\bar{x}$. In this situation, bootstrapping becomes useful. If we are estimating the population mean $T(F_\theta)$ by using the sample median (which is appropriate when we know the distribution we were sampling from is symmetric), then the estimate of the squared bias in the sample median is given by

$$(\hat{\psi} - T(\hat{F}))^2 = \left(\hat{x}_{0.5} - \bar{x}\right)^2$$

because $\hat{\psi} = \hat{x}_{0.5}$ and $T(\hat{F}) = \bar{x}$ (the mean of the empirical distribution is $\bar{x}$). This should be close to 0, or else our assumption of a symmetric distribution would seem to be incorrect. To calculate (6.4.5), we have to generate m samples of size n from $\{x_1, \ldots, x_n\}$ (with replacement) and calculate $\hat{x}_{0.5}$ for each sample.

To illustrate, suppose we have a sample of size $n = 15$, given by the following table.

−2.0	−0.2	−5.2	−3.5	−3.9
−0.6	−4.3	−1.7	−9.5	1.6
−2.9	0.9	−1.0	−2.0	3.0

Then, using the definition of $\hat{x}_{0.5}$ given by (5.5.4) (denoted $\check{x}_{0.5}$ there), $\hat{\psi} = -2.000$ and $\bar{x} = -2.087$. The estimate of the squared bias (6.4.3) equals $(-2.000 + 2.087)^2 = 7.569 \times 10^{-3}$, which is appropriately small. Using a statistical package, we generated $m = 10^3$ samples of size $n = 15$ from the distribution that has probability $1/15$ at each of the sample points and obtained

$$\widehat{\mathrm{Var}}_{\hat{F}}(\hat{\psi}) = 0.770866.$$

Based on $m = 10^4$ samples, we obtained

$$\widehat{\mathrm{Var}}_{\hat{F}}(\hat{\psi}) = 0.718612,$$

and based on $m = 10^5$ samples we obtained

$$\widehat{\mathrm{Var}}_{\hat{F}}(\hat{\psi}) = 0.704928.$$

Because these estimates appear to be stabilizing, we take this as our estimate. So in this case, the bootstrap estimate of the MSE of the sample median at the true value of θ is given by

$$\widehat{\mathrm{MSE}}_\theta(\hat{\psi}) = 0.007569 + 0.704928 = 0.71250.$$

Note that the estimated MSE of the sample average is given by $s^2 = 0.62410$, so the sample mean and sample median appear to be providing similar accuracy in this problem. In Figure 6.4.1, we have plotted a density histogram of the sample medians obtained from the $m = 10^5$ bootstrap samples. Note that the histogram is very skewed. See Appendix B for more details on how these computations were carried out.

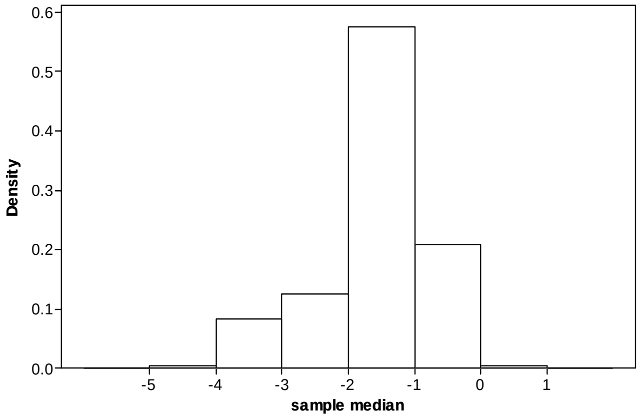

Figure 6.4.1: A density histogram of $m = 10^5$ sample medians, each obtained from a bootstrap sample of size $n = 15$ from the data in Example 6.4.2.

Even with the very small sample size here, it was necessary to use the computer to carry out our calculations. To evaluate (6.4.4) exactly would have required computing the median of 15^{15} (roughly 4.4×10^{17}) samples, which is clearly impossible even using a computer. So the bootstrap is a very useful device. ∎

The validity of the bootstrapping technique depends on $\hat{\psi}$ having its first two moments. So the family $\{P_\theta : \theta \in \Omega\}$ must be appropriately restricted, but we can see that the technique is very general.

Broadly speaking, it is not clear how to choose m. Perhaps the most direct method is to implement bootstrapping for successively higher values of m and stop when we see that the results stabilize for several values. This is what we did in Example 6.4.2, but it must be acknowledged that this approach is not foolproof, as we could have a sample $(x_1, \ldots, x_n)$ such that the estimate (6.4.5) is very slowly convergent.

Bootstrap Confidence Intervals

Bootstrap methods have also been devised to obtain approximate γ-confidence intervals for characteristics such as $\psi(\theta) = T(F_\theta)$. One very simple method, is to simply form the *bootstrap t γ-confidence interval*

$$\hat{\psi} \pm t_{(1+\gamma)/2}(n-1)\sqrt{\widehat{\mathrm{Var}}_{\hat{F}}(\hat{\psi})},$$

where $t_{(1+\gamma)/2}(n-1)$ is the $(1+\gamma)/2$th quantile of the $t(n-1)$ distribution. Another possibility is to compute a *bootstrap percentile confidence interval* given by

$$(\hat{\psi}_{(1+\gamma)/2}, \hat{\psi}_{(1+\gamma)/2}),$$

where $\hat{\psi}_p$ denotes the pth empirical quantile of $\hat{\psi}$ in the bootstrap sample of m.

It should be noted that to be applicable, these intervals require some conditions to hold. In particular, $\hat{\psi}$ should be at least approximately unbiased for $\psi(\theta)$ and the bootstrap distribution should be approximately normal. Looking at the plot of the bootstrap distribution in Figure 6.4.1 we can see that the median does not have an approximately normal bootstrap distribution, so these intervals are not applicable with the median.

Consider the following example.

EXAMPLE 6.4.3 *The 0.25-Trimmed Mean as an Estimator of the Population Mean*
One of the virtues of the sample median as an estimator of the population mean is that it is not affected by extreme values in the sample. On the other hand, the sample median discards all but one or two of the data values and so seems to be discarding a lot of information. Estimators known as trimmed means can be seen as an attempt at retaining the virtues of the median while at the same time not discarding too much information. Let $\lfloor x \rfloor$ denote the greatest integer less than or equal to $x \in R^1$.

Definition 6.4.1 For $\alpha \in [0, 1]$, a *sample α-trimmed mean* is given by

$$\bar{x}_\alpha = \frac{1}{n - 2\lfloor \alpha n \rfloor} \sum_{i=\lfloor \alpha n \rfloor + 1}^{n - \lfloor \alpha n \rfloor} x_{(i)},$$

where $x_{(i)}$ is the ith-order statistic.

Thus for a sample α-trimmed mean, we toss out (approximately) αn of the smallest data values and αn of the largest data values and calculate the average of the $n - 2\alpha n$ of the data values remaining. We need the greatest integer function because in general, αn will not be an integer. Note that the sample mean arises with $\alpha = 0$ and the sample median arises with $\alpha = 0.5$.

