Chapter 8

Optimal Inferences

CHAPTER OUTLINE

In Chapter 5, we introduced the basic ingredient of statistical inference — the statistical model. In Chapter 6, inference methods were developed based on the model alone via the likelihood function. In Chapter 7, we added the prior distribution on the model parameter, which led to the posterior distribution as the basis for deriving inference methods.

With both the likelihood and the posterior, however, the inferences were derived largely based on intuition. For example, when we had a characteristic of interest $\psi(\theta)$, there was nothing in the theory in Chapters 6 and 7 that forced us to choose a particular estimator, confidence or credible interval, or testing procedure. A complete theory of statistical inference, however, would totally prescribe our inferences.

One attempt to resolve this issue is to introduce a performance measure on inferences and then choose an inference that does best with respect to this measure. For example, we might choose to measure the performance of estimators by their mean-squared error (MSE) and then try to obtain an estimator that had the smallest possible MSE. This is the optimality approach to inference, and it has been quite successful in a number of problems. In this chapter, we will consider several successes for the optimality approach to deriving inferences.

Sometimes the performance measure we use can be considered to be based on what is called a *loss function*. Loss functions form the basis for yet another approach to statistical inference called *decision theory*. While it is not always the case that a performance measure is based on a loss function, this holds in some of the most important problems in statistical inference. Decision theory provides a general framework in which to discuss these problems. A brief introduction to decision theory is provided in Section 8.4 as an advanced topic.

8.1 Optimal Unbiased Estimation

Suppose we want to estimate the real-valued characteristic $\psi(\theta)$ for the statistical model $\{f_\theta : \theta \in \Omega\}$. If we have observed the data s, an estimate is a value $T(s)$ that the statistician hopes will be close to the true value of $\psi(\theta)$. We refer to T as an *estimator* of ψ. The error in the estimate is given by $|T(s) - \psi(\theta)|$. For a variety of reasons (mostly to do with mathematics) it is more convenient to consider the squared error $(T(s) - \psi(\theta))^2$.

Of course, we would like this squared error to be as small as possible. Because we do not know the true value of θ, this leads us to consider the distributions of the squared error, when s has distribution given by f_θ, for each $\theta \in \Omega$. We would then like to choose the estimator T so that these distributions are as concentrated as possible about 0. A convenient measure of the concentration of these distributions about 0 is given by their means, or

$$\text{MSE}_\theta(T) = E_\theta((T - \psi(\theta))^2), \tag{8.1.1}$$

called the *mean-squared error* (recall Definition 6.3.1).

An *optimal estimator* of $\psi(\theta)$ is then a T that minimizes (8.1.1) for every $\theta \in \Omega$. In other words, T would be optimal if, for any other estimator T^* defined on S, we have that

$$\text{MSE}_\theta(T) \le \text{MSE}_\theta(T^*)$$

for each θ. Unfortunately, it can be shown that, except in very artificial circumstances, there is no such T, so we need to modify our optimization problem.

This modification takes the form of restricting the estimators T that we will entertain as possible choices for the inference. Consider an estimator T such that $E_\theta(T)$ does not exist or is infinite. It can then be shown that (8.1.1) is infinite (see Challenge 8.1.26). So we will first restrict our search to those T for which $E_\theta(T)$ is finite for every θ.

Further restrictions on the types of estimators that we consider make use of the following result (recall also Theorem 6.3.1).

Theorem 8.1.1 If T is such that $E(T^2)$ is finite, then

$$E((T - c)^2) = \text{Var}(T) + (E(T) - c)^2,$$

This is minimized by taking $c = E(T)$.

PROOF We have that

$$\begin{aligned} E((T - c)^2) &= E((T - E(T) + E(T) - c)^2) \\ &= E((T - E(T))^2) + 2E(T - E(T))(E(T) - c) + (E(T) - c)^2 \\ &= \text{Var}(T) + (E(T) - c)^2, \end{aligned} \tag{8.1.2}$$

because $E(T - E(T)) = E(T) - E(T) = 0$. As $(E(T) - c)^2 \ge 0$, and $\text{Var}(T)$ does not depend on c, the value of (8.1.2) is minimized by taking $c = E(T)$. ▮

8.1.1 The Rao–Blackwell Theorem and Rao–Blackwellization

We will prove that, when we are looking for T to minimize (8.1.1), we can further restrict our attention to estimators T that depend on the data only through the value of a sufficient statistic. This simplifies our search, as sufficiency often results in a reduction of the dimension of the data (recall the discussion and examples in Section 6.1.1). First, however, we need the following property of sufficiency.

Theorem 8.1.2 A statistic U is sufficient for a model if and only if the conditional distribution of the data s given $U = u$ is the same for every $\theta \in \Omega$.

PROOF See Section 8.5 for the proof of this result. ▮

The implication of this result is that information in the data s beyond the value of $U(s) = u$ can tell us nothing about the true value of θ, because this information comes from a distribution that does not depend on the parameter. Notice that Theorem 8.1.2 is a characterization of sufficiency, alternative to that provided in Section 6.1.1.

Consider a simple example that illustrates the content of Theorem 8.1.2.

EXAMPLE 8.1.1

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

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

Then $L(\cdot \mid 2) = L(\cdot \mid 3) = L(\cdot \mid 4)$, and so $U : S \longrightarrow \{0, 1\}$, given by $U(1) = 0$ and $U(2) = U(3) = U(4) = 1$ is a sufficient statistic.

As we must have $s = 1$ when we observe $U(s) = 0$, the conditional distribution of the response s, given $U(s) = 0$, is degenerate at 1 (i.e., all the probability mass is at the point 1) for both $\theta = a$ and $\theta = b$. When $\theta = a$, the conditional distribution of the response s, given $U(s) = 1$, places 1/3 of its mass at each of the points in $\{2, 3, 4\}$ and similarly when $\theta = b$. So given $U(s) = 1$, the conditional distributions are as in the following table.

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

Thus, we see that indeed the conditional distributions are independent of θ. ▮

We now combine Theorems 8.1.1 and 8.1.2 to show that we can restrict our attention to estimators T that depend on the data only through the value of a sufficient statistic U. By Theorem 8.1.2 we can denote the conditional probability measure for s, given $U(s) = u$, by $P(\cdot \mid U = u)$, i.e., this probability measure does not depend on θ, as it is the same for every $\theta \in \Omega$.

For estimator T of $\psi(\theta)$, such that $E_\theta(T)$ is finite for every θ, put $T_U(s)$ equal to the conditional expectation of T given the value of $U(s)$, namely,

$$T_U(s) = E_{P(\cdot \mid U = U(s))}(T),$$

i.e., T_U is the average value of T when we average using $P(\cdot \mid U = U(s))$. Notice that $T_U(s_1) = T_U(s_2)$ whenever $U(s_1) = U(s_2)$ (this is because $P(\cdot \mid U = U(s_1)) = P(\cdot \mid U = U(s_2))$), and so T_U depends on the data s only through the value of $U(s)$.

Theorem 8.1.3 (*Rao–Blackwell*) Suppose that U is a sufficient statistic and $E_\theta(T^2)$ is finite for every θ. Then $\text{MSE}_\theta(T_U) \le \text{MSE}_\theta(T)$ for every $\theta \in \Omega$.

PROOF Let $P_{\theta,U}$ denote the marginal probability measure of U induced by P_θ. By the theorem of total expectation (see Theorem 3.5.2), we have that

$$\text{MSE}_\theta(T) = E_{P_{\theta,U}}\left(E_{P(\cdot \mid U=u)}((T - \psi(\theta))^2)\right),$$

where $E_{P(\cdot \mid U=u)}((T - \psi(\theta))^2)$ denotes the conditional MSE of T, given $U = u$. Now by Theorem 8.1.1,

$$E_{P(\cdot \mid U=u)}((T - \psi(\theta))^2) = \text{Var}_{P(\cdot \mid U=u)}(T) + (E_{P(\cdot \mid U=u)}(T) - \psi(\theta))^2. \quad (8.1.3)$$

As both terms in (8.1.3) are nonnegative, and recalling the definition of T_U, we have

$$\begin{aligned}\text{MSE}_\theta(T) &= E_{P_{\theta,U}}(\text{Var}_{P(\cdot \mid U=u)}(T)) + E_{P_{\theta,U}}((T_U(s) - \psi(\theta))^2)\\ &\ge E_{P_{\theta,U}}((T_U(s) - \psi(\theta))^2).\end{aligned}$$

Now $(T_U(s) - \psi(\theta))^2 = E_{P(\cdot \mid U=u)}((T_U(s) - \psi(\theta))^2)$ (Theorem 3.5.4) and so, by the theorem of total expectation,

$$\begin{aligned}E_{P_{\theta,U}}((T_U(s) - \psi(\theta))^2) &= E_{P_{\theta,U}}\left(E_{P(\cdot \mid U=u)}((T_U(s) - \psi(\theta))^2)\right)\\ &= E_{P_\theta}((T_U(s) - \psi(\theta))^2) = \text{MSE}_\theta(T_U)\end{aligned}$$

and the theorem is proved. ∎

Theorem 8.1.3 shows that we can always improve on (or at least make no worse) any estimator T that possesses a finite second moment, by replacing $T(s)$ by the estimate $T_U(s)$. This process is sometimes referred to as the *Rao-Blackwellization* of an estimator.

Notice that putting $E = E_\theta$ and $c = \psi(\theta)$ in Theorem 8.1.1 implies that

$$\text{MSE}_\theta(T) = \text{Var}_\theta(T) + (E_\theta(T) - \psi(\theta))^2. \quad (8.1.4)$$

So the MSE of T can be decomposed as the sum of the variance of T plus the squared bias of T (this was also proved in Theorem 6.3.1).

Theorem 8.1.1 has another important implication, for (8.1.4) is minimized by taking $\psi(\theta) = E_\theta(T)$. This indicates that, on average, the estimator T comes closer (in terms of squared error) to $E_\theta(T)$ than to any other value. So, if we are sampling from the distribution specified by θ, $T(s)$ is a natural estimate of $E_\theta(T)$. Therefore, for a general characteristic $\psi(\theta)$, it makes sense to restrict attention to estimators that have bias equal to 0. This leads to the following definition.

Definition 8.1.1 An estimator T of $\psi(\theta)$ is *unbiased* if $E_\theta(T) = \psi(\theta)$ for every $\theta \in \Omega$.

Notice that, for unbiased estimators with finite second moment, (8.1.4) becomes

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

Therefore, our search for an optimal estimator has become the search for an unbiased estimator with smallest variance. If such an estimator exists, we give it a special name.

Definition 8.1.2 An unbiased estimator of $\psi(\theta)$ with smallest variance for each $\theta \in \Omega$ is called a *uniformly minimum variance unbiased (UMVU) estimator.*

It is important to note that the Rao–Blackwell theorem (Theorem 8.1.3) also applies to unbiased estimators. This is because the Rao–Blackwellization of an unbiased estimator yields an unbiased estimator, as the following result demonstrates.

Theorem 8.1.4 (*Rao–Blackwell for unbiased estimators*) If T has finite second moment, is unbiased for $\psi(\theta)$, and U is a sufficient statistic, then $E_\theta(T_U) = \psi(\theta)$ for every $\theta \in \Omega$ (so T_U is also unbiased for $\psi(\theta)$) and $\text{Var}_\theta(T_U) \le \text{Var}_\theta(T)$.

PROOF Using the theorem of total expectation (Theorem 3.5.2), we have

$$E_\theta(T_U) = E_{P_{\theta,U}}(T_U) = E_{P_{\theta,U}}\left(E_{P(\cdot\,|\,U=u)}(T)\right) = E_\theta(T) = \psi(\theta).$$

So T_U is unbiased for $\psi(\theta)$ and $\text{MSE}_\theta(T) = \text{Var}_\theta(T)$, $\text{MSE}_\theta(T_U) = \text{Var}_\theta(T_U)$. Applying Theorem 8.1.3 gives $\text{Var}_\theta\,(T_U) \le \text{Var}_\theta\,(T)$. ∎

There are many situations in which the theory of unbiased estimation leads to good estimators. However, the following example illustrates that in some problems, there are no unbiased estimators and hence the theory has some limitations.

EXAMPLE 8.1.2 *The Nonexistence of an Unbiased Estimator*
Suppose that $(x_1, \ldots, x_n)$ is a sample from the Bernoulli(θ) and we wish to find a UMVU estimator of $\psi(\theta) = \theta/(1-\theta)$, the odds in favor of a success occurring. From Theorem 8.1.4, we can restrict our search to unbiased estimators T that are functions of the sufficient statistic $n\bar{x}$.

Such a T satisfies $E_\theta(T(n\bar{X})) = \theta/(1-\theta)$ for every $\theta \in [0,1]$. Recalling that $n\bar{X} \sim \text{Binomial}(n,\theta)$, this implies that

$$\frac{\theta}{1-\theta} = \sum_{k=0}^{n} T\,(k) \binom{n}{k} \theta^k\,(1-\theta)^{n-k}$$

for every $\theta \in [0,1]$. By the binomial theorem, we have

$$(1-\theta)^{n-k} = \sum_{l=0}^{n-k} \binom{n-k}{l} (-1)^l\,\theta^l.$$

Substituting this into the preceding expression for $\theta/(1-\theta)$ and writing this in terms of powers of θ leads to

$$\frac{\theta}{1-\theta} = \sum_{m=0}^{n} \left(\sum_{k=0}^{m} T(k) \binom{n}{k} (-1)^{m-k} \right) \theta^m. \tag{8.1.5}$$

Now the left-hand side of (8.1.5) goes to ∞ as $\theta \to 1$, but the right-hand side is a polynomial in θ, which is bounded in $[0, 1]$. Therefore, an unbiased estimator of ψ cannot exist. ∎

If a characteristic $\psi(\theta)$ has an unbiased estimator, then it is said to be *U-estimable*. It should be kept in mind, however, that just because a parameter is not U-estimable does not mean that we cannot estimate it! For example, ψ in Example 8.1.2, is a 1–1 function of θ, so the MLE of ψ is given by $\bar{x}/(1-\bar{x})$ (see Theorem 6.2.1); this seems like a sensible estimator, even if it is biased.

8.1.2 Completeness and the Lehmann–Scheffé Theorem

In certain circumstances, if an unbiased estimator exists, and is a function of a sufficient statistic U, then there is only one such estimator — so it must be UMVU. We need the concept of completeness to establish this.

Definition 8.1.3 A statistic U is *complete* if any function h of U, which satisfies $E_\theta(h(U)) = 0$ for every $\theta \in \Omega$, also satisfies $h(U(s)) = 0$ with probability 1 for each $\theta \in \Omega$ (i.e., $P_\theta(\{s : h(U(s)) = 0\}) = 1$ for every $\theta \in \Omega$).

In probability theory, we treat two functions as equivalent if they differ only on a set having probability content 0, as the probability of the functions taking different values at an observed response value is 0. So in Definition 8.1.3, we need not distinguish between h and the constant 0. Therefore, a statistic U is complete if the only unbiased estimator of 0, based on U, is given by 0 itself.

We can now derive the following result.

Theorem 8.1.5 (*Lehmann–Scheffé*) If U is a complete sufficient statistic, and if T depends on the data only through the value of U, has finite second moment for every θ, and is unbiased for $\psi(\theta)$, then T is UMVU.