For the data in Example 6.4.1 and $\alpha = 0.25$, we have $(0.25)15 = 3.75$, so we discard the three smallest and three largest observations leaving the nine data values $-3.9, -3.5, -2.9, -2.0, -2.0, -1.7, -1.0, -0.6, -0.2$. The average of these nine values gives $\hat{\psi} = \bar{x}_{0.25} = -1.97778$, which we note is close to both the sample median and the sample mean.

Now suppose we use a 0.25-trimmed mean as an estimator $\hat{\psi}$ of a population mean where we believe the population distribution is symmetric. Consider the data in Example 6.4.1 and suppose we generated $m = 10^4$ bootstrap samples. We have plotted a histogram of the 10^4 values of $\hat{\psi}$ in Figure 6.4.2. Notice that it is very normal looking, so we feel justified in using the confidence intervals associated with the bootstrap. In this case, we obtained

$$\sqrt{\widehat{\text{Var}}_{\hat{F}}(\hat{\psi})} = 0.7380,$$

so the bootstrap t 0.95-confidence interval for the mean is given by $-1.97778 + (2.14479)(0.7380) \simeq (-3.6, -0.4)$. Sorting the bootstrap sample gives a bootstrap percentile 0.95-confidence interval as $(-3.36667, -0.488889) \simeq (-3.4, -0.5)$ which shows that the two intervals are very similar.

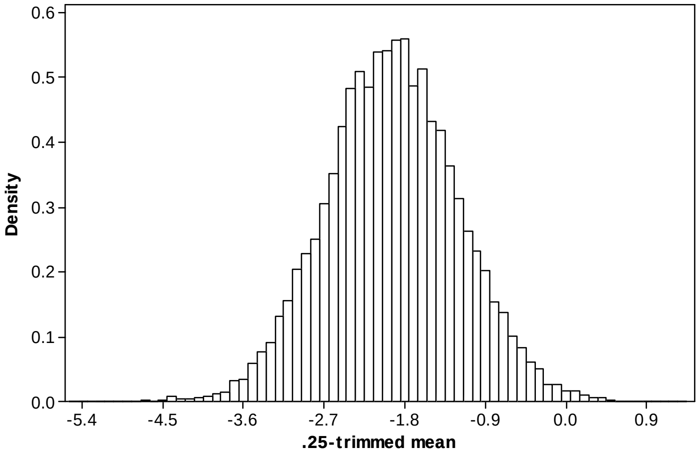

Figure 6.4.2: A density histogram of $m = 10^4$ sample 0.25-trimmed means, each obtained from a bootstrap sample of size $n = 15$ from the data in Example 6.4.3

■

More details about the bootstrap can be found in *An Introduction to the Bootstrap*, by B. Efron and R. J. Tibshirani (Chapman and Hall, New York, 1993).

6.4.3 The Sign Statistic and Inferences about Quantiles

Suppose that $\{P_\theta : \theta \in \Omega\}$ is the set of all distributions on R^1 such that the associated distribution functions are continuous. Suppose we want to make inferences about a pth quantile of P_θ. We denote this quantile by $x_p(\theta)$ so that, when the distribution function associated with P_θ is denoted by F_θ, we have $p = F_\theta(x_p(\theta))$. Note that continuity implies there is always a solution in x to $p = F_\theta(x)$, and that $x_p(\theta)$ is the smallest solution.

Recall the definitions and discussion of estimation of these quantities in Example 5.5.2 based on a sample $(x_1, \ldots, x_n)$. For simplicity, let us restrict attention to the cases where $p = i/n$ for some $i \in \{1, \ldots, n\}$. In this case, we have that $\hat{x}_p = x_{(i)}$ is the natural estimate of x_p.

Now consider assessing the evidence in the data concerning the hypothesis $H_0 : x_p(\theta) = x_0$. For testing this hypothesis, we can use the *sign test statistic*, given by $S = \sum_{i=1}^{n} I_{(-\infty,x_0]}(x_i)$. So S is the number of sample values less than or equal to x_0.

Notice that when H_0 is true, $I_{(-\infty,x_0]}(x_1), \ldots, I_{(-\infty,x_0]}(x_n)$ is a sample from the Bernoulli(p) distribution. This implies that, when H_0 is true, $S \sim$ Binomial(n, p).

Therefore, we can test H_0 by computing the observed value of S, denoted S_o, and seeing whether this value lies in a region of low probability for the Binomial(n, p) distribution. Because the binomial distribution is unimodal, the regions of low probability correspond to the left and right tails of this distribution. See, for example, Figure 6.4.3, where we have plotted the probability function of a Binomial(20, 0.7) distribution.

The P-value is therefore obtained by computing the probability of the set

$$\left\{ i : \binom{n}{i} p^i (1-p)^{n-i} \le \binom{n}{S_o} p^{S_o} (1-p)^{n-S_o} \right\} \tag{6.4.6}$$

using the Binomial(n, p) probability distribution. This is a measure of how far out in the tails the observed value S_o is (see Figure 6.4.3). Notice that this P-value is completely independent of θ and is thus valid for the entire model. Tables of binomial probabilities (Table D.6 in Appendix D), or built-in functions available in most statistical packages, can be used to calculate this P-value.

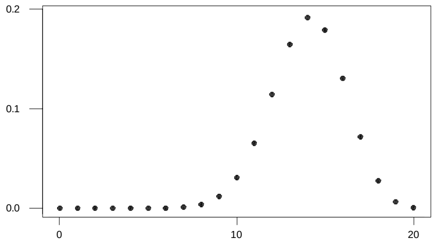

Figure 6.4.3: Plot of the Binomial(20, 0.7) probability function.

When n is large, we have that, under H_0,

$$Z = \frac{S - np}{\sqrt{np(1-p)}} \xrightarrow{D} N(0,1)$$

as $n \to \infty$. Therefore, an approximate P-value is given by

$$2\left\{1 - \Phi\left(\left|\frac{S_o - 0.5 - np}{\sqrt{np\,(1-p)}}\right|\right)\right\}$$

(as in Example 6.3.11), where we have replaced S_o by $S_o - 0.5$ as a correction for continuity (see Example 4.4.9 for discussion of the correction for continuity).