PROOF Suppose that T^* is also an unbiased estimator of $\psi(\theta)$. By Theorem 8.1.4 we can assume that T^* depends on the data only through the value of U. Then there exist functions h and h^* such that $T(s) = h(U(s))$ and $T^*(s) = h^*(U(s))$ and

$$0 = E_\theta(T) - E_\theta(T^*) = E_\theta(h(U)) - E_\theta(h^*(U)) = E_\theta(h(U) - h^*(U)).$$

By the completeness of U, we have that $h(U) = h^*(U)$ with probability 1 for each $\theta \in \Omega$, which implies that $T = T^*$ with probability 1 for each $\theta \in \Omega$. This says there is essentially only one unbiased estimator for $\psi(\theta)$ based on U, and so it must be UMVU. ∎

The Rao–Blackwell theorem for unbiased estimators (Theorem 8.1.4), together with the Lehmann–Scheffé theorem, provide a method for obtaining a UMVU estimator of $\psi(\theta)$. Suppose we can find an unbiased estimator T that has finite second moment. If we also have a complete sufficient statistic U, then by Theorem 8.1.4 $T_U(s) = E_{P(\cdot\,|\,U=U(s))}(T)$ is unbiased for $\psi(\theta)$ and depends on the data only through the value of U, because $T_U(s_1) = T_U(s_2)$ whenever $U(s_1) = U(s_2)$. Therefore, by Theorem 8.1.5, T_U is UMVU for $\psi(\theta)$.

It is not necessary, in a given problem, that a complete sufficient statistic exist. In fact, it can be proved that the only candidate for this is a minimal sufficient statistic (recall the definition in Section 6.1.1). So in a given problem, we must obtain a minimal sufficient statistic and then determine whether or not it is complete. We illustrate this via an example.

EXAMPLE 8.1.3 *Location Normal*
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 $\sigma_0^2 > 0$ is known. In Example 6.1.7, we showed that $\bar{x}$ is a minimal sufficient statistic for this model.

In fact, $\bar{x}$ is also complete for this model. The proof of this is a bit involved and is presented in Section 8.5.

Given that $\bar{x}$ is a complete, minimal sufficient statistic, this implies that $T(\bar{x})$ is a UMVU estimator of its mean $E_\mu(T(\bar{X}))$ whenever T has a finite second moment for every $\mu \in R^1$. In particular, $\bar{x}$ is the UMVU estimator of μ because $E_\mu(\bar{X}) = \mu$ and $E_\mu(\bar{X}^2) = (\sigma_0^2/n) + \mu^2 < \infty$. Furthermore, $\bar{x} + \sigma_0 z_p$ is the UMVU estimator of $E_\mu(\bar{X} + \sigma_0 z_p) = \mu + \sigma_0 z_p$ (the pth quantile of the true distribution). ∎

The arguments needed to show the completeness of a minimal sufficient statistic in a problem are often similar to the one required in Example 8.1.3 (see Challenge 8.1.27). Rather than pursue such technicalities here, we quote some important examples in which the minimal sufficient statistic is complete.

EXAMPLE 8.1.4 *Location-Scale Normal*
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 and is given by $(\mu, \sigma^2) \in R^1 \times (0, \infty)$.

We showed, in Example 6.1.8, that $(\bar{x}, s^2)$ is a minimal sufficient statistic for this model. In fact, it can be shown that $(\bar{x}, s^2)$ is a complete minimal sufficient statistic. Therefore, $T(\bar{x}, s^2)$ is a UMVU estimator of $E_\theta(T(\bar{X}, S^2))$ whenever the second moment of $T(\bar{x}, s^2)$ is finite for every (μ, σ^2). In particular, $\bar{x}$ is the UMVU estimator of μ and s^2 is UMVU for σ^2. ∎

EXAMPLE 8.1.5 *Distribution-Free Models*
Suppose that $(x_1, \ldots, x_n)$ is a sample from some continuous distribution on R^1. The statistical model comprises all continuous distributions on R^1.

It can be shown that the order statistics $(x_{(1)}, \ldots, x_{(n)})$ make up a complete minimal sufficient statistic for this model. Therefore, $T(x_{(1)}, \ldots, x_{(n)})$ is UMVU for

$$E_\theta(T(X_{(1)}, \ldots, X_{(n)}))$$

whenever

$$E_\theta(T^2(X_{(1)}, \ldots, X_{(n)})) < \infty \tag{8.1.6}$$

for every continuous distribution. In particular, if $T : R^n \to R^1$ is bounded, then this is the case. For example, if

$$T(x_{(1)}, \ldots, x_{(n)}) = \frac{1}{n} \sum_{i=1}^{n} I_A(x_{(i)}),$$

the relative frequency of the event A in the sample, then $T(x_{(1)}, \ldots, x_{(n)})$ is UMVU for $E_\theta(T(X_{(1)}, \ldots, X_{(n)})) = P_\theta(A)$.

Now change the model assumption so that $(x_1, \ldots, x_n)$ is a sample from some continuous distribution on R^1 that possesses its first m moments. Again, it can be shown that the order statistics make up a complete minimal sufficient statistic. Therefore, $T(x_{(1)}, \ldots, x_{(n)})$ is UMVU for $E_\theta(T(X_{(1)}, \ldots, X_{(n)}))$ whenever (8.1.6) holds for every continuous distribution possessing its first m moments. For example, if $m = 2$, then this implies that $T(x_{(1)}, \ldots, x_{(n)}) = \bar{x}$ is UMVU for $E_\theta(\bar{X})$. When $m = 4$, we have that s^2 is UMVU for the population variance (see Exercise 8.1.2). ∎

8.1.3 The Cramer–Rao Inequality (Advanced)

There is a fundamental inequality that holds for the variance of an estimator T. This is given by the *Cramer–Rao inequality* (sometimes called the *information inequality*). It is a corollary to the following inequality.

Theorem 8.1.6 (*Covariance inequality*) Suppose $T, U_\theta : S \to R^1$ and $E_\theta(T^2) < \infty$, $0 < E_\theta(U_\theta^2) < \infty$ for every $\theta \in \Omega$. Then

$$\text{Var}_\theta(T) \ge \frac{(\text{Cov}_\theta(T, U_\theta))^2}{\text{Var}_\theta(U_\theta)}$$

for every $\theta \in \Omega$. Equality holds if and only if

$$T(s) = E_\theta(T) + \frac{\text{Cov}_\theta(T, U_\theta)}{\text{Var}_\theta(U_\theta)}(U_\theta(s) - E_\theta(U_\theta(s)))$$

with probability 1 for every $\theta \in \Omega$ (i.e., if and only if $T(s)$ and $U_\theta(s)$ are linearly related).

PROOF This result follows immediately from the Cauchy–Schwartz inequality (Theorem 3.6.3). ∎

Now suppose that Ω is an open subinterval of R^1 and we take

$$U_\theta(s) = S(\theta \,|\, s) = \frac{\partial \ln f_\theta(s)}{\partial \theta}, \tag{8.1.7}$$

i.e., U_θ is the score function. Assume that the conditions discussed in Section 6.5 hold, so that $E_\theta(S(\theta \,|\, s)) = 0$ for all θ, and, Fisher's information $I(\theta) = \text{Var}_\theta(S(\theta \,|\, s))$ is finite. Then using

$$\frac{\partial \ln f_\theta(s)}{\partial \theta} = \frac{\partial f_\theta(s)}{\partial \theta}\frac{1}{f_\theta(s)},$$

we have

$$\begin{aligned}
&\mathrm{Cov}_\theta(T, U_\theta) \\
&= E_\theta\left(T(s)\frac{\partial \ln f_\theta(s)}{\partial \theta}\right) = E_\theta\left(T(s)\frac{\partial f_\theta(s)}{\partial \theta}\frac{1}{f_\theta(s)}\right) \\
&= \sum_s \left(T(s)\frac{\partial f_\theta(s)}{\partial \theta}\frac{1}{f_\theta(s)}\right) f_\theta(s) = \frac{\partial}{\partial \theta}\sum_s T(s)\, f_\theta(s) = \frac{\partial E_\theta(T)}{\partial \theta}, \qquad (8.1.8)
\end{aligned}$$

in the discrete case, where we have assumed conditions like those discussed in Section 6.5, so we can pull the partial derivative through the sum. A similar argument gives the equality (8.1.8) in the continuous case as well.

The covariance inequality, applied with U_θ specified as in (8.1.7) and using (8.1.8), gives the following result.

Corollary 8.1.1 (*Cramer–Rao or information inequality*) Under conditions,

$$\mathrm{Var}_\theta(T) \geq \left(\frac{\partial E_\theta(T)}{\partial \theta}\right)^2 (I(\theta))^{-1}$$

for every $\theta \in \Omega$. Equality holds if and only if

$$T(s) = E_\theta(T) + \frac{\partial E_\theta(T)}{\partial \theta}(I(\theta))^{-1} S(\theta \,|\, s)$$

with probability 1 for every $\theta \in \Omega$.

The Cramer–Rao inequality provides a fundamental lower bound on the variance of an estimator T. From (8.1.4), we know that the variance is a relevant measure of the accuracy of an estimator only when the estimator is unbiased, so we restate Corollary 8.1.1 for this case.

Corollary 8.1.2 Under the conditions of Corollary 8.1.1, when T is an unbiased estimator of $\psi(\theta)$,

$$\mathrm{Var}_\theta(T) \geq (\psi'(\theta))^2 (I(\theta))^{-1}$$

for every $\theta \in \Omega$. Equality holds if and only if

$$T(s) = \psi(\theta) + \psi'(\theta)(I(\theta))^{-1} S(\theta \,|\, s) \qquad (8.1.9)$$

with probability 1 for every $\theta \in \Omega$.

Notice that when $\psi(\theta) = \theta$, then Corollary 8.1.2 says that the variance of the unbiased estimator T is bounded below by the reciprocal of the Fisher information. More generally, when ψ is a 1–1, smooth transformation, we have (using Challenge 6.5.19) that the variance of an unbiased T is again bounded below by the reciprocal of the Fisher information, but this time the model uses the parameterization in terms of $\psi(\theta)$.

Corollary 8.1.2 has several interesting implications. First, if we obtain an unbiased estimator T with variance at the lower bound, then we know immediately that it is UMVU. Second, we know that any unbiased estimator that achieves the lower bound is of the form given in (8.1.9). Note that the right-hand side of (8.1.9) must be independent of θ in order for this to be an estimator. If this is not the case, then there are no UMVU estimators whose variance achieves the lower bound. The following example demonstrates that there are cases in which UMVU estimators exist, but their variance does not achieve the lower bound.

EXAMPLE 8.1.6 *Poisson*(λ) *Model*

Suppose that $(x_1, \ldots, x_n)$ is a sample from the Poisson(λ) distribution where $\lambda > 0$ is unknown. The log-likelihood is given by $l(\lambda \,|\, x_1, \ldots, x_n) = n\bar{x} \ln \lambda - n\lambda$, so the score function is given by $S(\lambda \,|\, x_1, \ldots, x_n) = n\bar{x}/\lambda - n$. Now

$$\frac{\partial S(\lambda \,|\, x_1, \ldots, x_n)}{\partial \lambda} = -\frac{n\bar{x}}{\lambda^2},$$

and thus

$$I(\lambda) = E_\lambda\left(\frac{n\bar{x}}{\lambda^2}\right) = \frac{n}{\lambda}.$$

Suppose we are estimating λ. Then the Cramer–Rao lower bound is given by $I^{-1}(\lambda) = \lambda/n$. Noting that $\bar{x}$ is unbiased for λ and that $\text{Var}_\lambda(\bar{X}) = \lambda/n$, we see immediately that $\bar{x}$ is UMVU and achieves the lower bound.

Now suppose that we are estimating $\psi(\lambda) = e^{-\lambda} = P_\lambda(\{0\})$. The Cramer–Rao lower bound equals $\lambda e^{-2\lambda}/n$ and

$$\begin{aligned}\psi(\lambda) + \psi'(\lambda)I^{-1}(\lambda)S(\lambda \,|\, x_1, \ldots, x_n) &= e^{-\lambda} - e^{-\lambda}\left(\frac{\lambda}{n}\right)\left(\frac{n\bar{x}}{\lambda} - n\right) \\ &= e^{-\lambda}(1 - \bar{x} + \lambda),\end{aligned}$$

which is clearly not independent of λ. So there does not exist a UMVU estimator for ψ that attains the lower bound.

Does there exist a UMVU estimator for ψ? Observe that when $n = 1$, then $I_{\{0\}}(x_1)$ is an unbiased estimator of ψ. As it turns out, $\bar{x}$ is (for every n) a complete minimal sufficient statistic for this model, so by the Lehmann–Scheffé theorem $I_{\{0\}}(x_1)$ is UMVU for ψ. Furthermore, $I_{\{0\}}(X_1)$ has variance

$$P_\lambda(X_1 = 0)\,(1 - P_\lambda(X_1 = 0)) = e^{-\lambda}(1 - e^{-\lambda})$$

since $I_{\{0\}}(X_1) \sim$ Bernoulli$(e^{-\lambda})$. This implies that $e^{-\lambda}(1 - e^{-\lambda}) > \lambda e^{-2\lambda}$.

In general, we have that

$$\frac{1}{n}\sum_{i=1}^{n} I_{\{0\}}(x_i)$$

is an unbiased estimator of ψ, but it is not a function of $\bar{x}$. Thus we cannot apply the Lehmann–Scheffé theorem, but we can Rao–Blackwellize this estimator. Therefore, the UMVU estimator of ψ is given by

$$\frac{1}{n}\sum_{i=1}^{n} E(I_{\{0\}}(X_i)\,|\,\bar{X}=\bar{x}).$$

To determine this estimator in closed form, we reason as follows. The conditional probability function of $(X_1,\ldots,X_n)$ given $\bar{X}=\bar{x}$, because $n\bar{X}$ is distributed Poisson$(n\lambda)$, is

$$\left\{\frac{\lambda^{x_1}}{x_1!}\cdots\frac{\lambda^{x_n}}{x_n!}e^{-n\lambda}\right\}\left\{\frac{(n\lambda)^{n\bar{x}}}{(n\bar{x})!}e^{-n\lambda}\right\}^{-1}=\binom{n\bar{x}}{x_1\ \ldots\ x_n}\left(\frac{1}{n}\right)^{x_1}\cdots\left(\frac{1}{n}\right)^{x_n},$$

i.e., $(X_1,\ldots,X_n)$ given $\bar{X}=\bar{x}$ is distributed Multinomial$(n\bar{x},1/n,\ldots,1/n)$. Accordingly, the UMVU estimator is given by

$$E(I_{\{0\}}(X_1)\,|\,\bar{X}=\bar{x})=P(X_1=0\,|\,\bar{X}=\bar{x})=\left(1-\frac{1}{n}\right)^{n\bar{x}}$$

because $X_i\,|\,\bar{X}=\bar{x}\sim$ Binomial$(n\bar{x},1/n)$ for each $i=1,\ldots,n$.

Certainly, it is not at all obvious from the functional form that this estimator is unbiased, let alone UMVU. So this result can be viewed as a somewhat remarkable application of the theory. ▮

Recall now Theorems 6.5.2 and 6.5.3. The implications of these results, with some additional conditions, are that the MLE of θ is asymptotically unbiased for θ and that the asymptotic variance of the MLE is at the information lower bound. This is often interpreted to mean that, with large samples, the MLE makes full use of the information about θ contained in the data.

Summary of Section 8.1

- An estimator comes closest (using squared distance) on average to its mean (see Theorem 8.1.1), so we can restrict attention to unbiased estimators for quantities of interest.
- The Rao–Blackwell theorem says that we can restrict attention to functions of a sufficient statistic when looking for an estimator minimizing MSE.
- When a sufficient statistic is complete, then any function of that sufficient statistic is UMVU for its mean.
- The Cramer–Rao lower bound gives a lower bound on the variance of an unbiased estimator and a method for obtaining an estimator that has variance at this lower bound when such an estimator exists.

EXERCISES

8.1.1 Suppose that a statistical model is given by the two distributions in the following table.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$
$f_a(s)$	$1/3$	$1/6$	$1/12$	$5/12$
$f_b(s)$	$1/2$	$1/4$	$1/6$	$1/12$

If $T : \{1, 2, 3, 4\} \to \{1, 2, 3, 4\}$ is defined by $T(1) = T(2) = 1$ and $T(s) = s$ otherwise, then prove that T is a sufficient statistic. Derive the conditional distributions of s given $T(s)$ and show that these are independent of θ.