A special case arises when $p = 1/2$, i.e., when we are making inferences about an unknown population median $x_{0.5}(\theta)$. In this case, the distribution of S under H_0 is Binomial$(n, 1/2)$. Because the Binomial$(n, 1/2)$ is unimodal and symmetrical about $n/2$, (6.4.6) becomes

$$\{i : |S_o - n/2| \le |i - n/2|\}\,.$$

If we want a γ-confidence interval for $x_{0.5}(\theta)$, then we can use the equivalence between tests, which we always reject when the P-value is less than or equal to $1 - \gamma$, and γ-confidence intervals (see Example 6.3.12). For this, let j be the smallest integer greater than $n/2$ satisfying

$$P\left(\{i : |i - n/2| \ge j - n/2\}\right) \le 1 - \gamma\,, \tag{6.4.7}$$

where P is the Binomial$(n, 1/2)$ distribution. If $S \in \{i : |i - n/2| \ge j - n/2\}$, we will reject $H_0 : x_{0.5}(\theta) = x_0$ at the $1 - \gamma$ level and will not otherwise. This leads to the γ-confidence interval, namely, the set of all those values $x_{0.5}$ such that the null hypothesis $H_0 : x_{0.5}(\theta) = x_{0.5}$ is not rejected at the $1 - \gamma$ level, equaling

$$\begin{aligned} C(x_1, \ldots, x_n) &= \left\{x_0 : \left|\sum_{i=1}^{n} I_{(-\infty, x_0]}(x_i) - n/2\right| < j - n/2\right\} \\ &= \left\{x_0 : n - j < \sum_{i=1}^{n} I_{(-\infty, x_0]}(x_i) < j\right\} = [x_{(n-j+1)}, x_{(j)}) \end{aligned} \tag{6.4.8}$$

because, for example, $n - j < \sum_{i=1}^{n} I_{(-\infty, x_0]}(x_i)$ if and only if $x_0 \ge x_{(n-j+1)}$.

EXAMPLE 6.4.4 *Application of the Sign Test*
Suppose we have the following sample of size $n = 10$ from a continuous random variable X, and we wish to test the hypothesis $H_0 : x_{0.5}(\theta) = 0$.

0.44	−0.06	0.43	−0.16	−2.13
1.15	1.08	5.67	−4.97	0.11

The boxplot in Figure 6.4.4 indicates that it is very unlikely that this sample came from a normal distribution, as there are two extreme observations. So it is appropriate to measure the location of the distribution of X by the median.

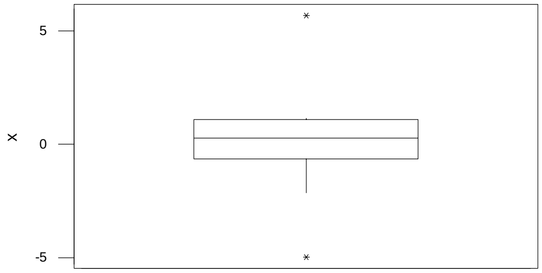

Figure 6.4.4: Boxplot of the data in Example 6.4.4.

In this case, the sample median (using (5.5.4)) is given by $(0.11 + 0.43)/2 = 0.27$. The sign statistic for the null is given by

$$S = \sum_{i=1}^{10} I_{(-\infty,0]}(x_i) = 4.$$

The P-value is given by

$$P(\{i : |4-5| \le |i-5|\}) = P(\{i : |i-5| \ge 1\}) = 1 - P(\{i : |i-5| < 1\})$$
$$= 1 - P(\{5\}) = 1 - \binom{10}{5}\left(\frac{1}{2}\right)^{10} = 1 - 0.24609 = 0.75391,$$

and we have no reason to reject the null hypothesis.

Now suppose that we want a 0.95-confidence interval for the median. Using software (or Table D.6), we calculate

$$\begin{array}{ll} \binom{10}{5}\left(\frac{1}{2}\right)^{10} = 0.24609 & \binom{10}{4}\left(\frac{1}{2}\right)^{10} = 0.20508 \\ \binom{10}{3}\left(\frac{1}{2}\right)^{10} = 0.11719 & \binom{10}{2}\left(\frac{1}{2}\right)^{10} = 4.3945 \times 10^{-2} \\ \binom{10}{1}\left(\frac{1}{2}\right)^{10} = 9.7656 \times 10^{-3} & \binom{10}{0}\left(\frac{1}{2}\right)^{10} = 9.7656 \times 10^{-4}. \end{array}$$

We will use these values to compute the value of j in (6.4.7).

We can use the symmetry of the Binomial(10, 1/2) distribution about $n/2$ to compute the values of $P(\{i : |i - n/2| \ge j - n/2\})$ as follows. For $j = 10$, we have that (6.4.7) equals

$$P(\{i : |i-5| \ge 5\}) = P(\{0, 10\}) = 2\binom{10}{0}\left(\frac{1}{2}\right)^{10} = 1.9531 \times 10^{-3},$$

and note that $1.9531 \times 10^{-3} < 1 - 0.95 = 0.05$. For $j = 9$, we have that (6.4.7) equals

$$\begin{aligned} P(\{i : |i-5| \ge 4\}) &= P(\{0, 1, 9, 10\}) = 2\binom{10}{0}\left(\frac{1}{2}\right)^{10} + 2\binom{10}{1}\left(\frac{1}{2}\right)^{10} \\ &= 2.148\,4 \times 10^{-2}, \end{aligned}$$

which is also less than 0.05. For $j = 8$, we have that (6.4.7) equals

$$\begin{aligned} P(\{i : |i-5| \geq 3\}) &= P(\{0, 1, 2, 8, 9, 10\}) \\ &= 2\binom{10}{0}\left(\frac{1}{2}\right)^{10} + 2\binom{10}{1}\left(\frac{1}{2}\right)^{10} + 2\binom{10}{2}\left(\frac{1}{2}\right)^{10} \\ &= 0.10938, \end{aligned}$$

and this is greater than 0.05. Therefore, the appropriate value is $j = 9$, and a 0.95-confidence interval for the median is given by $\left[x_{(2)}, x_{(9)}\right) = [-0.16, 1.15)$. ∎

There are many other distribution-free methods for a variety of statistical situations. While some of these are discussed in the problems, we leave a thorough study of such methods to further courses in statistics.

Summary of Section 6.4

- Distribution-free methods of statistical inference are appropriate methods when we feel we can make only very minimal assumptions about the distribution from which we are sampling.
- The method of moments, bootstrapping, and methods of inference based on the sign statistic are three distribution-free methods that are applicable in different circumstances.

EXERCISES

6.4.1 Suppose we obtained the following sample from a distribution that we know has its first six moments. Determine an approximate 0.95-confidence interval for μ_3.

3.27	−1.24	3.97	2.25	3.47	−0.09	7.45	6.20	3.74	4.12
1.42	2.75	−1.48	4.97	8.00	3.26	0.15	−3.64	4.88	4.55

6.4.2 Determine the method of moments estimator of the population variance. Is this estimator unbiased for the population variance? Justify your answer.

6.4.3 (*Coefficient of variation*) The coefficient of variation for a population measurement with nonzero mean is given by σ/μ, where μ is the population mean and σ is the population standard deviation. What is the method of moments estimate of the coefficient of variation? Prove that the coefficient of variation is invariant under rescalings of the distribution, i.e., under transformations of the form $T(x) = cx$ for constant $c > 0$. It is this invariance that leads to the coefficient of variation being an appropriate measure of sampling variability in certain problems, as it is independent of the units we use for the measurement.

6.4.4 For the context described in Exercise 6.4.1, determine an approximate 0.95-confidence interval for $\exp(\mu_1)$.

6.4.5 Verify that the third moment of an $N(\mu, \sigma^2)$ distribution is given by $\mu_3 = \mu^3 + 3\mu\sigma^2$. Because the normal distribution is specified by its first two moments, any characteristic of the normal distribution can be estimated by simply plugging in the MLE estimates of μ and σ^2. Compare the method of moments estimator of μ_3 with this plug-in MLE estimator, i.e., determine whether they are the same or not.

6.4.6 Suppose we have the sample data 1.48, 4.10, 2.02, 56.59, 2.98, 1.51, 76.49, 50.25, 43.52, 2.96. Consider this as a sample from a normal distribution with unknown mean and variance, and assess the hypothesis that the population median (which is the same as the mean in this case) is 3. Also carry out a sign test that the population median is 3 and compare the results. Plot a boxplot for these data. Does this support the assumption that we are sampling from a normal distribution? Which test do you think is more appropriate? Justify your answer.

6.4.7 Determine the empirical distribution function based on the sample given below.

1.06	−1.28	0.40	1.36	−0.35
−1.42	0.44	−0.58	−0.24	−1.34
0.00	−1.02	−1.35	2.05	1.06
0.98	0.38	2.13	−0.03	−1.29

Using the empirical cdf, determine the sample median, the first and third quartiles, and the interquartile range. What is your estimate of $F(2)$?

6.4.8 Suppose you obtain the sample of $n = 3$ distinct values given by 1, 2, and 3.

(a) Write down all possible bootstrap samples.

(b) If you are bootstrapping the sample median, what are the possible values for the sample median for a bootstrap sample?

(c) If you are bootstrapping the sample mean, what are the possible values for the sample mean for a bootstrap sample?

(d) What do you conclude about the bootstrap distribution of the sample median compared to the bootstrap distribution of the sample mean?

6.4.9 Explain why the central limit theorem justifies saying that the bootstrap distribution of the sample mean is approximately normal when n and m are large. What result justifies the approximate normality of the bootstrap distribution of a function of the sample mean under certain conditions?

6.4.10 For the data in Exercise 6.4.1, determine an approximate 0.95-confidence interval for the population median when we assume the distribution we are sampling from is symmetric with finite first and second moments. (Hint: Use large sample results.)

6.4.11 Suppose you have a sample of n distinct values and are interested in the bootstrap distribution of the *sample range* given by $x_{(n)} - x_{(1)}$. What is the maximum number of values that this statistic can take over all bootstrap samples? What are the largest and smallest values that the sample range can take in a bootstrap sample? Do you think the bootstrap distribution of the sample range will be approximately normal? Justify your answer.

6.4.12 Suppose you obtain the data 1.1, −1.0, 1.1, 3.1, 2.2, and 3.1. How many distinct bootstrap samples are there?

COMPUTER EXERCISES

6.4.13 For the data of Exercise 6.4.7, assess the hypothesis that the population median is 0. State a 0.95-confidence interval for the population median. What is the exact coverage probability of this interval?

6.4.14 For the data of Exercise 6.4.7, assess the hypothesis that the first quartile of the distribution we are sampling from is -1.0.

6.4.15 With a bootstrap sample size of $m = 1000$, use bootstrapping to estimate the MSE of the plug-in MLE estimator of μ_3 for the normal distribution, using the sample data in Exercise 6.4.1. Determine whether $m = 1000$ is a large enough sample for accurate results.

6.4.16 For the data of Exercise 6.4.1, use the plug-in MLE to estimate the first quartile of an $N(\mu, \sigma^2)$ distribution. Use bootstrapping to estimate the MSE of this estimate for $m = 10^3$ and $m = 10^4$ (use (5.5.3) to compute the first quartile of the empirical distribution).

6.4.17 For the data of Exercise 6.4.1, use the plug-in MLE to estimate $F(3)$ for an $N(\mu, \sigma^2)$ distribution. Use bootstrapping to estimate the MSE of this estimate for $m = 10^3$ and $m = 10^4$.

6.4.18 For the data of Exercise 6.4.1, form a 0.95-confidence interval for μ assuming that this is a sample from an $N(\mu, \sigma^2)$ distribution. Also compute a 0.95-confidence interval for μ based on the sign statistic, a bootstrap t 0.95-confidence interval, and a bootstrap percentile 0.95-confidence interval using $m = 10^3$ for the bootstrapping. Compare the four intervals.

6.4.19 For the data of Exercise 6.4.1, use the plug-in MLE to estimate the first *quintile*, i.e., $x_{0.2}$, of an $N(\mu, \sigma^2)$ distribution. Plot a density histogram estimate of the bootstrap distribution of this estimator for $m = 10^3$ and compute a bootstrap t 0.95-confidence interval for $x_{0.2}$, if you think it is appropriate.

6.4.20 For the data of Exercise 6.4.1, use the plug-in MLE to estimate μ_3 of an $N(\mu, \sigma^2)$ distribution. Plot a density histogram estimate of the bootstrap distribution of this estimator for $m = 10^3$ and compute a bootstrap percentile 0.95-confidence interval for μ_3, if you think it is appropriate.

PROBLEMS

6.4.21 Prove that when $(x_1, \ldots, x_n)$ is a sample of distinct values from a distribution on R^1, then the ith moment of the empirical distribution on R^1 (i.e., the distribution with cdf given by $\hat{F}$) is m_i.

6.4.22 Suppose that $(x_1, \ldots, x_n)$ is a sample from a distribution on R^1. Determine the general form of the ith moment of $\hat{F}$, i.e., in contrast to Problem 6.4.21, we are now allowing for several of the data values to be equal.

6.4.23 (*Variance stabilizing transformations*) From the delta theorem, we have that $\psi(M_1)$ is asymptotically normal with mean $\psi(\mu_1)$ and variance $(\psi'(\mu_1))^2\sigma^2/n$ when ψ is continuously differentiable, $\psi'(\mu_1) \neq 0$, and M_1 is asymptotically normal with mean μ_1 and variance σ^2/n. In some applications, it is important to choose the transformation ψ so that the asymptotic variance does not depend on the mean μ_1, i.e.,

$(\psi'(\mu_1))^2\sigma^2$ is constant as μ_1 varies (note that σ^2 may change as μ_1 changes). Such transformations are known as variance stabilizing transformations.

(a) If we are sampling from a Poisson(λ) distribution, then show that $\psi(x) = \sqrt{x}$ is variance stabilizing.

(b) If we are sampling from a Bernoulli(θ) distribution, show that $\psi(x) = \arcsin\sqrt{x}$ is variance stabilizing.