8.1.2 Suppose that $(x_1, \ldots, x_n)$ is a sample from a distribution with mean μ and variance σ^2. Prove that $s^2 = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$ is unbiased for σ^2.

8.1.3 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. Determine a UMVU estimator of the second moment $\mu^2 + \sigma_0^2$.

8.1.4 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. Determine a UMVU estimator of the first quartile $\mu + \sigma_0 z_{0.25}$.

8.1.5 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. Is $2\bar{x} + 3$ a UMVU estimator of anything? If so, what is it UMVU for? Justify your answer.

8.1.6 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) distribution, where $\theta \in [0, 1]$ is unknown. Determine a UMVU estimator of θ (use the fact that a minimal sufficient statistic for this model is complete).

8.1.7 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Gamma(α_0, β) distribution, where α_0 is known and $\beta > 0$ is unknown. Using the fact that $\bar{x}$ is a complete sufficient statistic (see Challenge 8.1.27), determine a UMVU estimator of β^{-1}.

8.1.8 Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu_0, \sigma^2)$ distribution, where μ_0 is known and $\sigma^2 > 0$ is unknown. Show that $\sum_{i=1}^n \left(x_i - \mu_0\right)^2$ is a sufficient statistic for this problem. Using the fact that it is complete, determine a UMVU estimator for σ^2.

8.1.9 Suppose a statistical model comprises all continuous distributions on R^1. Based on a sample of n, determine a UMVU estimator of $P((-1, 1))$, where P is the true probability measure. Justify your answer.

8.1.10 Suppose a statistical model comprises all continuous distributions on R^1 that have a finite second moment. Based on a sample of n, determine a UMVU estimator of μ^2, where μ is the true mean. Justify your answer. (Hint: Find an unbiased estimator for $n = 2$, Rao–Blackwellize this estimator for a sample of n, and then use the Lehmann–Scheffé theorem.)

8.1.11 The estimator determined in Exercise 8.1.10 is also unbiased for μ^2 when the statistical model comprises all continuous distributions on R^1 that have a finite first moment. Is this estimator still UMVU for μ^2?

PROBLEMS

8.1.12 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Uniform$[0, \theta]$ distribution, where $\theta > 0$ is unknown. Show that $x_{(n)}$ is a sufficient statistic and determine its distribution. Using the fact that $x_{(n)}$ is complete, determine a UMVU estimator of θ.

8.1.13 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) distribution, where $\theta \in [0, 1]$ is unknown. Then determine the conditional distribution of $(x_1, \ldots, x_n)$, given the value of the sufficient statistic $\bar{x}$.

8.1.14 Prove that $L(\theta, a) = (\theta - a)^2$ satisfies

$$L(\theta, \alpha a_1 + (1 - \alpha)a_2) \le \alpha L(\theta, a_1) + (1 - \alpha)\, L(\theta, a_2)$$

when a ranges in a subinterval of R^1. Use this result together with Jensen's inequality (Theorem 3.6.4) to prove the Rao–Blackwell theorem.

8.1.15 Prove that $L(\theta, a) = |\theta - a|$ satisfies

$$L(\theta, \alpha a_1 + (1 - \alpha)a_2) \le \alpha L(\theta, a_1) + (1 - \alpha)L(\theta, a_2)$$

when a ranges in a subinterval of R^1. Use this result together with Jensen's inequality (Theorem 3.6.4) to prove the Rao–Blackwell theorem for absolute error. (Hint: First show that $|x + y| \le |x| + |y|$ for any x and y.)

8.1.16 Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $(\mu, \sigma^2) \in R^1 \times (0, \infty)$ is unknown. Show that the optimal estimator (in the sense of minimizing the MSE), of the form cs^2 for σ^2, is given by $c = (n - 1)/\,(n + 1)$. Determine the bias of this estimator and show that it goes to 0 as $n \to \infty$.

8.1.17 Prove that if a statistic T is complete for a model and $U = h(T)$ for a 1–1 function h, then U is also complete.

8.1.18 Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $(\mu, \sigma^2) \in R^1 \times (0, \infty)$ is unknown. Derive a UMVU estimator of the standard deviation σ. (Hint: Calculate the expected value of the sample standard deviation s.)

8.1.19 Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $(\mu, \sigma^2) \in R^1 \times (0, \infty)$ is unknown. Derive a UMVU estimator of the first quartile $\mu + \sigma z_{0.25}$. (Hint: Problem 8.1.17.)

8.1.20 Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where $\theta \in \Omega = \{\mu_1, \mu_2\}$ is unknown and $\sigma_0^2 > 0$ is known. Establish that $\bar{x}$ is a minimal sufficient statistic for this model but that it is not complete.

8.1.21 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. Determine the information lower bound, for an unbiased estimator, when we consider estimating the second moment $\mu^2 + \sigma_0^2$. Does the UMVU estimator in Exercise 8.1.3 attain the information lower bound?

8.1.22 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Gamma(α_0, β) distribution, where α_0 is known and $\beta > 0$ is unknown. Determine the information lower bound for the estimation of β^{-1} using unbiased estimators, and determine if the UMVU estimator obtained in Exercise 8.1.7 attains this.

8.1.23 Suppose that $(x_1, \ldots, x_n)$ is a sample from the distribution with density f_θ $(x) = \theta x^{\theta - 1}$ for $x \in [0, 1]$ and $\theta > 0$ is unknown. Determine the information lower bound for estimating θ using unbiased estimators. Does a UMVU estimator with variance at the lower bound exist for this problem?

8.1.24 Suppose that a statistic T is a complete statistic based on some statistical model. A submodel is a statistical model that comprises only some of the distributions in the original model. Why is it not necessarily the case that T is complete for a submodel?

8.1.25 Suppose that a statistic T is a complete statistic based on some statistical model. If we construct a larger model that contains all the distributions in the original model and is such that any set that has probability content equal to 0 for every distribution in the original model also has probability content equal to 0 for every distribution in the larger model, then prove that T is complete for the larger model as well.

CHALLENGES

8.1.26 If X is a random variable such that $E(X)$ either does not exist or is infinite, then show that $E((X-c)^2) = \infty$ for any constant c.

8.1.27 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Gamma(α_0, β) distribution, where α_0 is known and $\beta > 0$ is unknown. Show that $\bar{x}$ is a complete minimal sufficient statistic.

8.2 Optimal Hypothesis Testing

Suppose we want to assess a hypothesis about the real-valued characteristic $\psi(\theta)$ for the model $\{f_\theta : \theta \in \Omega\}$. Typically, this will take the form $H_0 : \psi(\theta) = \psi_0$, where we have specified a value for ψ. After observing data s, we want to assess whether or not we have evidence against H_0.

In Section 6.3.3, we discussed methods for assessing such a hypothesis based on the plug-in MLE for $\psi(\theta)$. These involved computing a P-value as a measure of how surprising the data s are when the null hypothesis is assumed to be true. If s is surprising for each of the distributions f_θ for which $\psi(\theta) = \psi_0$, then we have evidence against H_0. The development of such procedures was largely based on the intuitive justification for the likelihood function.

8.2.1 The Power Function of a Test

Closely associated with a specific procedure for computing a P-value is the concept of a *power function* $\beta(\theta)$, as defined in Section 6.3.6. For this, we specified a *critical value* α, such that we declare the results of the test statistically significant whenever the P-value is less than or equal to α. The power $\beta(\theta)$ is then the probability of the P-value being less than or equal to α when we are sampling from f_θ. The greater the value of $\beta(\theta)$, when $\psi(\theta) \neq \psi_0$, the better the procedure is at detecting departures from H_0. The power function is thus a measure of the sensitivity of the testing procedure to detecting departures from H_0.

Recall the following fundamental example.

EXAMPLE 8.2.1 *Location Normal Model*
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 want to assess the null hypothesis $H_0 : \mu = \mu_0$. In Example 6.3.9, we showed that a sensible test for this problem is based on the z-statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}},$$

with $Z \sim N(0, 1)$ under H_0. The P-value is then given by

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

where Φ denotes the $N(0, 1)$ distribution function.

In Example 6.3.18, we showed that, for critical value α, the power function of the z-test is given by

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

because $\bar{X} \sim N(\mu, \sigma_0^2/n)$.

We see that specifying a value for α specifies a set of data values

$$R = \left\{(x_1, \ldots, x_n) : \Phi\left(\left|\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right|\right) > 1 - \frac{\alpha}{2}\right\}$$

such that the results of the test are determined to be statistically significant whenever $(x_1, \ldots, x_n) \in R$. Using the fact that Φ is 1–1 increasing, we can also write R as

$$\begin{aligned}R &= \left\{(x_1, \ldots, x_n) : \left|\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right\} \\ &= \left\{(x_1, \ldots, x_n) : \left|\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right| > z_{1-(\alpha/2)}\right\}.\end{aligned}$$

Furthermore, the power function is given by $\beta(\mu) = P_\mu(R)$ and $\beta(\mu_0) = P_{\mu_0}(R) = \alpha$. ▮

8.2.2 Type I and Type II Errors

We now adopt a different point of view. We are going to look for tests that are optimal for testing the null hypothesis $H_0 : \psi(\theta) = \psi_0$. First, we will assume that, having observed the data s, we will decide to either accept or reject H_0. If we reject H_0, then this is equivalent to accepting the alternative $H_a : \psi(\theta) \neq \psi_0$. Our performance measure for assessing testing procedures will then be the probability that the testing procedure makes an error.

There are two types of error. We can make a *type I error* — rejecting H_0 when it is true — or make a *type II error* — accepting H_0 when H_0 is false. Note that if we reject H_0, then this implies that we are accepting the *alternative hypothesis* $H_a : \psi(\theta) \neq \psi_0$.

It turns out that, except in very artificial circumstances, there are no testing procedures that simultaneously minimize the probabilities of making the two kinds of errors. Accordingly, we will place an upper bound α, called the *critical value*, on the probability of making a type I error. We then search among those tests whose probability of making a type I error is less than or equal to α, for a testing procedure that minimizes the probability of making a type II error.

Sometimes hypothesis testing problems for real-valued parameters are distinguished as being one-sided or two-sided. For example, if θ is real-valued, then $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ is a two-sided testing problem, while $H_0 : \theta \le \theta_0$ versus $H_a : \theta > \theta_0$ or $H_0 : \theta \ge \theta_0$ versus $H_a : \theta < \theta_0$ are examples of one-sided problems. Notice, however, that if we define

$$\psi(\theta) = I_{(\theta_0,\infty)}(\theta),$$

then $H_0 : \theta \le \theta_0$ versus $H_a : \theta > \theta_0$ is equivalent to the problem $H_0 : \psi(\theta) = 0$ versus $H_a : \psi(\theta) \neq 0$. Similarly, if we define

$$\psi(\theta) = I_{(-\infty,\theta_0)}(\theta),$$

then $H_0 : \theta \ge \theta_0$ versus $H_a : \theta < \theta_0$ is equivalent to the problem $H_0 : \psi(\theta) = 0$ versus $H_a : \psi(\theta) \neq 0$. So the formulation we have adopted for testing problems about a general ψ includes the one-sided problems as special cases.

8.2.3 Rejection Regions and Test Functions

One approach to specifying a testing procedure is to select a subset $R \subset S$ before we observe s. We then *reject* H_0 whenever $s \in R$ and accept H_0 whenever $s \notin R$. The set R is referred to as a *rejection region*. Putting an upper bound on the probability of rejecting H_0 when it is true leads to the following.

Definition 8.2.1 A rejection region R satisfying

$$P_\theta(R) \le \alpha \tag{8.2.1}$$

whenever $\psi(\theta) = \psi_0$ is called a *size α rejection region* for H_0.

So (8.2.1) expresses the bound on the probability of making a type I error.

Among all size α rejection regions R, we want to find the one (if it exists) that will minimize the probability of making a type II error. This is equivalent to finding the size α rejection region R that maximizes the probability of rejecting the null hypothesis when it is false. This probability can be expressed in terms of the power function of R and is given by $\beta(\theta) = P_\theta(R)$ whenever $\psi(\theta) \neq \psi_0$.

To fully specify the optimality approach to testing hypotheses, we need one additional ingredient. Observe that our search for an optimal size α rejection region R is equivalent to finding the indicator function I_R that satisfies $\beta(\theta) = E_\theta(I_R) = P_\theta(R) \le$

α, when $\psi(\theta) = \psi_0$, and maximizes $\beta(\theta) = E_\theta(I_R) = P_\theta(R)$, when $\psi(\theta) \neq \psi_0$. It turns out that, in a number of problems, there is no such rejection region.

On the other hand, there is often a solution to the more general problem of finding a function $\varphi : S \to [0, 1]$ satisfying

$$\beta(\theta) = E_\theta(\varphi) \leq \alpha, \tag{8.2.2}$$

when $\psi(\theta) = \psi_0$, and maximizes

$$\beta(\theta) = E_\theta(\varphi),$$

when $\psi(\theta) \neq \psi_0$. We have the following terminology.

Definition 8.2.2 We call $\varphi : S \to [0, 1]$ a *test function* and $\beta(\theta) = E_\theta(\varphi)$ the power function associated with the test function φ. If φ satisfies (8.2.2) when $\psi(\theta) = \psi_0$, it is called a *size α test function.* If φ satisfies $E_\theta(\varphi) = \alpha$ when $\psi(\theta) = \psi_0$, it is called an *exact size α test function.* A size α test function φ that maximizes $\beta(\theta) = E_\theta(\varphi)$ when $\psi(\theta) \neq \psi_0$ is called a *uniformly most powerful (UMP) size α test function.*

Note that $\varphi = I_R$ is a test function with power function given by $\beta(\theta) = E_\theta(I_R) = P_\theta(R)$.

For observed data s, we interpret $\varphi(s) = 0$ to mean that we accept H_0 and interpret $\varphi(s) = 1$ to mean that we reject H_0. In general, we interpret $\varphi(s)$ to be the conditional probability that we reject H_0 given the data s. Operationally, this means that, after we observe s, we generate a Bernoulli($\varphi(s)$) random variable. If we get a 1, we reject H_0; if we get a 0, we accept H_0. Therefore, by the theorem of total expectation, $E_\theta(\varphi)$ is the unconditional probability of rejecting H_0. The randomization that occurs when $0 < \varphi(s) < 1$ may seem somewhat counterintuitive, but it is forced on us by our search for a UMP size α test, as we can increase power by doing this in certain problems.

8.2.4 The Neyman–Pearson Theorem

For a testing problem specified by a null hypothesis $H_0 : \psi(\theta) = \psi_0$ and a critical value α, we want to find a UMP size α test function φ. Note that a UMP size α test function φ_0 for $H_0 : \psi(\theta) = \psi_0$ is characterized (letting β_φ denote the power function of φ) by

$$\beta_{\varphi_0}(\theta) \leq \alpha,$$

when $\psi(\theta) = \psi_0$, and by

$$\beta_{\varphi_0}(\theta) \geq \beta_\varphi(\theta),$$

when $\psi(\theta) \neq \psi_0$, for any other size α test function φ.

Still, this optimization problem does not have a solution in general. In certain problems, however, an optimal solution can be found. The following result gives one such example. It is fundamental to the entire theory of optimal hypothesis testing.

Theorem 8.2.1 (*Neyman–Pearson*) Suppose that $\Omega = \{\theta_0, \theta_1\}$ and that we want to test $H_0 : \theta = \theta_0$. Then an exact size α test function φ_0 exists of the form

$$\varphi_0(s) = \begin{cases} 1 & f_{\theta_1}(s)/f_{\theta_0}(s) > c_0 \\ \gamma & f_{\theta_1}(s)/f_{\theta_0}(s) = c_0 \\ 0 & f_{\theta_1}(s)/f_{\theta_0}(s) < c_0 \end{cases} \tag{8.2.3}$$

for some $\gamma \in [0, 1]$ and $c_0 \geq 0$. This test is UMP size α.