(c) If we are sampling from a distribution on $(0, \infty)$ whose variance is proportional to the square of its mean (like the Gamma(α, β) distribution), then show that $\psi(x) = \ln(x)$ is variance stabilizing.

CHALLENGES

6.4.24 Suppose that X has an absolutely continuous distribution on R^1 with density f that is symmetrical about its median. Assuming that the median is 0, prove that $|X|$ and

$$\text{sgn}(X) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

are independent, with $|X|$ having density $2f$ and sgn(X) uniformly distributed on $\{-1, 1\}$.

6.4.25 (*Fisher signed deviation statistic*) Suppose that $(x_1, \ldots, x_n)$ is a sample from an absolutely continuous distribution on R^1 with density that is symmetrical about its median. Suppose we want to assess the hypothesis $H_0 : x_{0.5}(\theta) = x_0$.

One possibility for this is to use the Fisher signed deviation test based on the statistic S^+. The observed value of S^+ is given by $S_o^+ = \sum_{i=1}^n |x_i - x_0| \,\text{sgn}\,(x_i - x_0)$. We then assess H_0 by comparing S_o^+ with the conditional distribution of S^+ given the absolute deviations $|x_1 - x_0|, \ldots, |x_n - x_0|$. If a value S_o^+ occurs near the smallest or largest possible value for S^+, under this conditional distribution, then we assert that we have evidence against H_0. We measure this by computing the P-value given by the conditional probability of obtaining a value as far, or farther, from the center of the conditional distribution of S^+ using the conditional mean as the center. This is an example of a *randomization test*, as the distribution for the test statistic is determined by randomly modifying the observed data (in this case, by randomly changing the signs of the deviations of the x_i from x_0).

(a) Prove that $S_o^+ = n(\bar{x} - x_0)$.

(b) Prove that the P-value described above does not depend on which distribution we are sampling from in the model. Prove that the conditional mean of S^+ is 0 and the conditional distribution of S^+ is symmetric about this value.

(c) Use the Fisher signed deviation test statistic to assess the hypothesis $H_0 : x_{0.5}(\theta) = 2$ when the data are 2.2, 1.5, 3.4, 0.4, 5.3, 4.3, 2.1, with the results declared to be statistically significant if the P-value is less than or equal to 0.05. (Hint: Based on the results obtained in part (b), you need only compute probabilities for the extreme values of S^+.)

(d) Show that using the Fisher signed deviation test statistic to assess the hypothesis $H_0 : x_{0.5}(\theta) = x_0$ is equivalent to the following randomized t-test statistic hypothesis assessment procedure. For this, we compute the conditional distribution of

$$T = \frac{\left(\bar{X} - x_0\right)}{S/\sqrt{n}}$$

when the $|X_i - x_0| = |x_i - x_0|$ are fixed and the $\operatorname{sgn}(X_i - x_0)$ are i.i.d. uniform on $\{-1, 1\}$. Compare the observed value of the t-statistic with this distribution, as we did for the Fisher signed deviation test statistic. (Hint: Show that $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - x_0)^2 - n (\bar{x} - x_0)^2$ and that large absolute values of T correspond to large absolute values of S^+.)

6.5 Large Sample Behavior of the MLE (Advanced)

As we saw in Examples 6.3.7 and 6.3.11, implementing exact sampling procedures based on the MLE can be difficult. In those examples, because the MLE was the sample average and we could use the central limit theorem, large sample theory allowed us to work out approximate procedures. In fact, there is some general large sample theory available for the MLE that allows us to obtain approximate sampling inferences. This is the content of this section. The results we develop are all for the case when θ is one-dimensional. Similar results exist for the higher-dimensional problems, but we leave those to a later course.

In Section 6.3, the basic issue was the need to measure the accuracy of the MLE. One approach is to plot the likelihood and examine how concentrated it is about its peak, with a more highly concentrated likelihood implying greater accuracy for the MLE. There are several problems with this. In particular, the appearance of the likelihood will depend greatly on how we choose the scales for the axes. With appropriate choices, we can make a likelihood look as concentrated or as diffuse as we want. Also, when θ is more than two-dimensional, we cannot even plot the likelihood. One solution, when the likelihood is a smooth function of θ, is to compute a numerical measure of how concentrated the log-likelihood is at its peak. The quantity typically used for this is called the observed Fisher information.

Definition 6.5.1 The *observed Fisher information* is given by

$$\hat{I}(s) = -\left.\frac{\partial^2 l(\theta \mid s)}{\partial \theta^2}\right|_{\theta=\hat{\theta}(s)}, \tag{6.5.1}$$

where $\hat{\theta}(s)$ is the MLE.

The larger the observed Fisher information is, the more peaked the likelihood function is at its maximum value. We will show that the observed Fisher information is estimating a quantity of considerable importance in statistical inference.

Suppose that response X is real-valued, θ is real-valued, and the model $\{f_\theta : \theta \in \Omega\}$ satisfies the following regularity conditions:

$$\frac{\partial^2 \ln f_\theta(x)}{\partial\theta^2} \text{ exists for each } x, \tag{6.5.2}$$

$$E_\theta(S(\theta \mid X)) = \int_{-\infty}^{\infty} \frac{\partial \ln f_\theta(x)}{\partial\theta} f_\theta(x)\,dx = 0, \tag{6.5.3}$$

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}\left\{\frac{\partial \ln f_\theta(x)}{\partial\theta} f_\theta(x)\right\} dx = 0, \tag{6.5.4}$$

and

$$\int_{-\infty}^{\infty} \left|\frac{\partial^2 \ln f_\theta(x)}{\partial\theta^2}\right| f_\theta(x)\,dx < \infty. \tag{6.5.5}$$

Note that we have

$$\frac{\partial f_\theta(x)}{\partial\theta} = \frac{\partial \ln f_\theta(x)}{\partial\theta} f_\theta(x),$$

so we can write (6.5.3) equivalently as

$$\int_{-\infty}^{\infty} \frac{\partial f_\theta(x)}{\partial\theta}\,dx = 0.$$

Also note that (6.5.4) can be written as

$$\begin{aligned}0 &= \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}\left\{\frac{\partial l(\theta \mid x)}{\partial\theta} f_\theta(x)\right\} dx \\ &= \int_{-\infty}^{\infty} \left\{\frac{\partial^2 l(\theta \mid x)}{\partial\theta^2} + \left(\frac{\partial l(\theta \mid x)}{\partial\theta}\right)^2\right\} f_\theta(x)\,dx \\ &= \int_{-\infty}^{\infty} \left\{\frac{\partial^2 l(\theta \mid x)}{\partial\theta^2} + S^2(\theta \mid x)\right\} f_\theta(x)\,dx = E_\theta\left(\frac{\partial^2 l(\theta \mid X)}{\partial\theta^2} + S^2(\theta \mid X)\right).\end{aligned}$$

This together with (6.5.3) and (6.5.5), implies that we can write (6.5.4) equivalently as

$$\operatorname{Var}_\theta(S(\theta \mid X)) = E_\theta(S^2(\theta \mid X)) = E_\theta\left(-\frac{\partial^2}{\partial\theta^2} l(\theta \mid X)\right).$$

We give a name to the quantity on the left.