PROOF See Section 8.5 for the proof of this result. ▮

The following result can be established by a simple extension of the proof of the Neyman–Pearson theorem.

Corollary 8.2.1 If φ is a UMP size α test, then $\varphi(s) = \varphi_0(s)$ everywhere except possibly on the *boundary* $B = \{s : f_{\theta_1}(s)/f_{\theta_0}(s) = c_0\}$. Furthermore, φ has exact size α unless the power of a UMP size α test equals 1.

PROOF See Challenge 8.2.22. ▮

Notice the intuitive nature of the test given by the Neyman–Pearson theorem, for (8.2.3) indicates that we categorically reject H_0 as being true when the likelihood ratio of θ_1 versus θ_0 is greater than the constant c_0, and we accept H_0 when it is smaller. When the likelihood ratio equals c_0, we randomly decide to reject H_0 with probability γ. Also, Corollary 8.2.1 says that a UMP size α test is basically unique, although there are possibly different randomization strategies on the boundary.

The proof of the Neyman–Pearson theorem reveals that c_0 is the smallest real number such that

$$P_{\theta_0}\left(\frac{f_{\theta_1}(s)}{f_{\theta_0}(s)} > c_0\right) \leq \alpha \tag{8.2.4}$$

and

$$\gamma = \begin{cases} \dfrac{\alpha - P_{\theta_0}\left(\frac{f_{\theta_1}(s)}{f_{\theta_0}(s)} > c_0\right)}{P_{\theta_0}\left(\frac{f_{\theta_1}(s)}{f_{\theta_0}(s)} = c_0\right)} & P_{\theta_0}\left(\frac{f_{\theta_1}(s)}{f_{\theta_0}(s)} = c_0\right) \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{8.2.5}$$

We use (8.2.4) and (8.2.5) to calculate c_0 and γ, and so determine the UMP size α test, in a particular problem.

Note that the test is nonrandomized whenever $P_{\theta_0}(f_{\theta_1}(s)/f_{\theta_0}(s) > c_0) = \alpha$, as then $\gamma = 0$, i.e., we categorically accept or reject H_0 after seeing the data. This always occurs whenever the distribution of $f_{\theta_1}(s)/f_{\theta_0}(s)$ is continuous when $s \sim P_{\theta_0}$. Interestingly, it can happen that the distribution of the ratio is not continuous even when the distribution of s is continuous (see Problem 8.2.17).

Before considering some applications of the Neyman–Pearson theorem, we establish the analog of the Rao–Blackwell theorem for hypothesis testing problems. Given the value of the sufficient statistic $U(s) = u$, we denote the conditional probability measure for the response s by $P(\cdot \mid U = u)$ (by Theorem 8.1.2, this probability measure does not depend on θ). For test function φ put $\varphi_U(s)$ equal to the conditional expectation of φ given the value of $U(s)$, namely,

$$\varphi_U(s) = E_{P(\cdot \mid U=U(s))}(\varphi).$$

Theorem 8.2.2 Suppose that U is a sufficient statistic and φ is a size α test function for $H_0 : \psi(\theta) = \psi_0$. Then φ_U is a size α test function for $H_0 : \psi(\theta) = \psi_0$ that depends on the data only through the value of U. Furthermore, φ and φ_U have the same power function.

PROOF It is clear that $\varphi_U(s_1) = \varphi_U(s_2)$ whenever $U(s_1) = U(s_2)$, and so φ_U depends on the data only through the value of U. Now let $P_{\theta,U}$ denote the marginal probability measure of U induced by P_θ. Then by the theorem of total expectation, we have $E_\theta(\varphi) = E_{P_{\theta,U}}(E_{P(\cdot \mid U=u)}(\varphi)) = E_{P_{\theta,U}}(\varphi_U) = E_\theta(\varphi_U)$. Now $E_\theta(\varphi) \le \alpha$ when $\psi(\theta) = \psi_0$, which implies that $E_\theta(\varphi_U) \le \alpha$ when $\psi(\theta) = \psi_0$, and $\beta(\theta) = E_\theta(\varphi) = E_\theta(\varphi_U)$ when $\psi(\theta) \ne \psi_0$. ∎

This result allows us to restrict our search for a UMP size α test to those test functions that depend on the data only through the value of a sufficient statistic.

We now consider some applications of the Neyman–Pearson theorem. The following example shows that this result can lead to solutions to much more general problems than the simple case being addressed.

EXAMPLE 8.2.2 *Optimal Hypothesis Testing in the Location Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where $\mu \in \Omega = \{\mu_0, \mu_1\}$ and $\sigma_0^2 > 0$ is known, and we want to test $H_0 : \mu = \mu_0$ versus $H_a : \mu = \mu_1$. The likelihood function is given by

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

and $\bar{x}$ is a sufficient statistic for this restricted model.

By Theorem 8.2.2, we can restrict our attention to test functions that depend on the data through $\bar{x}$. Now $\bar{X} \sim N(\mu, \sigma_0^2/n)$ so that

$$\begin{aligned}
\frac{f_{\mu_1}(\bar{x})}{f_{\mu_0}(\bar{x})} &= \frac{\exp\left(-\frac{n}{2\sigma_0^2}(\bar{x} - \mu_1)^2\right)}{\exp\left(-\frac{n}{2\sigma_0^2}(\bar{x} - \mu_0)^2\right)} \\
&= \exp\left(-\frac{n}{2\sigma_0^2}\left(\bar{x}^2 - 2\bar{x}\mu_1 + \mu_1^2 - \bar{x}^2 + 2\bar{x}\mu_0 - \mu_0^2\right)\right) \\
&= \exp\left(\frac{n}{\sigma_0^2}(\mu_1 - \mu_0)\,\bar{x}\right)\exp\left(-\frac{n}{2\sigma_0^2}\left(\mu_1^2 - \mu_0^2\right)\right).
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&P_{\mu_0}\left(\frac{f_{\mu_1}(\bar{X})}{f_{\mu_0}(\bar{X})} > c_0\right) \\
&= P_{\mu_0}\left(\exp\left(\frac{n}{\sigma_0^2}\left(\mu_1-\mu_0\right)\bar{X}\right)\exp\left(-\frac{n}{2\sigma_0^2}\left(\mu_1^2-\mu_0^2\right)\right) > c_0\right) \\
&= P_{\mu_0}\left(\exp\left(\frac{n}{\sigma_0^2}\left(\mu_1-\mu_0\right)\bar{X}\right) > c_0\exp\left(\frac{n}{2\sigma_0^2}\left(\mu_1^2-\mu_0^2\right)\right)\right) \\
&= P_{\mu_0}\left(\left(\mu_1-\mu_0\right)\bar{X} > \frac{\sigma_0^2}{n}\ln\left\{c_0\exp\left(\frac{n}{2\sigma_0^2}\left(\mu_1^2-\mu_0^2\right)\right)\right\}\right) \\
&= \begin{cases} P_{\mu_0}\left(\frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} > c_0'\right) & \mu_1 > \mu_0 \\ P_{\mu_0}\left(\frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} < c_0'\right) & \mu_1 < \mu_0, \end{cases}
\end{aligned}
$$

where

$$c_0' = \frac{\sqrt{n}}{\sigma_0}\left\{\frac{\sigma_0^2}{n\left(\mu_1-\mu_0\right)}\ln\left\{c_0\exp\left(\frac{n}{2\sigma_0^2}\left(\mu_1^2-\mu_0^2\right)\right)\right\} - \mu_0\right\}.$$

Using (8.2.4), when $\mu_1 > \mu_0$, we select c_0 so that $c_0' = z_{1-\alpha}$; when $\mu_1 < \mu_0$, we select c_0 so that $c_0' = z_\alpha$. These choices imply that

$$P_{\mu_0}\left(\frac{f_{\mu_1}(\bar{X})}{f_{\mu_0}(\bar{X})} > c_0\right) = \alpha$$

and, by (8.2.5), $\gamma = 0$.

So the UMP size α test is nonrandomized. When $\mu_1 > \mu_0$, the test is given by

$$\varphi_0\left(\bar{x}\right) = \begin{cases} 1 & \bar{x} \ge \mu_0 + \frac{\sigma_0}{\sqrt{n}}z_{1-\alpha} \\ 0 & \bar{x} < \mu_0 + \frac{\sigma_0}{\sqrt{n}}z_{1-\alpha}. \end{cases} \tag{8.2.6}$$

When $\mu_1 < \mu_0$, the test is given by

$$\varphi_0^*\left(\bar{x}\right) = \begin{cases} 1 & \bar{x} \le \mu_0 + \frac{\sigma_0}{\sqrt{n}}z_{\alpha} \\ 0 & \bar{x} > \mu_0 + \frac{\sigma_0}{\sqrt{n}}z_{\alpha}. \end{cases} \tag{8.2.7}$$

Notice that the test function in (8.2.6) does not depend on μ_1 in any way. The subsequent implication is that this test function is UMP size α for $H_0 : \mu = \mu_0$ versus $H_a : \mu = \mu_1$ for any $\mu_1 > \mu_0$. This implies that φ_0 is UMP size α for $H_0 : \mu = \mu_0$ versus the alternative $H_a : \mu > \mu_0$.

Furthermore, we have

$$\begin{aligned}\beta_{\varphi_0}(\mu) &= P_\mu\left(\bar{X} \geq \mu_0 + \frac{\sigma_0}{\sqrt{n}} z_{1-\alpha}\right) = P_\mu\left(\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \geq \frac{\mu_0-\mu}{\sigma_0/\sqrt{n}} + z_{1-\alpha}\right) \\ &= 1-\Phi\left(\frac{\mu_0-\mu}{\sigma_0/\sqrt{n}} + z_{1-\alpha}\right).\end{aligned}$$

Note that this is increasing in μ, which implies that φ_0 is a size α test function for $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$. Observe that, if φ is a size α test function for $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$, then it is also a size α test for $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$. From this, we conclude that φ_0 is UMP size α for $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$. Similarly (see Problem 8.2.12), it can be shown that φ_0^* in (8.2.7) is UMP size α for $H_0 : \mu \geq \mu_0$ versus $H_a : \mu < \mu_0$.

We might wonder if a UMP size α test exists for the two-sided problem $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$. Suppose that φ is a size α UMP test for this problem. Then φ is also size α for $H_0 : \mu = \mu_0$ versus $H_a : \mu = \mu_1$ when $\mu_1 > \mu_0$. Using Corollary 8.2.1 and the preceding developments (which also shows that there does not exist a test of the form (8.2.3) having power equal to 1 for this problem), this implies that $\varphi = \varphi_0$ (the boundary B has probability 0 here). But φ is also UMP size α for $H_0 : \mu = \mu_0$ versus $H_a : \mu = \mu_1$ when $\mu_1 < \mu_0$; thus, by the same reasoning, $\varphi = \varphi_0^*$. But clearly $\varphi_0 \neq \varphi_0^*$, so there is no UMP size α test for the two-sided problem.

Intuitively, we would expect that the size α test given by

$$\varphi(\bar{x}) = \begin{cases} 1 & \left|\frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}}\right| \geq z_{1-\alpha/2} \\ & \\ 0 & \left|\frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}}\right| < z_{1-\alpha/2} \end{cases} \tag{8.2.8}$$

would be a good test to use, but it is not UMP size α. It turns out, however, that the test in (8.2.8) is UMP size α among all tests satisfying $\beta_\varphi(\mu_0) \leq \alpha$ and $\beta_\varphi(\mu) \geq \alpha$ when $\mu \neq \mu_0$. ▮

Example 8.2.2 illustrated a hypothesis testing problem for which no UMP size α test exists. Sometimes, however, by requiring that the test possess another very natural property, we can obtain an optimal test.

Definition 8.2.3 A test φ that satisfies $\beta_\varphi(\theta) \leq \alpha$, when $\psi(\theta) = \psi_0$, and $\beta_\varphi(\theta) \geq \alpha$, when $\psi(\theta) \neq \psi_0$, is said to be an *unbiased size α test* for the hypothesis testing problem $H_0 : \psi(\theta) = \psi_0$.

So (8.2.8) is a UMP unbiased size α test. An unbiased test has the property that the probability of rejecting the null hypothesis, when the null hypothesis is false, is always greater than the probability of rejecting the null hypothesis, when the null hypothesis is true. This seems like a very reasonable property. In particular, it can be proved that any UMP size α is always an unbiased size α test (Problem 8.2.14). We do not pursue the theory of unbiased tests further in this text.

We now consider an example which shows that we cannot dispense with the use of randomized tests.

EXAMPLE 8.2.3 *Optimal Hypothesis Testing in the Bernoulli(θ) Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) distribution, where $\theta \in \Omega = \{\theta_0, \theta_1\}$, and we want to test $H_0 : \theta = \theta_0$ versus $H_a : \theta = \theta_1$, where $\theta_1 > \theta_0$. Then $n\bar{x}$ is a minimal sufficient statistic and, by Theorem 8.2.2, we can restrict our attention to test functions that depend on the data only through $n\bar{x}$.

Now $n\bar{X} \sim$ Binomial(n, θ) , so

$$\frac{f_{\theta_1}(n\bar{x})}{f_{\theta_0}(n\bar{x})} = \frac{\theta_1^{n\bar{x}}(1-\theta_1)^{n-n\bar{x}}}{\theta_0^{n\bar{x}}(1-\theta_0)^{n-n\bar{x}}} = \left(\frac{\theta_1}{\theta_0}\right)^{n\bar{x}} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-n\bar{x}}.$$

Therefore,

$$\begin{aligned}
&P_{\theta_0}\left(\frac{f_{\theta_1}(n\bar{X})}{f_{\theta_0}(n\bar{X})} > c_0\right) \\
&= P_{\theta_0}\left(\left(\frac{\theta_1}{\theta_0}\right)^{n\bar{X}} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-n\bar{X}} > c_0\right) \\
&= P_{\theta_0}\left(\left(\frac{\theta_1}{1-\theta_1}\frac{1-\theta_0}{\theta_0}\right)^{n\bar{X}} > c_0\left(\frac{1-\theta_1}{1-\theta_0}\right)^{-n}\right) \\
&= P_{\theta_0}\left(n\bar{X}\left[\ln\left(\frac{\theta_1}{1-\theta_1}\frac{1-\theta_0}{\theta_0}\right)\right] > \ln c_0\left(\frac{1-\theta_1}{1-\theta_0}\right)^{-n}\right) \\
&= P_{\theta_0}\left(n\bar{X} > \frac{\ln c_0\left(\frac{1-\theta_1}{1-\theta_0}\right)^{-n}}{\ln\left(\frac{\theta_1}{1-\theta_1}\frac{1-\theta_0}{\theta_0}\right)}\right) = P_{\theta_0}(n\bar{X} > c_0')
\end{aligned}$$

because

$$\ln\left(\frac{\theta_1}{1-\theta_1}\frac{1-\theta_0}{\theta_0}\right) > 0$$

as $\theta/(1-\theta)$ is increasing in θ, which implies $\theta_1/(1-\theta_1) > \theta_0/(1-\theta_0)$.

Now, using (8.2.4), we choose c_0 so that c_0' is an integer satisfying

$$P_{\theta_0}(n\bar{X} > c_0') \le \alpha \text{ and } P_{\theta_0}(n\bar{X} > c_0' - 1) > \alpha.$$

Because $n\bar{X} \sim$ Binomial(n, θ_0) is a discrete distribution, we see that, in general, we will not be able to achieve $P_{\theta_0}(n\bar{X} > c_0') = \alpha$ exactly. So, using (8.2.5),

$$\gamma = \frac{\alpha - P_{\theta_0}(n\bar{X} > c_0')}{P_{\theta_0}(n\bar{X} = c_0')}$$

will not be equal to 0. Then

$$\varphi_0(n\bar{x}) = \begin{cases} 1 & n\bar{x} > c_0' \\ \gamma & n\bar{x} = c_0' \\ 0 & n\bar{x} < c_0' \end{cases}$$

is UMP size α for $H_0 : \theta = \theta_0$ versus $H_a : \theta = \theta_1$. Note that we can use statistical software (or Table D.6) for the binomial distribution to obtain c_0'.

For example, suppose $n = 6$ and $\theta_0 = 0.25$. The following table gives the values of the Binomial$(6, 0.25)$ distribution function to three decimal places.

x	0	1	2	3	4	5	6
$F(x)$	0.178	0.534	0.831	0.962	0.995	1.000	1.000

Therefore, if $\alpha = 0.05$, we have that $c_0' = 3$ because $P_{0.25}\left(n\bar{X} > 3\right) = 1 - 0.962 = 0.038$ and $P_{0.25}\left(n\bar{X} > 2\right) = 1 - 0.831 = 0.169$. This implies that

$$\gamma = \frac{0.05 - (1 - 0.962)}{0.962} = 0.012.$$

So with this test, we reject $H_0 : \theta = \theta_0$ categorically if the number of successes is greater than 3, accept $H_0 : \theta = \theta_0$ categorically when the number of successes is less than 3, and when the number of 1's equals 3, we randomly reject $H_0 : \theta = \theta_0$ with probability 0.012 (e.g., generate $U \sim$ Uniform[0, 1] and reject whenever $U \le 0.012$).

Notice that the test φ_0 does not involve θ_1, so indeed it is UMP size α for $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$. Furthermore, using Problem 8.2.18, we have

$$\begin{aligned} P_\theta(n\bar{X} &> c_0') = \sum_{k=c_0'+1}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} \\ &= 1 - \frac{\Gamma(n+1)}{\Gamma\left(c_0'+1\right)\Gamma\left(n-c_0'\right)} \int_\theta^1 u^{c_0'} (1-u)^{n-c_0'-1}\, du. \end{aligned}$$

Because

$$\int_\theta^1 u^{c_0'} (1-u)^{n-c_0'-1}\, du$$

is decreasing in θ, we must have that $P_\theta(n\bar{X} > c_0')$ is increasing in θ. Arguing as in Example 8.2.2, we conclude that φ_0 is UMP size α for $H_0 : \theta \le \theta_0$ versus $H_a : \theta > \theta_0$.

Similarly, we obtain a UMP size α test for $H_0 : \theta \le \theta_0$ versus $H_a : \theta > \theta_0$. As in Example 8.2.2, there is no UMP size α test for $H_0 : \theta = \theta_0$ versus $H_a : \theta \ne \theta_0$, but there is a UMP unbiased size α test for this problem. ▮

8.2.5 Likelihood Ratio Tests (Advanced)

In the examples considered so far, the Neyman–Pearson theorem has led to solutions to problems in which H_0 or H_a are not just single values of the parameter, even though the theorem was only stated for the single-value case. We also noted, however, that this is not true in general (for example, the two-sided problems discussed in Examples 8.2.2 and 8.2.3).

The method of *generalized likelihood ratio tests* for $H_0 : \psi(\theta) = \psi_0$ has been developed to deal with the general case. This is motivated by the Neyman–Pearson theorem, for observe that in (8.2.3),

$$\frac{f_{\theta_1}(s)}{f_{\theta_0}(s)} = \frac{L(\theta_1 \,|\, s)}{L(\theta_0 \,|\, s)}.$$

Therefore, (8.2.3) can be thought of as being based on the ratio of the likelihood at θ_1 to the likelihood at θ_0, and we reject $H_0 : \theta = \theta_0$ when the likelihood gives much more support to θ_1 than to θ_0. The amount of the additional support required for rejection is determined by c_0. The larger c_0 is, the larger the likelihood $L(\theta_1 \,|\, s)$ has to be relative to $L(\theta_0 \,|\, s)$ before we reject $H_0 : \theta = \theta_0$.

Denote the overall MLE of θ by $\hat{\theta}(s)$, and the MLE, when $\theta \in H_0$, by $\hat{\theta}_{H_0}(s)$. So we have

$$L(\theta \,|\, s) \le L(\hat{\theta}_{H_0}(s) \,|\, s)$$

for all θ such that $\psi(\theta) = \psi_0$. The generalized likelihood ratio test then rejects H_0 when

$$\frac{L(\hat{\theta}(s) \,|\, s)}{L(\hat{\theta}_{H_0}(s) \,|\, s)} \tag{8.2.9}$$

is large, as this indicates evidence against H_0 being true.

How do we determine when (8.2.9) is large enough to reject? Denoting the observed data by s_0, we do this by computing the P-values

$$P_\theta\left(\frac{L(\hat{\theta}(s) \,|\, s)}{L(\hat{\theta}_{H_0}(s) \,|\, s)} > \frac{L(\hat{\theta}(s_0) \,|\, s_0)}{L(\hat{\theta}_{H_0}(s_0) \,|\, s_0)}\right) \tag{8.2.10}$$

when $\theta \in H_0$. Small values of (8.2.10) are evidence against H_0. Of course, when $\psi(\theta) = \psi_0$ for more than one value of θ, then it is not clear which value of (8.2.10) to use. It can be shown, however, that under conditions such as those discussed in Section 6.5, if s corresponds to a sample of n values from a distribution, then

$$2 \ln \frac{L(\hat{\theta}(s) \,|\, s)}{L(\hat{\theta}_{H_0}(s) \,|\, s)} \xrightarrow{D} \chi^2(\dim \Omega - \dim H_0)$$

as $n \to \infty$, whenever the true value of θ is in H_0. Here, $\dim \Omega$ and $\dim H_0$ are the dimensions of these sets. This leads us to a test that rejects H_0 whenever

$$2 \ln \frac{L(\hat{\theta}(s) \,|\, s)}{L(\hat{\theta}_{H_0}(s) \,|\, s)} \tag{8.2.11}$$

is greater than a particular quantile of the $\chi^2(\dim \Omega - \dim H_0)$ distribution.

For example, suppose that in a location-scale normal model, we are testing $H_0 : \mu = \mu_0$. Then $\Omega = R^1 \times [0, \infty)$, $H_0 = \{\mu_0\} \times [0, \infty)$, $\dim \Omega = 2$, $\dim H_0 = 1$, and, for a size 0.05 test, we reject whenever (8.2.11) is greater than $\chi^2_{0.95}(1)$. Note that, strictly speaking, likelihood ratio tests are not derived via optimality considerations. We will not discuss likelihood ratio tests further in this text.

Summary of Section 8.2

- In searching for an optimal hypothesis testing procedure, we place an upper bound on the probability of making a type I error (rejecting H_0 when it is true) and search for a test that minimizes the probability of making a type II error (accepting H_0 when it is false).
- The Neyman–Pearson theorem prescribes an optimal size α test when H_0 and H_a each specify a single value for the full parameter θ.
- Sometimes the Neyman–Pearson theorem leads to solutions to hypothesis testing problems when the null or alternative hypotheses allow for more than one possible value for θ, but in general we must resort to likelihood ratio tests for such problems.

EXERCISES

8.2.1 Suppose that a statistical model is given by the two distributions in the following table.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$
$f_a(s)$	$1/3$	$1/6$	$1/12$	$5/12$
$f_b(s)$	$1/2$	$1/4$	$1/6$	$1/12$

Determine the UMP size 0.10 test for testing $H_0 : \theta = a$ versus $H_a : \theta = b$. What is the power of this test? Repeat this with the size equal to 0.05.

8.2.2 Suppose for the hypothesis testing problem of Exercise 8.2.1, a statistician decides to generate $U \sim$ Uniform[0, 1] and reject H_0 whenever $U \leq 0.05$. Show that this test has size 0.05. Explain why this is not a good choice of test and why the test derived in Exercise 8.2.1 is better. Provide numerical evidence for this.

8.2.3 Suppose an investigator knows that an industrial process yields a response variable that follows an $N(1, 2)$ distribution. Some changes have been made in the industrial process, and the investigator believes that these have possibly made a change in the mean of the response (not the variance), increasing its value. The investigator wants the probability of a type I error occurring to be less than 1%. Determine an appropriate testing procedure for this problem based on a sample of size 10.

8.2.4 Suppose you have a sample of 20 from an $N(\mu, 1)$ distribution. You form a 0.975-confidence interval for μ and use it to test $H_0 : \mu = 0$ by rejecting H_0 whenever 0 is not in the confidence interval.

(a) What is the size of this test?

(b) Determine the power function of this test.

8.2.5 Suppose you have a sample of size $n = 1$ from a Uniform$[0, \theta]$ distribution, where $\theta > 0$ is unknown. You test $H_0 : \theta \leq 1$ by rejecting H_0 whenever the sampled value is greater than 1.

(a) What is the size of this test?

(b) Determine the power function of this test.

8.2.6 Suppose you are testing a null hypothesis $H_0 : \theta = 0$, where $\theta \in R^1$. You use a size 0.05 testing procedure and accept H_0. You feel you have a fairly large sample, but when you compute the power at ± 0.2, you obtain a value of 0.10 where 0.2 represents the smallest difference from 0 that is of practical importance. Do you believe it makes sense to conclude that the null hypothesis is true? Justify your conclusion.

8.2.7 Suppose you want to test the null hypothesis $H_0 : \mu = 0$ based on a sample of n from an $N(\mu, 1)$ distribution, where $\mu \in \{0, 2\}$. How large does n have to be so that the power at $\mu = 2$, of the optimal size 0.05 test, is equal to 0.99?

8.2.8 Suppose we have available two different test procedures in a problem and these have the same power function. Explain why, from the point of view of optimal hypothesis testing theory, we should not care which test is used.

8.2.9 Suppose you have a UMP size α test φ for testing the hypothesis $H_0 : \psi(\theta) = \psi_0$, where ψ is real-valued. Explain how the graph of the power function of another size α test that was not UMP would differ from the graph of the power function of φ.

COMPUTER EXERCISES

8.2.10 Suppose you have a coin and you want to test the hypothesis that the coin is fair, i.e., you want to test $H_0 : \theta = 1/2$ where θ is the probability of getting a head on a single toss. You decide to reject H_0 using the rejection region $R = \{0, 1, 7, 8\}$ based on $n = 10$ tosses. Tabulate the power function for this procedure for $\theta \in \{0, 1/8, 2/8, \ldots, 7/8, 1\}$.

8.2.11 On the same graph, plot the power functions for the two-sided z-test of $H_0 : \mu = 0$ for samples of sizes $n = 1, 4, 10, 20$, and 100 based on $\alpha = 0.05$.

(a) What do you observe about these graphs?

(b) Explain how these graphs demonstrate the unbiasedness of this test.

PROBLEMS

8.2.12 Prove that φ_0^* in (8.2.7) is UMP size α for $H_0 : \mu \geq \mu_0$ versus $H_a : \mu < \mu_0$.

8.2.13 Prove that the test function $\varphi(s) = \alpha$ for every $s \in S$ is an exact size α test function. What is the interpretation of this test function?

8.2.14 Using the test function in Problem 8.2.13, show that a UMP size α test is also a UMP unbiased size α test.

8.2.15 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Gamma(α_0, β) distribution, where α_0 is known and $\beta > 0$ is unknown. Determine the UMP size α test for testing $H_0 : \beta = \beta_0$ versus $H_a : \beta = \beta_1$, where $\beta_1 > \beta_0$. Is this test UMP size α for $H_0 : \beta \leq \beta_0$ versus $H_a : \beta > \beta_0$?

8.2.16 Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu_0, \sigma^2)$ distribution, where μ_0 is known and $\sigma^2 > 0$ is unknown. Determine the UMP size α test for testing $H_0 : \sigma^2 = \sigma_0^2$ versus $H_a : \sigma^2 = \sigma_1^2$ where $\sigma_0^2 < \sigma_1^2$. Is this test UMP size α for $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_a : \sigma^2 > \sigma_0^2$?

8.2.17 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Uniform$[0, \theta]$ distribution, where $\theta > 0$ is unknown. Determine the UMP size α test for testing $H_0 : \theta = \theta_0$ versus $H_a : \theta = \theta_1$, where $\theta_0 < \theta_1$. Is this test function UMP size α for $H_0 : \theta \leq \theta_0$ versus $H_a : \theta > \theta_0$?

8.2.18 Suppose that F is the distribution function for the Binomial(n, θ) distribution. Then prove that

$$F(x) = \frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x)} \int_{\theta}^{1} y^{x}(1-y)^{n-x-1}\, dy$$

for $x = 0, 1, \ldots, n-1$. This establishes a relationship between the binomial probability distribution and the beta function. (Hint: Integration by parts.)

8.2.19 Suppose that F is the distribution function for the Poisson(λ) distribution. Then prove that

$$F(x) = \frac{1}{x!} \int_{\lambda}^{\infty} y^{x} e^{-y}\, dy$$

for $x = 0, 1, \ldots$. This establishes a relationship between the Poisson probability distribution and the gamma function. (Hint: Integration by parts.)

8.2.20 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Poisson(λ) distribution, where $\lambda > 0$ is unknown. Determine the UMP size α test for $H_0 : \lambda = \lambda_0$ versus $H_a : \lambda = \lambda_1$, where $\lambda_0 < \lambda_1$. Is this test function UMP size α for $H_0 : \lambda \le \lambda_0$ versus $H_a : \lambda > \lambda_0$? (Hint: You will need the result of Problem 8.2.19.)

8.2.21 Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $(\mu, \sigma^2) \in R^1 \times (0, \infty)$ is unknown. Derive the form of the exact size α likelihood ratio test for testing $H_0 : \mu = \mu_0$ versus $H_0 : \mu \neq \mu_0$.

8.2.22 (*Optimal confidence intervals*) Suppose that for model $\{f_\theta : \theta \in \Omega\}$ we have a UMP size α test function φ_{ψ_0} for $H_0 : \psi(\theta) = \psi_0$, for each possible value of ψ_0. Suppose further that each φ_{ψ_0} only takes values in $\{0, 1\}$, i.e., each φ_{ψ_0} is a nonrandomized size α test function.

(a) Prove that

$$C(s) = \{\psi_0 : \varphi_{\psi_0}(s) = 0\}$$

satisfies

$$P_\theta(\psi(\theta) \in C(s)) \ge 1 - \alpha$$

for every $\theta \in \Omega$. Conclude that $C(s)$ is a $(1 - \alpha)$-confidence set for $\psi(\theta)$.

(b) If C^* is a $(1 - \alpha)$-confidence set for $\psi(\theta)$, then prove that the test function defined by

$$\varphi^*_{\psi_0}(s) = \begin{cases} 1 & \psi_0 \notin C(s) \\ \\ 0 & \psi_0 \in C(s) \end{cases}$$

is size α for $H_0 : \psi(\theta) = \psi_0$.

(c) Suppose that for each value ψ_0, the test function φ_{ψ_0} is UMP size α for testing $H_0 : \psi(\theta) = \psi_0$ versus $H_0 : \psi(\theta) \neq \psi_0$. Then prove that

$$P_\theta(\psi(\theta^*) \in C(s)) \qquad (8.2.12)$$

is minimized, when $\psi(\theta) \neq \psi_0$, among all $(1 - \alpha)$-confidence sets for $\psi(\theta)$. The probability (8.2.12) is the probability of C containing the false value $\psi(\theta^*)$, and a

$(1-\alpha)$-confidence region that minimizes this probability when $\psi(\theta) \neq \psi_0$ is called a *uniformly most accurate (UMA)* $(1-\alpha)$-confidence region for $\psi(\theta)$.