Definition 6.5.2 The function $I(\theta) = \operatorname{Var}_\theta(S(\theta \mid X))$, is called the *Fisher information* of the model.

Our developments above have proven the following result.

Theorem 6.5.1 If (6.5.2) and (6.5.3) are satisfied, then $E_\theta(S(\theta \mid X)) = 0$. If, in addition, (6.5.4) and (6.5.5) are satisfied, then

$$I(\theta) = \operatorname{Var}_\theta(S(\theta \mid X)) = E_\theta\left(-\frac{\partial^2 l(\theta \mid X)}{\partial\theta^2}\right).$$

Now we see why $\hat{I}$ is called the observed Fisher information, as it is a natural estimate of the Fisher information at the true value θ. We note that there is another natural estimate of the Fisher information at the true value, given by $I(\hat{\theta})$. We call this the *plug-in Fisher information*.

When we have a sample $(x_1, \ldots, x_n)$ from f_θ, then

$$S(\theta \mid x_1, \ldots, x_n) = \frac{\partial}{\partial\theta} \ln \prod_{i=1}^{n} f_\theta(x_i) = \sum_{i=1}^{n} \frac{\partial \ln f_\theta(x_i)}{\partial\theta} = \sum_{i=1}^{n} S(\theta \mid x_i).$$

So, if (6.5.3) holds for the basic model, then $E_\theta(S(\theta \mid X_1, \ldots, X_n)) = 0$ and (6.5.3) also holds for the sampling model. Furthermore, if (6.5.4) holds for the basic model, then

$$\begin{aligned} 0 &= \sum_{i=1}^{n} E_\theta\left(\frac{\partial^2}{\partial\theta^2} \ln f_\theta(X_i)\right) + \sum_{i=1}^{n} E_\theta(S^2(\theta \mid X_i)) \\ &= E_\theta\left(\frac{\partial^2}{\partial\theta^2} l(\theta \mid X_1, \ldots, X_n)\right) + \mathrm{Var}_\theta(S(\theta \mid X_1, \ldots, X_n)), \end{aligned}$$

which implies

$$\mathrm{Var}_\theta(S(\theta \mid X_1, \ldots, X_n)) = -E_\theta\left(\frac{\partial^2}{\partial\theta^2} l(\theta \mid X_1, \ldots, X_n)\right) = nI(\theta),$$

because $l(\theta \mid x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f_\theta(x_i)$. Therefore, (6.5.4) holds for the sampling model as well, and the Fisher information for the sampling model is given by the sample size times the Fisher information for the basic model. We have established the following result.

Corollary 6.5.1 Under i.i.d. sampling from a model with Fisher information $I(\theta)$. the Fisher information for a sample of size n is given by $nI(\theta)$.

The conditions necessary for Theorem 6.5.1 to apply do not hold in general and have to be checked in each example. There are, however, many models where these conditions do hold.

EXAMPLE 6.5.1 *Nonexistence of the Fisher Information*
If $X \sim U[0, \theta]$, then $f_\theta(x) = \theta^{-1} I_{[0,\theta]}(x)$, which is not differentiable at $\theta = x$ for any x. Indeed, if we ignored the lack of differentiability at $\theta = x$ and wrote

$$\frac{\partial f_\theta(x)}{\partial\theta} = -\frac{1}{\theta^2} I_{[0,\theta]}(x),$$

then

$$\int_{-\infty}^{\infty} \frac{\partial f_\theta(x)}{\partial\theta}\, dx = -\int_{-\infty}^{\infty} \frac{1}{\theta^2} I_{[0,\theta]}(x)\, dx = -\frac{1}{\theta} \neq 0.$$

So we cannot define the Fisher information for this model. ■

EXAMPLE 6.5.2 *Location Normal*

Suppose we have a sample $(x_1, \ldots, x_n)$ from an $N(\theta, \sigma_0^2)$ distribution where $\theta \in R^1$ is unknown and σ_0^2 is known. We saw in Example 6.2.2 that

$$S(\theta \mid x_1, \ldots, x_n) = \frac{n}{\sigma_0^2}(\bar{x} - \theta)$$

and therefore

$$\frac{\partial^2}{\partial\theta^2} l(\theta \mid x_1, \ldots, x_n) = -\frac{n}{\sigma_0^2},$$

$$nI(\theta) = E_\theta\left(-\frac{\partial^2}{\partial\theta^2} l(\theta \mid X_1, \ldots, X_n)\right) = \frac{n}{\sigma_0^2}.$$

We also determined in Example 6.2.2 that the MLE is given by $\hat{\theta}(x_1, \ldots, x_n) = \bar{x}$. Then the plug-in Fisher information is

$$nI(\bar{x}) = \frac{n}{\sigma_0^2},$$

while the observed Fisher information is

$$\hat{I}(x_1, \ldots, x_n) = -\left.\frac{\partial^2 l(\theta \mid x_1, \ldots, x_n)}{\partial\theta^2}\right|_{\theta=\bar{x}} = \frac{n}{\sigma_0^2}.$$

In this case, there is no need to estimate the Fisher information, but it is comforting that both of our estimates give the exact value. ▮

We now state, without proof, some theorems about the large sample behavior of the MLE under repeated sampling from the model. First, we have a result concerning the consistency of the MLE as an estimator of the true value of θ.

Theorem 6.5.2 Under regularity conditions (like those specified above) for the model $\{f_\theta : \theta \in \Omega\}$, the MLE $\hat{\theta}$ exists a.s. and $\hat{\theta} \overset{a.s.}{\to} \theta$ as $n \to \infty$.

PROOF See *Approximation Theorems of Mathematical Statistics*, by R. J. Serfling (John Wiley & Sons, New York, 1980), for the proof of this result. ▮

We see that Theorem 6.5.2 serves as a kind of strong law for the MLE. It also turns out that when the sample size is large, the sampling distribution of the MLE is approximately normal.

Theorem 6.5.3 Under regularity conditions (like those specified above) for the model $\{f_\theta : \theta \in \Omega\}$, then $(nI(\theta))^{1/2}(\hat{\theta} - \theta) \overset{D}{\to} N(0, 1)$ as $n \to \infty$.

PROOF See *Approximation Theorems of Mathematical Statistics*, by R. J. Serfling (John Wiley & Sons, New York, 1980), for the proof of this result. ▮

We see that Theorem 6.5.3 serves as a kind of central limit theorem for the MLE. To make this result fully useful to us for inference, we need the following corollary to this theorem.