CHALLENGES

8.2.23 Prove Corollary 8.2.1 in the discrete case.

8.3 Optimal Bayesian Inferences

We now add the prior probability measure Π with density π. As we will see, this completes the specification of an optimality problem, as now there is always a solution. Solutions to Bayesian optimization problems are known as *Bayes rules*.

In Section 8.1, the unrestricted optimization problem was to find the estimator T of $\psi(\theta)$ that minimizes $\text{MSE}_\theta(T) = E_\theta((T-\psi(\theta))^2)$, for each $\theta \in \Omega$. The Bayesian version of this problem is to minimize

$$E_\Pi(\text{MSE}_\theta(T)) = E_\Pi(E_\theta((T-\psi(\theta))^2)). \tag{8.3.1}$$

By the theorem of total expectation (Theorem 3.5.2), (8.3.1) is the expected value of the squared error $(T(s)-\psi(\theta))^2$ under the joint distribution on (θ, s) induced by the conditional distribution for s, given θ (the sampling model), and by the marginal distribution for θ (the prior distribution of θ). Again, by the theorem of total expectation, we can write this as

$$E_\Pi(\text{MSE}_\theta(T)) = E_M(E_{\Pi(\cdot\,|\,s)}((T-\psi(\theta))^2)), \tag{8.3.2}$$

where $\Pi(\cdot\,|\,s)$ denotes the posterior probability measure for θ, given the data s (the conditional distribution of θ given s), and M denotes the prior predictive probability measure for s (the marginal distribution of s).

We have the following result.

Theorem 8.3.1 When (8.3.1) is finite, a Bayes rule is given by

$$T(s) = E_{\Pi(\cdot\,|\,s)}\left(\psi(\theta)\right),$$

namely, the posterior expectation of $\psi(\theta)$.

PROOF First, consider the expected posterior squared error

$$E_{\Pi(\cdot\,|\,s)}\left((T'(s)-\psi(\theta))^2\right)$$

of an estimate $T'(s)$. By Theorem 8.1.1 this is minimized by taking $T'(s)$ equal to $T(s) = E_{\Pi(\cdot\,|\,s)}\left(\psi(\theta)\right)$ (note that the "random" quantity here is θ).

Now suppose that T' is any estimator of $\psi(\theta)$. Then we have just shown that

$$0 \le E_{\Pi(\cdot\,|\,s)}\left((T(s)-\psi(\theta))^2\right) \le E_{\Pi(\cdot\,|\,s)}\left((T'(s)-\psi(\theta))^2\right)$$

and thus,

$$\begin{aligned} E_{\Pi}\,(\mathrm{MSE}_{\theta}(T)) &= E_M\left(E_{\Pi(\cdot\,|\,s)}((T(s)-\psi(\theta))^2)\right) \\ &\le E_M\left(E_{\Pi(\cdot\,|\,s)}((T'(s)-\psi(\theta))^2)\right) = E_{\Pi}(\mathrm{MSE}_{\theta}(T')). \end{aligned}$$

Therefore, T minimizes (8.3.1) and is a Bayes rule. ▮

So we see that, under mild conditions, the optimal Bayesian estimation problem always has a solution and there is no need to restrict ourselves to unbiased estimators, etc.

For the hypothesis testing problem $H_0 : \psi(\theta) = \psi_0$, we want to find the test function φ that minimizes the prior probability of making an error (type I or type II). Such a φ is a Bayes rule. We have the following result.

Theorem 8.3.2 A Bayes rule for the hypothesis testing problem $H_0 : \psi(\theta) = \psi_0$ is given by

$$\varphi_0(s) = \begin{cases} 1 & \Pi(\{\psi(\theta) = \psi_0\}\,|\,s) \le \Pi(\{\psi(\theta) \ne \psi_0\}\,|\,s) \\ 0 & \text{otherwise.} \end{cases}$$

PROOF Consider test function φ and let $I_{\{\psi(\theta)=\psi_0\}}(\theta)$ denote the indicator function of the set $\{\theta : \psi(\theta) = \psi_0\}$ (so $I_{\{\psi(\theta)=\psi_0\}}(\theta) = 1$ when $\psi(\theta) = \psi_0$ and equals 0 otherwise). Observe that $\varphi(s)$ is the probability of rejecting H_0, having observed s, which is an error when $I_{\{\psi(\theta)=\psi_0\}}(\theta) = 1$; $1-\varphi(s)$ is the probability of accepting H_0, having observed s, which is an error when $I_{\{\psi(\theta)=\psi_0\}}(\theta) = 0$. Therefore, given s and θ, the probability of making an error is

$$e(\theta, s) = \varphi(s)I_{\{\psi(\theta)=\psi_0\}}(\theta) + (1-\varphi(s))\,(1 - I_{\{\psi(\theta)=\psi_0\}}(\theta)).$$

By the theorem of total expectation, the prior probability of making an error (taking the expectation of $e(\theta, s)$ under the joint distribution of (θ, s)) is

$$E_M\left(E_{\Pi(\cdot\,|\,s)}\,(e(\theta, s))\right). \tag{8.3.3}$$

As in the proof of Theorem 8.3.1, if we can find φ that minimizes $E_{\Pi(\cdot\,|\,s)}\,(e(\theta, s))$ for each s, then φ also minimizes (8.3.3) and is a Bayes rule.

Using Theorem 3.5.4 to pull $\varphi(s)$ through the conditional expectation, and the fact that $E_{\Pi(\cdot\,|\,s)}\,(I_A(\theta)) = \Pi\,(A\,|\,s)$ for any event A, then

$$E_{\Pi(\cdot\,|\,s)}\,(e(\theta, s)) = \varphi(s)\Pi(\{\psi(\theta) = \psi_0\}\,|\,s) + (1-\varphi(s))\,(1 - \Pi(\{\psi(\theta) = \psi_0\}\,|\,s)).$$

Because $\varphi(s) \in [0, 1]$, we have

$$\begin{aligned} &\min\{\Pi(\{\psi(\theta) = \psi_0\}\,|\,s), 1 - \Pi(\{\psi(\theta) = \psi_0\}\,|\,s)\} \\ &\le \varphi(s)\Pi(\{\psi(\theta) = \psi_0\}\,|\,s) + (1-\varphi(s))\,(1 - \Pi(\{\psi(\theta) = \psi_0\}\,|\,s)). \end{aligned}$$

Therefore, the minimum value of $E_{\Pi(\cdot\,|\,s)}\,(e(\theta, s))$ is attained by $\varphi(s) = \varphi_0(s)$. ∎

Observe that Theorem 8.3.2 says that the Bayes rule rejects H_0 whenever the posterior probability of the null hypothesis is less than or equal to the posterior probability of the alternative. This is an intuitively satisfying result.

The following problem does arise with this approach, however. We have

$$\begin{aligned}\Pi(\{\psi(\theta) = \psi_0\}\,|\,s) &= \frac{E_\Pi(I_{\{\theta:\psi(\theta)=\psi_0\}}(\theta)\, f_\theta(s))}{m(s)} \\ &\le \frac{\max_{\{\theta:\psi(\theta)=\psi_0\}} f_\theta(s)\Pi(\{\psi(\theta) = \psi_0\})}{m(s)}. \qquad (8.3.4)\end{aligned}$$

When $\Pi(\{\psi(\theta) = \psi_0\}) = 0$, (8.3.4) implies that $\Pi(\{\psi(\theta) = \psi_0\}\,|\,s = 0)$ for every s. Therefore, using the Bayes rule, we would always reject H_0 no matter what data s are obtained, which does not seem sensible. As discussed in Section 7.2.3, we have to be careful to make sure we use a prior Π that assigns positive mass to H_0 if we are going to use the optimal Bayes approach to a hypothesis testing problem.

Summary of Section 8.3

- Optimal Bayesian procedures are obtained by minimizing the expected performance measure using the posterior distribution.
- In estimation problems, when using squared error as the performance measure, the posterior mean is optimal.
- In hypothesis testing problems, when minimizing the probability of making an error as the performance measure, then computing the posterior probability of the null hypothesis and accepting H_0 when this is greater than 1/2 is optimal.

EXERCISES

8.3.1 Suppose that $S = \{1, 2, 3\}, \Omega = \{1, 2\}$, with data distributions given by the following table. We place a uniform prior on θ and want to estimate θ.

	$s = 1$	$s = 2$	$s = 3$
$f_1(s)$	$1/6$	$1/6$	$2/3$
$f_2(s)$	$1/4$	$1/4$	$1/2$

Using a Bayes rule, test the hypothesis $H_0 : \theta = 2$ when $s = 2$ is observed.

8.3.2 For the situation described in Exercise 8.3.1, determine the Bayes rule estimator of θ when using expected squared error as our performance measure for estimators.

8.3.3 Suppose that we have a sample $(x_1, \ldots, x_n)$ from an $N(\mu, \sigma_0^2)$ distribution, where μ is unknown and σ_0^2 is known, and we want to estimate μ using expected squared error as our performance measure for estimators. If we use the prior distribution $\mu \sim N(\mu, \tau_0^2)$, then determine the Bayes rule for this problem. Determine the limiting Bayes rule as $\tau_0 \to \infty$.

8.3.4 Suppose that we observe a sample $(x_1, \ldots, x_n)$ from a Bernoulli(θ) distribution, where θ is completely unknown, and we want to estimate θ using expected squared error as our performance measure for estimators. If we use the prior distribution $\theta \sim$ Beta(α, β), then determine a Bayes rule for this problem.

8.3.5 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Gamma(α_0, β) distribution, where α_0 is known, and $\beta \sim$ Gamma(τ_0, υ_0), where τ_0 and υ_0 are known. If we want to estimate β using expected squared error as our performance measure for estimators, then determine the Bayes rule. Use the weak (or strong) law of large numbers to determine what this estimator converges to as $n \to \infty$.

8.3.6 For the situation described in Exercise 8.3.5, determine the Bayes rule for estimating β^{-1} when using expected squared error as our performance measure for estimators.

8.3.7 Suppose that we have a sample $(x_1, \ldots, x_n)$ from an $N(\mu, \sigma_0^2)$ distribution, where μ is unknown and σ_0^2 is known, and we want to find the test of $H_0 : \mu = \mu_0$ that minimizes the prior probability of making an error (type I or type II). If we use the prior distribution $\mu \sim p_0 I_{\{\mu_0\}} + (1 - p_0) N(\mu_0, \tau_0^2)$, where $p_0 \in (0, 1)$ is known (i.e., the prior is a mixture of a distribution degenerate at μ_0 and an $N(\mu_0, \tau_0^2)$ distribution), then determine the Bayes rule for this problem. Determine the limiting Bayes rule as $\tau_0 \to \infty$. (Hint: Make use of the computations in Example 7.2.13.)

8.3.8 Suppose that we have a sample $(x_1, \ldots, x_n)$ from a Bernoulli(θ) distribution, where θ is unknown, and we want to find the test of $H_0 : \theta = \theta_0$ that minimizes the prior probability of making an error (type I or type II). If we use the prior distribution $\theta \sim p_0 I_{\{\theta_0\}} + (1 - p_0)$Uniform$[0, 1]$, where $p_0 \in (0, 1)$ is known (i.e., the prior is a mixture of a distribution degenerate at θ_0 and a uniform distribution), then determine the Bayes rule for this problem.

PROBLEMS

8.3.9 Suppose that $\Omega = \{\theta_1, \theta_2\}$, that we put a prior π on Ω, and that we want to estimate θ. Suppose our performance measure for estimators is the probability of making an incorrect choice of θ. If the model is denoted $\{f_\theta : \theta \in \Omega\}$, then obtain the form of the Bayes rule when data s are observed.

8.3.10 For the situation described in Exercise 8.3.1, use the Bayes rule obtained via the method of Problem 8.3.9 to estimate θ when $s = 2$. What advantage does this estimate have over that obtained in Exercise 8.3.2?

8.3.11 Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution where $(\mu, \sigma^2) \in R^1 \times (0, \infty)$ is unknown, and want to estimate μ using expected squared error as our performance measure for estimators. Using the prior distribution given by

$$\mu \,|\, \sigma^2 \sim N(\mu_0, \tau_0^2 \sigma^2),$$

and using

$$\frac{1}{\sigma^2} \sim \text{Gamma}\left(\alpha_0, \beta_0\right),$$

where μ_0, τ_0^2, α_0, and β_0 are fixed and known, then determine the Bayes rule for μ.

8.3.12 (*Model selection*) Generalize Problem 8.3.9 to the case $\Omega = \{\theta_1, \ldots, \theta_k\}$.

CHALLENGES

8.3.13 In Section 7.2.4, we described the Bayesian prediction problem. Using the notation found there, suppose we wish to predict $t \in R^1$ using a predictor $\tilde{T}(s)$. If we assess the accuracy of a predictor by

$$E((\tilde{T}(s) - t)^2) = E_{\Pi}(E_{P_\theta}(E_{Q_\theta(\cdot\,|\,s)}((\tilde{T}(s) - t)^2))),$$

then determine the prior predictor that minimizes this quantity (assume all relevant expectations are finite). If we observe s_0, then determine the best predictor. (Hint: Assume all the probability measures are discrete.)

8.4 Decision Theory (Advanced)

To determine an optimal inference, we chose a performance measure and then attempted to find an inference, of a given type, that has optimal performance with respect to this measure. For example, when considering estimates of a real-valued characteristic of interest $\psi(\theta)$, we took the performance measure to be MSE and then searched for the estimator that minimizes this for each value of θ.

Decision theory is closely related to the optimal approach to deriving inferences, but it is a little more specialized. In the decision framework, we take the point of view that, in any statistical problem, the statistician is faced with making a decision, e.g., deciding on a particular value for $\psi(\theta)$. Furthermore, associated with a decision is the notion of a loss incurred whenever the decision is incorrect. A decision rule is a procedure, based on the observed data s, that the statistician uses to select a decision. The decision problem is then to find a decision rule that minimizes the average loss incurred.

There are a number of real-world contexts in which losses are an obvious part of the problem, e.g., the monetary losses associated with various insurance plans that an insurance company may consider offering. So the decision theory approach has many applications. It is clear in many practical problems, however, that losses (as well as performance measures) are somewhat arbitrary components of a statistical problem, often chosen simply for convenience. In such circumstances, the approaches to deriving inferences described in Chapters 6 and 7 are preferred by many statisticians.

So the *decision theory model* for inference adds another ingredient to the sampling model (or to the sampling model and prior) to derive inferences — the loss function. To formalize this, we conceive of a set of possible actions or decisions that the statistician could take after observing the data s. This set of possible actions is denoted by $\mathcal{A}$ and is called the *action space*. To connect these actions with the statistical model, there is a *correct action function* $A : \Omega \to \mathcal{A}$ such that $A(\theta)$ is the correct action to take when θ is the true value of the parameter. Of course, because we do not know θ, we do not know the correct action $A(\theta)$, so there is uncertainty involved in our decision. Consider a simple example.

EXAMPLE 8.4.1

Suppose you are told that an urn containing 100 balls has either 50 white and 50 black balls or 60 white and 40 black balls. Five balls are drawn from the urn without replacement and their colors are observed. The statistician's job is to make a decision about the true proportion of white balls in the urn based on these data.

The statistical model then comprises two distributions $\{P_1, P_2\}$ where, using parameter space $\Omega = \{1, 2\}$, P_1 is the Hypergeometric(100, 50, 5) distribution (see Example 2.3.7) and P_2 is the Hypergeometric(100, 60, 5) distribution. The action space is $\mathcal{A} = \{0.5, 0.6\}$, and $A : \Omega \to \mathcal{A}$ is given by $A(1) = 0.5$ and $A(2) = 0.6$. The data are given by the colors of the five balls drawn. ∎

We suppose now that there is also a loss or penalty $L(\theta, a)$ incurred when we select action $a \in \mathcal{A}$ and θ is true. If we select the correct action, then the loss is 0; it is greater than 0 otherwise.

Definition 8.4.1 A *loss function* is a function L defined on $\Omega \times \mathcal{A}$ and taking values in $[0, \infty)$ such that $L(\theta, a) = 0$ if and only if $a = A(\theta)$.

Sometimes the loss can be an actual monetary loss. Actually, decision theory is a little more general than what we have just described, as we can allow for negative losses (gains or profits), but the restriction to nonnegative losses is suitable for purely statistical applications.

In a specific problem, the statistician chooses a loss function that is believed to lead to reasonable statistical procedures. This choice is dependent on the particular application. Consider some examples.

EXAMPLE 8.4.2 (*Example 8.4.1 continued*)

Perhaps a sensible choice in this problem would be

$$L(\theta, a) = \begin{cases} 1 & \theta = 1, a = 0.6 \\ 2 & \theta = 2, a = 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

Here we have decided that selecting $a = 0.5$ when it is not correct is a more serious error than selecting $a = 0.6$ when it is not correct. If we want to treat errors symmetrically, then we could take

$$L(\theta, a) = I_{\{(1,0.6),(2,0.5)\}}(\theta, a),$$

i.e., the losses are 1 or 0. ∎

EXAMPLE 8.4.3 *Estimation as a Decision Problem*

Suppose we have a marginal parameter $\psi(\theta)$ of interest, and we want to specify an estimate $T(s)$ after observing $s \in S$. Here, the action space is $\mathcal{A} = \{\psi(\theta) : \theta \in \Omega\}$ and $A(\theta) = \psi(\theta)$. Naturally, we want $T(s) \in \mathcal{A}$.

For example, suppose $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $(\mu, \sigma^2) \in \Omega = R^1 \times R^+$ is unknown, and we want to estimate $\psi(\mu, \sigma^2) = \mu$. In this case, $\mathcal{A} = R^1$ and a possible estimator is the sample average $T(x_1, \ldots, x_n) = \bar{x}$.

There are many possible choices for the loss function. Perhaps a natural choice is to use

$$L(\theta, a) = |\psi(\theta) - a|, \tag{8.4.1}$$

the absolute deviation between $\psi(\theta)$ and a. Alternatively, it is common to use

$$L(\theta, a) = (\psi(\theta) - a)^2, \tag{8.4.2}$$

the squared deviations between $\psi(\theta)$ and a.

We refer to (8.4.2) as *squared error loss*. Notice that (8.4.2) is just the square of the Euclidean distance between $\psi(\theta)$ and a. It might seem more natural to actually use the distance (8.4.1) as the loss function. It turns out, however, that there are a number of mathematical conveniences that arise from using squared distance. ∎

EXAMPLE 8.4.4 *Hypothesis Testing as a Decision Problem*
In this problem, we have a characteristic of interest $\psi(\theta)$ and want to assess the plausibility of the value ψ_0 after viewing the data s. In a hypothesis testing problem, this is written as $H_0 : \psi(\theta) = \psi_0$ versus $H_a : \psi(\theta) \neq \psi_0$. As in Section 8.2, we refer to H_0 as the null hypothesis and to H_a as the alternative hypothesis.

The purpose of a hypothesis testing procedure is to decide which of H_0 or H_a is true based on the observed data s. So in this problem, the action space is $\mathcal{A} = \{H_0, H_a\}$ and the correct action function is

$$A(\theta) = \begin{cases} H_0 & \psi(\theta) = \psi_0 \\ H_a & \psi(\theta) \neq \psi_0. \end{cases}$$

An alternative, and useful, way of thinking of the two hypotheses is as subsets of Ω. We write $H_0 = \psi^{-1}\{\psi_0\}$ as the subset of all θ values that make the null hypothesis true, and $H_a = H_0^c$ is the subset of all θ values that make the null hypothesis false. Then, based on the data s, we want to decide if the true value of θ is in H_0 or if θ is in H_a. If H_0 (or H_a) is composed of a single point, then it is called a *simple hypothesis* or a *point hypothesis*; otherwise, it is referred to as a *composite hypothesis*.

For example, suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution where $\theta = (\mu, \sigma^2) \in \Omega = R^1 \times R^+$, $\psi(\theta) = \mu$, and we want to test the null hypothesis $H_0 : \mu = \mu_0$ versus the alternative $H_a : \mu \neq \mu_0$. Then $H_0 = \{\mu_0\} \times R^+$ and $H_a = \{\mu_0\}^c \times R^+$. For the same model, let

$$\psi(\theta) = I_{(-\infty, \mu_0] \times R^+}(\mu, \sigma^2),$$

i.e., ψ is the indicator function for the subset $(-\infty, \mu_0] \times R^+$. Then testing $H_0 : \psi = 1$ versus the alternative $H_a : \psi = 0$ is equivalent to testing that the mean is less than or equal to μ_0 versus the alternative that it is greater than μ_0. This one-sided hypothesis testing problem is often denoted as $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$.

There are a number of possible choices for the loss function, but the most commonly used is of the form

$$L(\theta, a) = \begin{cases} 0 & \theta \in H_0, a = H_0 \text{ or } \theta \in H_a, a = H_a \\ b & \theta \notin H_0, a = H_0 \\ c & \theta \notin H_a, a = H_a. \end{cases}$$

If we reject H_0 when H_0 is true (a type I error), we incur a loss of c; if we accept H_0 when H_0 is false (a type II error), we incur a loss of b. When $b = c$, we can take $b = c = 1$ and produce the commonly used *0–1 loss function*. ∎

A statistician faced with a decision problem — i.e., a model, action space, correct action function, and loss function — must now select a rule for choosing an element of the action space $\mathcal{A}$ when the data s are observed. A *decision function* is a procedure that specifies how an action is to be selected in the action space $\mathcal{A}$.

Definition 8.4.2 A *nonrandomized decision function* d is a function $d : S \to \mathcal{A}$.

So after observing s, we decide that the appropriate action is $d(s)$.

Actually, we will allow our decision procedures to be a little more general than this, as we permit a random choice of an action after observing s.

Definition 8.4.3 A *decision function* δ is such that $\delta(s, \cdot)$ is a probability measure on the action space $\mathcal{A}$ for each $s \in S$ (so $\delta(s, A)$ is the probability that the action taken is in $A \subset \mathcal{A}$).

Operationally, after observing s, a random mechanism with distribution specified by $\delta(s, \cdot)$ is used to select the action from the set of possible actions. Notice that if $\delta(s, \cdot)$ is a probability measure degenerate at the point $d(s)$ (so $\delta(s, \{d(s)\}) = 1$) for each s, then δ is equivalent to the nonrandomized decision function d and conversely (see Problem 8.4.8).

The use of randomized decision procedures may seem rather unnatural, but, as we will see, sometimes they are an essential ingredient of decision theory. In many estimation problems, the use of randomized procedures provides no advantage, but this is not the case in hypothesis testing problems. We let D denote the set of all decision functions δ for the specific problem of interest.

The decision problem is to choose a decision function $\delta \in D$. The selected δ will then be used to generate decisions in applications. We base this choice on how the various decision functions δ perform with respect to the loss function. Intuitively, we want to choose δ to make the loss as small as possible. For a particular δ, because $s \sim f_\theta$ and $a \sim \delta(s, \cdot)$, the loss $L(\theta, a)$ is a random quantity. Therefore, rather than minimizing specific losses, we speak instead about minimizing some aspect of the distribution of the losses for each $\theta \in \Omega$. Perhaps a reasonable choice is to minimize the average loss. Accordingly, we define the risk function associated with $\delta \in D$ as the average loss incurred by δ. The risk function plays a central role in determining an appropriate decision function for a problem.

Definition 8.4.4 The *risk function* associated with decision function δ is given by

$$R_\delta(\theta) = E_\theta(E_{\delta(s,\cdot)}(L(\theta, a))). \tag{8.4.3}$$

Notice that to calculate the risk function we first calculate the average of $L(\theta, a)$, based on s fixed and $a \sim \delta(s, \cdot)$. Then we average this conditional average with respect to $s \sim f$. By the theorem of total expectation, this is the average loss. When $\delta(s, \cdot)$ is degenerate at $d(s)$ for each $s,$ then (8.4.3) simplifies (see Problem 8.4.8) to

$$R_\delta(\theta) = E_\theta(L(\theta, d(s))).$$

Consider the following examples.

EXAMPLE 8.4.5

Suppose that $S = \{1, 2, 3\}$, $\Omega = \{1, 2\}$, and the distributions are given by the following table.

	$s = 1$	$s = 2$	$s = 3$
$f_1(s)$	$1/3$	$1/3$	$1/3$
$f_2(s)$	$1/2$	$1/2$	0

Further suppose that $\mathcal{A} = \Omega$, $A(\theta) = \theta$, and the loss function is given by $L(\theta, a) = 1$ when $\theta \neq a$ but is 0 otherwise.

Now consider the decision function δ specified by the following table.

	$a = 1$	$a = 2$
$\delta(1, \{a\})$	$1/4$	$3/4$
$\delta(2, \{a\})$	$1/4$	$3/4$
$\delta(3, \{a\})$	1	0

So when we observe $s = 1$, we randomly choose the action $a = 1$ with probability 1/4 and choose the action $a = 2$ with probability 3/4, etc. Notice that this decision function does the sensible thing and selects the decision $a = 1$ when we observe $s = 3,$ as we know unequivocally that $\theta = 1$ in this case.

We have

$$\begin{aligned} E_{\delta(1,\cdot)}(L(\theta, a)) &= \frac{1}{4}L(\theta, 1) + \frac{3}{4}L(\theta, 2) \\ E_{\delta(2,\cdot)}(L(\theta, a)) &= \frac{1}{4}L(\theta, 1) + \frac{3}{4}L(\theta, 2) \\ E_{\delta(3,\cdot)}(L(\theta, a)) &= L(\theta, 1), \end{aligned}$$

so the risk function of δ is then given by

$$\begin{aligned} R_\delta(1) &= E_1(E_{\delta(s,\cdot)}(L(1, a))) \\ &= \frac{1}{3}\left(\frac{1}{4}L(1,1) + \frac{3}{4}L(1,2)\right) + \frac{1}{3}\left(\frac{1}{4}L(1,1) + \frac{3}{4}L(1,2)\right) + \frac{1}{3}L(1,1) \\ &= \frac{3}{12} + \frac{3}{12} + 0 = \frac{1}{2} \end{aligned}$$

and

$$\begin{aligned} R_\delta(2) &= E_2(E_{\delta(s,\cdot)}(L(2, a))) \\ &= \frac{1}{2}\left(\frac{1}{4}L(2,1) + \frac{3}{4}L(2,2)\right) + \frac{1}{2}\left(\frac{1}{4}L(2,1) + \frac{3}{4}L(2,2)\right) + 0L(2,1) \\ &= \frac{1}{8} + \frac{1}{8} + 0 = \frac{1}{4}. \blacksquare \end{aligned}$$

EXAMPLE 8.4.6 *Estimation*
We will restrict our attention to nonrandomized decision functions and note that these are also called estimators. The risk function associated with estimator T and loss function (8.4.1) is given by

$$R_T(\theta) = E_\theta\left(|\psi(\theta) - T|\right)$$

and is called the *mean absolute deviation* (MAD). The risk function associated with the estimator T and loss function (8.4.2) is given by

$$R_T(\theta) = E_\theta((\psi(\theta) - T)^2)$$

and is called the MSE.

We want to choose the estimator T to minimize $R_T(\theta)$ for every $\theta \in \Omega$. Note that, when using (8.4.2), this decision problem is exactly the same as the optimal estimation problem discussed in Section 8.1. ▮

EXAMPLE 8.4.7 *Hypothesis Testing*
We note that for a given decision function δ for this problem, and a data value s, the distribution $\delta(s, \cdot)$ is characterized by $\varphi(s) = \delta(s, H_a)$, which is the probability of rejecting H_0 when s has been observed. This is because the probability measure $\delta(s, \cdot)$ is concentrated on two points, so we need only give its value at one of these to completely specify it. We call φ the *test function* associated with δ and observe that a decision function for this problem is also specified by a test function φ.

We have immediately that

$$E_{\delta(s,\cdot)}(L(\theta, a)) = (1 - \varphi(s))\, L(\theta, H_0) + \varphi(s) L(\theta, H_a). \tag{8.4.4}$$

Therefore, when using the 0–1 loss function,

$$\begin{aligned} R_\delta(\theta) &= E_\theta\left((1 - \varphi(s))\, L(\theta, H_0) + \varphi(s) L(\theta, H_a)\right) \\ &= L(\theta, H_0) + E_\theta(\varphi(s))\left(L(\theta, H_a) - L(\theta, H_0)\right) \\ &= \begin{cases} E_\theta(\varphi(s)) & \theta \in H_0 \\ 1 - E_\theta(\varphi(s)) & \theta \in H_a. \end{cases} \end{aligned}$$

Recall that in Section 6.3.6, we introduced the power function associated with a hypothesis assessment procedure that rejected H_0 whenever the P-value was smaller than some prescribed value. The power function, evaluated at θ, is the probability that such a procedure rejects H_0 when θ is the true value. Because $\varphi(s)$ is the conditional probability, given s, that H_0 is rejected, the theorem of total expectation implies that $E_\theta(\varphi(s))$ equals the unconditional probability that we reject H_0 when θ is the true value. So in general, we refer to the function

$$\beta_\varphi(\theta) = E_\theta(\varphi(s))$$

as the *power function* of the decision procedure δ or, equivalently, as the power function of the test function φ.

Therefore, minimizing the risk function in this case is equivalent to choosing φ to minimize $\beta_\varphi(\theta)$ for every $\theta \in H_0$ and to maximize $\beta_\varphi(\theta)$ for every $\theta \in H_a$. Accordingly, this decision problem is exactly the same as the optimal inference problem discussed in Section 8.2. ▮

Once we have written down all the ingredients for a decision problem, it is then clear what form a solution to the problem will take. In particular, any decision function δ_0 that satisfies

$$R_{\delta_0}(\theta) \le R_\delta(\theta)$$

for every $\theta \in \Omega$ and $\delta \in D$ is an *optimal decision function* and is a solution. If two decision functions have the same risk functions, then, from the point of view of decision theory, they are equivalent. So it is conceivable that there might be more than one solution to a decision problem.

Actually, it turns out that an optimal decision function exists only in extremely unrealistic cases, namely, the data always tell us categorically what the correct decision is (see Problem 8.4.9). We do not really need statistical inference for such situations. For example, suppose we have two coins — coin A has two heads and coin B has two tails. As soon as we observe an outcome from a coin toss, we know exactly which coin was tossed and there is no need for statistical inference.

Still, we can identify some decision rules that we do not want to use. For example, if $\delta \in D$ is such that there exists $\delta_0 \in D$ satisfying $R_{\delta_0}(\theta) \le R_\delta(\theta)$ for every θ, and if there is at least one θ for which $R_{\delta_0}(\theta) < R_\delta(\theta)$, then naturally we strictly prefer δ_0 to δ.

Definition 8.4.5 A decision function δ is said to be *admissible* if there is no δ_0 that is strictly preferred to it.

A consequence of decision theory is that we should use only admissible decision functions. Still, there are many admissible decision functions and typically none is optimal. Furthermore, a procedure that is only admissible may be a very poor choice (see Challenge 8.4.11).

There are several routes out of this impasse for decision theory. One approach is to use *reduction principles*. By this we mean that we look for an optimal decision function in some subclass $D_0 \subset D$ that is considered appropriate. So we then look for a $\delta_0 \in D_0$ such that $R_{\delta_0}(\theta) \le R_\delta(\theta)$ for every $\theta \in \Omega$ and $\delta \in D_0$, i.e., we look for an optimal decision function in D_0. Consider the following example.