Corollary 6.5.2 When I is a continuous function of θ, then

$$\left(nI(\hat{\theta})\right)^{1/2}\left(\hat{\theta}-\theta\right)\overset{D}{\to}N(0,1).$$

In Corollary 6.5.2, we have estimated the Fisher information $I(\theta)$ by the plug-in Fisher estimation $I(\hat{\theta})$. Often it is very difficult to evaluate the function I. In such a case, we instead estimate $nI(\theta)$ by the observed Fisher information $\hat{I}(x_1,\ldots,x_n)$. A result such as Corollary 6.5.2 again holds in this case.

From Corollary 6.5.2, we can devise large sample approximate inference methods based on the MLE. For example, the approximate standard error of the MLE is

$$(nI(\hat{\theta}))^{-1/2}.$$

An approximate γ-confidence interval is given by

$$\hat{\theta}\pm(nI(\hat{\theta}))^{-1/2}z_{(1+\gamma)/2}.$$

Finally, if we want to assess the hypothesis $H_0:\theta=\theta_0$, we can do this by computing the approximate P-value

$$2\left\{1-\Phi\left((nI(\theta_0))^{1/2}\left|\hat{\theta}-\theta_0\right|\right)\right\}.$$

Notice that we are using Theorem 6.5.3 for the P-value, rather than Corollary 6.5.2, as, when H_0 is true, we know the asymptotic variance of the MLE is $(nI(\theta_0))^{-1}$. So we do not have to estimate this quantity.

When evaluating I is difficult, we can replace $nI(\hat{\theta})$ by $\hat{I}(x_1,\ldots,x_n)$ in the above expressions for the confidence interval and P-value. We now see very clearly the significance of the observed information. Of course, as we move from using $nI(\theta)$ to $nI(\hat{\theta})$ to $\hat{I}(x_1,\ldots,x_n)$, we expect that larger sample sizes n are needed to make the normality approximation accurate.

We consider some examples.

EXAMPLE 6.5.3 *Location Normal Model*
Using the Fisher information derived in Example 6.5.2, the approximate γ-confidence interval based on the MLE is

$$\hat{\theta}\pm(nI(\hat{\theta}))^{-1/2}z_{(1+\gamma)/2}=\bar{x}\pm(\sigma_0/\sqrt{n})z_{(1+\gamma)/2}.$$

This is just the z-confidence interval derived in Example 6.3.6. Rather than being an approximate γ-confidence interval, the coverage is exact in this case. Similarly, the approximate P-value corresponds to the z-test and the P-value is exact. ∎

EXAMPLE 6.5.4 *Bernoulli Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) distribution, where $\theta \in [0, 1]$ is unknown. The likelihood function is given by

$$L(\theta \,|\, x_1, \ldots, x_n) = \theta^{n\bar{x}} (1-\theta)^{n(1-\bar{x})},$$

and the MLE of θ is $\bar{x}$. The log-likelihood is

$$l(\theta \,|\, x_1, \ldots, x_n) = n\bar{x} \ln \theta + n(1-\bar{x}) \ln(1-\theta),$$

the score function is given by

$$S(\theta \,|\, x_1, \ldots, x_n) = \frac{n\bar{x}}{\theta} - \frac{n(1-\bar{x})}{1-\theta},$$

and

$$\frac{\partial}{\partial\theta} S(\theta \,|\, x_1, \ldots, x_n) = -\frac{n\bar{x}}{\theta^2} - \frac{n(1-\bar{x})}{(1-\theta)^2}.$$

Therefore, the Fisher information for the sample is

$$nI(\theta) = E_\theta\left(-\frac{\partial}{\partial\theta} S(\theta \,|\, X_1, \ldots, X_n)\right) = E_\theta\left(\frac{n\bar{X}}{\theta^2} + \frac{n\left(1-\bar{X}\right)}{(1-\theta)^2}\right) = \frac{n}{\theta(1-\theta)},$$

and the plug-in Fisher information is

$$nI(\bar{x}) = \frac{n}{\bar{x}(1-\bar{x})}.$$

Note that the plug-in Fisher information is the same as the observed Fisher information in this case.

So an approximate γ-confidence interval is given by

$$\hat{\theta} \pm (nI(\hat{\theta}))^{-1/2} z_{(1+\gamma)/2} = \bar{x} \pm z_{(1+\gamma)/2}\sqrt{\bar{x}(1-\bar{x})/n},$$

which is precisely the interval obtained in Example 6.3.7 using large sample considerations based on the central limit theorem. Similarly, we obtain the same P-value as in Example 6.3.11 when testing $H_0 : \theta = \theta_0$. ▮

EXAMPLE 6.5.5 *Poisson Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from a Poisson(λ) distribution, where $\lambda > 0$ is unknown. The likelihood function is given by

$$L(\lambda \,|\, x_1, \ldots, x_n) = \lambda^{n\bar{x}} e^{-n\lambda}.$$

The log-likelihood is

$$l(\lambda \,|\, x_1, \ldots, x_n) = n\bar{x} \ln \lambda - n\lambda,$$

the score function is given by

$$S(\lambda \,|\, x_1, \ldots, x_n) = \frac{n\bar{x}}{\lambda} - n,$$

and

$$\frac{\partial}{\partial\lambda}S(\lambda\,|\,x_1,\ldots,x_n) = -\frac{n\bar{x}}{\lambda^2}.$$

From this we deduce that the MLE of λ is $\hat{\lambda} = \bar{x}$.

Therefore, the Fisher information for the sample is

$$nI(\lambda) = E_\lambda\left(-\frac{\partial}{\partial\lambda}S(\lambda\,|\,X_1,\ldots,X_n)\right) = E_\lambda\left(\frac{n\bar{X}}{\lambda^2}\right) = \frac{n}{\lambda},$$

and the plug-in Fisher information is

$$nI(\bar{x}) = \frac{n}{\bar{x}}.$$

Note that the plug-in Fisher information is the same as the observed Fisher information in this case.

So an approximate γ-confidence interval is given by

$$\hat{\lambda} \pm (nI(\hat{\lambda}))^{-1/2}z_{(1+\gamma)/2} = \bar{x} \pm z_{(1+\gamma)/2}\sqrt{\bar{x}/n}.$$

Similarly, the approximate P-value for testing $H_0 : \lambda = \lambda_0$ is given by

$$2\left\{1 - \Phi\left((nI(\lambda_0))^{1/2}\left|\hat{\lambda} - \lambda_0\right|\right)\right\} = 2\left\{1 - \Phi\left((n/\lambda_0)^{1/2}\left|\bar{x} - \lambda_0\right|\right)\right\}.$$

Note that we have used the Fisher information evaluated at λ_0 for this test. ∎

Summary of Section 6.5

- Under regularity conditions on the statistical model with parameter θ, we can define the Fisher information $I(\theta)$ for the model.
- Under regularity conditions on the statistical model, it can be proved that, when θ is the true value of the parameter, the MLE is consistent for θ and the MLE is approximately normally distributed with mean given by θ and with variance given by $(nI(\theta))^{-1}$.
- The Fisher information $I(\theta)$ can be estimated by plugging in the MLE or by using the observed Fisher information. These estimates lead to practically useful inferences for θ in many problems.

EXERCISES

6.5.1 If $(x_1,\ldots,x_n)$ is a sample from an $N(\mu_0,\sigma^2)$ distribution, where μ_0 is known and $\sigma^2 \in (0,\infty)$ is unknown, determine the Fisher information.

6.5.2 If $(x_1,\ldots,x_n)$ is a sample from a Gamma(α_0,θ) distribution, where α_0 is known and $\theta \in (0,\infty)$ is unknown, determine the Fisher information.

6.5.3 If $(x_1,\ldots,x_n)$ is a sample from a Pareto(α) distribution (see Exercise 6.2.9), where $\alpha > 0$ is unknown, determine the Fisher information.

6.5.4 Suppose the number of calls arriving at an answering service during a given hour of the day is Poisson(λ), where $\lambda \in (0, \infty)$ is unknown. The number of calls actually received during this hour was recorded for 20 days and the following data were obtained.

9	10	8	12	11	12	5	13	9	9
7	5	16	13	9	5	13	8	9	10

Construct an approximate 0.95-confidence interval for λ. Assess the hypothesis that this is a sample from a Poisson(11) distribution. If you are going to decide that the hypothesis is false when the P-value is less than 0.05, then compute an approximate power for this procedure when $\lambda = 10$.

6.5.5 Suppose the lifelengths in hours of lightbulbs from a manufacturing process are known to be distributed Gamma$(2, \theta)$, where $\theta \in (0, \infty)$ is unknown. A random sample of 27 bulbs was taken and their lifelengths measured with the following data obtained.

336.87	2750.71	2199.44	292.99	1835.55	1385.36	2690.52
710.64	2162.01	1856.47	2225.68	3524.23	2618.51	361.68
979.54	2159.18	1908.94	1397.96	914.41	1548.48	1801.84
1016.16	1666.71	1196.42	1225.68	2422.53	753.24	

Determine an approximate 0.90-confidence interval for θ.

6.5.6 Repeat the analysis of Exercise 6.5.5, but this time assume that the lifelengths are distributed Gamma$(1, \theta)$. Comment on the differences in the two analyses.

6.5.7 Suppose that incomes (measured in thousands of dollars) above \$20K can be assumed to be Pareto(α), where $\alpha > 0$ is unknown, for a particular population. A sample of 20 is taken from the population and the following data obtained.

21.265	20.857	21.090	20.047	20.019	32.509	21.622	20.693
20.109	23.182	21.199	20.035	20.084	20.038	22.054	20.190
20.488	20.456	20.066	20.302				

Construct an approximate 0.95-confidence interval for α. Assess the hypothesis that the mean income in this population is \$25K.

6.5.8 Suppose that $(x_1, \ldots, x_n)$ is a sample from an Exponential(θ) distribution. Construct an approximate left-sided γ-confidence interval for θ. (See Problem 6.3.25.)

6.5.9 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Geometric(θ) distribution. Construct an approximate left-sided γ-confidence interval for θ. (See Problem 6.3.25.)

6.5.10 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Negative-Binomial(r, θ) distribution. Construct an approximate left-sided γ-confidence interval for θ. (See Problem 6.3.25.)

PROBLEMS

6.5.11 In Exercise 6.5.1, verify that (6.5.2), (6.5.3), (6.5.4), and (6.5.5) are satisfied.

6.5.12 In Exercise 6.5.2, verify that (6.5.2), (6.5.3), (6.5.4), and (6.5.5) are satisfied.

6.5.13 In Exercise 6.5.3, verify that (6.5.2), (6.5.3), (6.5.4), and (6.5.5) are satisfied.

6.5.14 Suppose that sampling from the model $\{f_\theta : \theta \in \Omega\}$ satisfies (6.5.2), (6.5.3), (6.5.4), and (6.5.5). Prove that $n^{-1}\hat{I} \overset{a.s.}{\to} I(\theta)$ as $n \to \infty$.

6.5.15 (MV) When $\theta = (\theta_1, \theta_2)$, then, under appropriate regularity conditions for the model $\{f_\theta : \theta \in \Omega\}$, the *Fisher information matrix* is defined by

$$I(\theta) = \begin{pmatrix} E_\theta\left(-\frac{\partial^2}{\partial\theta_1^2} l(\theta \,|\, X)\right) & E_\theta\left(-\frac{\partial^2}{\partial\theta_1 \partial\theta_2} l(\theta \,|\, X)\right) \\ \\ E_\theta\left(-\frac{\partial^2}{\partial\theta_1 \partial\theta_2} l(\theta \,|\, X)\right) & E_\theta\left(-\frac{\partial^2}{\partial\theta_2^2} l(\theta \,|\, X)\right) \end{pmatrix}.$$

If $(X_1, X_2, X_3) \sim \text{Multinomial}(1, \theta_1, \theta_2, \theta_3)$ (Example 6.1.5), then determine the Fisher information for this model. Recall that $\theta_3 = 1 - \theta_1 - \theta_2$ and so is determined from (θ_1, θ_2).

6.5.16 (MV) Generalize Problem 6.5.15 to the case where

$$(X_1, \ldots, X_k) \sim \text{Multinomial}(1, \theta_1, \ldots, \theta_k).$$

6.5.17 (MV) Using the definition of the Fisher information matrix in Exercise 6.5.15, determine the Fisher information for the Bivariate Normal$(\mu_1, \mu_2, 1, 1, 0)$ model, where $\mu_1, \mu_2 \in R^1$ are unknown.

6.5.18 (MV) Extending the definition in Exercise 6.5.15 to the three-dimensional case, determine the Fisher information for the Bivariate Normal$\left(\mu_1, \mu_2, \sigma^2, \sigma^2, 0\right)$ model where $\mu_1, \mu_2 \in R^1$, and $\sigma^2 > 0$ are unknown.

CHALLENGES

6.5.19 Suppose that model $\{f_\theta : \theta \in \Omega\}$ satisfies (6.5.2), (6.5.3), (6.5.4), (6.5.5), and has Fisher information $I(\theta)$. If $\Psi : \Omega \to R^1$ is 1–1, and Ψ and Ψ^{-1} are continuously differentiable, then, putting $\Upsilon = \{\Psi(\theta) : \theta \in \Omega\}$, prove that the model given by $\left\{g_\psi : \psi \in \Upsilon\right\}$ satisfies the regularity conditions and that its Fisher information at ψ is given by $I(\Psi^{-1}(\psi))((\Psi^{-1})'(\psi))^2$.

DISCUSSION TOPICS

6.5.20 The method of moments inference methods discussed in Section 6.4.1 are essentially large sample methods based on the central limit theorem. The large sample methods in Section 6.5 are based on the form of the likelihood function. Which methods do you think are more likely to be correct when we know very little about the form of the distribution from which we are sampling? In what sense will your choice be "more correct"?