EXAMPLE 8.4.8 *Size α Tests for Hypothesis Testing*
Consider a hypothesis testing problem H_0 versus H_a. Recall that in Section 8.2, we restricted attention to those test functions φ that satisfy $E_\theta(\varphi) \le \alpha$ for every $\theta \in H_0$. Such a φ is called a size α test function for this problem. So in this case, we are restricting to the class D_0 of all decision functions δ for this problem, which correspond to size α test functions.

In Section 8.2, we showed that sometimes there is an optimal $\delta \in D_0$. For example, when H_0 and H_a are simple, the Neyman–Pearson theorem (Theorem 8.2.1) provides an optimal φ; thus, δ, defined by $\delta(s, H_a) = \varphi(s)$, is optimal. We also showed in Section 8.2, however, that in general there is no optimal size α test function φ and so there is no optimal $\delta \in D_0$. In this case, further reduction principles are necessary. ∎

Another approach to selecting a $\delta \in D$ is based on choosing one particular real-valued characteristic of the risk function of δ and ordering the decision functions based on that. There are several possibilities.

One way is to introduce a prior π into the problem and then look for the decision procedure $\delta \in D$ that has smallest prior risk

$$r_\delta = E_\pi (R_\delta(\theta)).$$

We then look for a rule that has prior risk equal to $\min_{\delta \in D} r_\delta$ (or $\inf_{\delta \in D} r_\delta$). This approach is called *Bayesian decision theory.*

Definition 8.4.6 The quantity r_δ is called the *prior risk* of δ, $\min_{\delta \in D} r_\delta$ is called the *Bayes risk*, and a rule with prior risk equal to the Bayes risk is called a *Bayes rule*.

We derived Bayes rules for several problems in Section 8.3. Interestingly, Bayesian decision theory always effectively produces an answer to a decision problem. This is a very desirable property for any theory of statistics.

Another way to order decision functions uses the maximum (or supremum) risk. So for a decision function δ, we calculate

$$\max_{\theta \in \Omega} R_\delta(\theta)$$

(or $\sup_{\theta \in \Omega} R_\delta(\theta)$) and then select a $\delta \in D$ that minimizes this quantity. Such a δ has the smallest, largest risk or the smallest, worst behavior.

Definition 8.4.7 A decision function δ_0 satisfying

$$\max_{\theta \in \Omega} R_{\delta_0}(\theta) = \min_{\delta \in D} \max_{\theta \in \Omega} R_\delta(\theta) \tag{8.1}$$

is called a *minimax decision function*.

Again, this approach will always effectively produce an answer to a decision problem (see Problem 8.4.10).

Much more can be said about decision theory than this brief introduction to the basic concepts. Many interesting, general results have been established for the decision theoretic approach to statistical inference.

Summary of Section 8.4

- The decision theoretic approach to statistical inference introduces an action space $\mathcal{A}$ and a loss function L.
- A decision function δ prescribes a probability distribution $\delta(s, \cdot)$ on $\mathcal{A}$. The statistician generates a decision in $\mathcal{A}$ using this distribution after observing s.
- The problem in decision theory is to select δ; for this, the risk function $R_\delta(\theta)$ is used. The value $R_\delta(\theta)$ is the average loss incurred when using the decision function δ, and the goal is to minimize risk.
- Typically, no optimal decision function δ exists. So, to select a δ, various reduction criteria are used to reduce the class of possible decision functions, or the decision functions are ordered using some real-valued characteristic of their risk functions, e.g., maximum risk or average risk with respect to some prior.

EXERCISES

8.4.1 Suppose we observe a sample $(x_1, \ldots, x_n)$ from a Bernoulli(θ) distribution, where θ is completely unknown, and we want to estimate θ using squared error loss. Write out all the ingredients of this decision problem. Calculate the risk function of the estimator $T(x_1, \ldots, x_n) = \bar{x}$. Graph the risk function when $n = 10$.

8.4.2 Suppose we have a sample $(x_1, \ldots, x_n)$ from a Poisson(λ) distribution, where λ is completely unknown, and we want to estimate λ using squared error loss. Write out all the ingredients of this decision problem. Consider the estimator $T(x_1, \ldots, x_n) = \bar{x}$ and calculate its risk function. Graph the risk function when $n = 25$.

8.4.3 Suppose we have a sample $(x_1, \ldots, x_n)$ from an $N(\mu, \sigma_0^2)$ distribution, where μ is unknown and σ_0^2 is known, and we want to estimate μ using squared error loss. Write out all the ingredients of this decision problem. Consider the estimator $T\ (x_1, \ldots, x_n) = \bar{x}$ and calculate its risk function. Graph the risk function when $n = 25$, $\sigma_0^2 = 2$.

8.4.4 Suppose we observe a sample $(x_1, \ldots, x_n)$ from a Bernoulli(θ) distribution, where θ is completely unknown, and we want to test the null hypothesis that $\theta = 1/2$ versus the alternative that it is not equal to this quantity, and we use 0-1 loss. Write out all the ingredients of this decision problem. Suppose we reject the null hypothesis whenever we observe $n\bar{x} \in \{0, 1, n-1, n\}$. Determine the form of the test function and its associated power function. Graph the power function when $n = 10$.

8.4.5 Consider the decision problem with sample space $S = \{1, 2, 3, 4\}$, parameter space $\Omega = \{a, b\}$, with the parameter indexing the distributions given in the following table.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$
$f_a(s)$	$1/4$	$1/4$	0	$1/2$
$f_b(s)$	$1/2$	0	$1/4$	$1/4$

Suppose that the action space $\mathcal{A} = \Omega$, with $A(\theta) = \theta$, and the loss function is given by $L(\theta, a) = 1$ when $a \neq A(\theta)$ and is equal to 0 otherwise.

(a) Calculate the risk function of the deterministic decision function given by $d(1) = d(2) = d(3) = a$ and $d(4) = b$.

(b) Is d in part (a) optimal?

COMPUTER EXERCISES

8.4.6 Suppose we have a sample $(x_1, \ldots, x_n)$ from a Poisson(λ) distribution, where λ is completely unknown, and we want to test the hypothesis that $\lambda \leq \lambda_0$ versus the alternative that $\lambda > \lambda_0$, using the 0–1 loss function. Write out all the ingredients of this decision problem. Suppose we decide to reject the null hypothesis whenever $n\bar{x} > \lfloor n\lambda_0 + 2\sqrt{n\lambda_0} \rfloor$ and randomly reject the null hypothesis with probability 1/2 when $n\bar{x} = \lfloor n\lambda_0 + 2\sqrt{n\lambda_0} \rfloor$. Determine the form of the test function and its associated power function. Graph the power function when $\lambda_0 = 1$ and $n = 5$.

8.4.7 Suppose we have a sample $(x_1, \ldots, x_n)$ from an $N(\mu, \sigma_0^2)$ distribution, where μ is unknown and σ_0^2 is known, and we want to test the null hypothesis that the mean response is μ_0 versus the alternative that the mean response is not equal to μ_0, using the 0–1 loss function. Write out all the ingredients of this decision problem. Suppose that we decide to reject whenever $\bar{x} \notin [\mu_0 - 2\sigma_0/\sqrt{n}, \mu_0 + 2\sigma_0/\sqrt{n}]$. Determine the form of the test function and its associated power function. Graph the power function when $\mu_0 = 0, \sigma_0 = 3$, and $n = 10$.

PROBLEMS

8.4.8 Prove that a decision function δ that gives a probability measure $\delta(s, \cdot)$ degenerate at $d(s)$ for each $s \in S$ is equivalent to specifying a function $d : S \to \mathcal{A}$ and conversely. For such a δ, prove that $R_\delta(\theta) = E_\theta(L(\theta, d(s)))$.

8.4.9 Suppose we have a decision problem and that each probability distribution in the model is discrete.

(a) Prove that δ is optimal in D if and only if $\delta(s, \cdot)$ is degenerate at $A(\theta)$ for each s for which $P_\theta(\{s\}) > 0$.

(b) Prove that if there exist $\theta_1, \theta_2 \in \Omega$ such that $A(\theta_1) \neq A(\theta_2)$, and $P_{\theta_1}, P_{\theta_2}$ are not concentrated on disjoint sets, then there is no optimal $\delta \in D$.

8.4.10 If decision function δ has constant risk and is admissible, then prove that δ is minimax.

CHALLENGES

8.4.11 Suppose we have a decision problem in which $\theta_0 \in \Omega$ is such that $P_{\theta_0}(C) = 0$ implies that $P_\theta(C) = 0$ for every $\theta \in \Omega$. Further assume that there is no optimal decision function (see Problem 8.4.9). Then prove that the nonrandomized decision function d given by $d(s) \equiv A(\theta_0)$ is admissible. What does this result tell you about the concept of admissibility?

DISCUSSION TOPICS

8.4.12 Comment on the following statement: A natural requirement for any theory of inference is that it produce an answer for every inference problem posed. Have we discussed any theories so far that you believe will satisfy this?

8.4.13 Decision theory produces a decision in a given problem. It says nothing about how likely it is that the decision is in error. Some statisticians argue that a valid approach to inference must include some quantification of our uncertainty concerning any statement we make about an unknown, as only then can a recipient judge the reliability of the inference. Comment on this.

8.5 Further Proofs (Advanced)

Proof of Theorem 8.1.2

We want to show that a statistic U is sufficient for a model if and only if the conditional distribution of the data s given $U = u$ is the same for every $\theta \in \Omega$.

We prove this in the discrete case so that $f_\theta(s) = P_\theta(\{s\})$. The general case requires more mathematics, and we leave that to a further course.

Let u be such that $P_\theta(U^{-1}\{u\}) > 0$ where $U^{-1}\{u\} = \{s : U(s) = u\}$, so $U^{-1}\{u\}$ is the set of values of s such that $U(s) = u$. We have

$$P_\theta(s = s_1 \mid U = u) = \frac{P_\theta(s = s_1, U = u)}{P_\theta(U = u)}. \tag{8.5.1}$$

Whenever $s_1 \notin U^{-1}\{u\}$,

$$P_\theta(s = s_1, U = u) = P_\theta(\{s_1\} \cap \{s : U(s) = u\}) = P_\theta(\phi) = 0$$

independently of θ. Therefore, $P_\theta(s = s_1 \mid U = u) = 0$ independently of θ.

So let us suppose that $s_1 \in U^{-1}\{u\}$. Then

$$P_\theta(s = s_1, U = u) = P_\theta(\{s_1\} \cap \{s : U(s) = u\}) = P_\theta(\{s_1\}) = f_\theta(s_1).$$

If U is a sufficient statistic, the factorization theorem (Theorem 6.1.1) implies $f_\theta(s) = h(s)g_\theta(U(s))$ for some h and g. Therefore, since

$$P_\theta(U = u) = \sum_{s \in U^{-1}\{u\}} f_\theta(s),$$

(8.5.1) equals

$$\frac{f_\theta(s_1)}{\sum_{s \in U^{-1}\{u\}} f_\theta\,(s)} = \frac{f_\theta(s_1)}{\sum_{s \in U^{-1}\{u\}} c(s, s_1)\, f_\theta(s_1)} = \frac{1}{\sum_{s \in U^{-1}\{u\}} c(s, s_1)}$$

where

$$\frac{f_\theta\,(s)}{f_\theta(s_1)} = \frac{h(s)}{h(s_1)} = c(s, s_1).$$

We conclude that (8.5.1) is independent of θ.

Conversely, if (8.5.1) is independent of θ, then for $s_1, s_2 \in U^{-1}\{u\}$ we have

$$P_\theta(U = u) = \frac{P_\theta(s = s_2)}{P_\theta(s = s_2 \mid U = u)}.$$

Thus

$$\begin{aligned} f_\theta(s_1) &= P_\theta(s = s_1) = P_\theta(s = s_1 \mid U = u) P_\theta(U = u) \\ &= P_\theta(s = s_1 \mid U = u) \frac{P_\theta(s = s_2)}{P_\theta(s = s_2 \mid U = u)} \\ &= \frac{P_\theta(s = s_1 \mid U = u)}{P_\theta(s = s_2 \mid U = u)} f_\theta(s_2) = c(s_1, s_2)\, f_\theta(s_2), \end{aligned}$$

where

$$c(s_1, s_2) = \frac{P_\theta(s = s_1 \mid U = u)}{P_\theta(s = s_2 \mid U = u)}.$$

By the definition of sufficiency in Section 6.1.1, this establishes the sufficiency of U. ∎

Establishing the Completeness of $\bar{x}$ in Example 8.1.3

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 $\sigma_0^2 > 0$ is known. In Example 6.1.7, we showed that $\bar{x}$ is a minimal sufficient statistic.

Suppose that the function h is such that $E_\mu(h(\bar{x})) = 0$ for every $\mu \in R^1$. Then defining

$$h^+(\bar{x}) = \max(0, h(\bar{x})) \text{ and } h^-(\bar{x}) = \max(0, -h(\bar{x})),$$

we have $h(\bar{x}) = h^+(\bar{x}) - h^-(\bar{x})$. Therefore, setting

$$c^+(\mu) = E_\mu(h^+(\bar{X})) \text{ and } c^-(\mu) = E_\mu(h^-(\bar{X})),$$

we must have

$$E_\mu(h(\bar{X})) = E_\mu(h^+(\bar{X})) - E_\mu(h^-(\bar{X})) = c^+(\mu) - c^-(\mu) = 0,$$

and so $c^+(\mu) = c^-(\mu)$. Because h^+ and h^- are nonnegative functions, we have that $c^+(\mu) \ge 0$ and $c^-(\mu) \ge 0$.

If $c^+(\mu) = 0$, then we have that $h^+(\bar{x}) = 0$ with probability 1, because a nonnegative function has mean 0 if and only if it is 0 with probability 1 (see Challenge 3.3.22). Then $h^-(\bar{x}) = 0$ with probability 1 also, and we conclude that $h(\bar{x}) = 0$ with probability 1.

If $c^+(\mu_0) > 0$, then $h^+(\bar{x}) > 0$ for all $\bar{x}$ in a set A having positive probability with respect to the $N(\mu_0, \sigma_0^2/n)$ distribution (otherwise $h^+(\bar{x}) = 0$ with probability 1, which implies, as above, that $c^+(\mu_0) = 0$). This implies that $c^+(\mu) > 0$ for every μ because every $N(\mu, \sigma_0^2/n)$ distribution assigns positive probability to A as well (you can think of A as a subinterval of R^1).

Now note that

$$g^+(\bar{x}) = h^+(\bar{x}) \frac{1}{\sqrt{2\pi}\sigma_0} \exp(-n\bar{x}^2/2\sigma_0^2)$$

is nonnegative and is strictly positive on A. We can write

$$\begin{aligned} c^+(\mu) = E_\mu(h^+(\bar{X})) &= \int_{-\infty}^{\infty} h^+(\bar{x}) \frac{1}{\sqrt{2\pi}\sigma_0} \exp(-n\,(\bar{x}-\mu)^2/2\sigma_0^2)\,d\bar{x} \\ &= \exp(-n\mu^2/2\sigma_0^2) \int_{-\infty}^{\infty} \exp(n\mu\bar{x}/\sigma_0^2) g^+(\bar{x})\,d\bar{x}. \end{aligned} \tag{8.5.2}$$

Setting $\mu = 0$ establishes that $0 < \int_{-\infty}^{\infty} g^+(\bar{x})\,d\bar{x} < \infty$, because $0 < c^+(\mu) < \infty$ for every μ. Therefore,

$$\frac{g^+(\bar{x})}{\int_{-\infty}^{\infty} g^+(\bar{x})\,d\bar{x}}$$

is a probability density of a distribution concentrated on $A^+ = \{\bar{x} : h(\bar{x}) > 0\}$. Furthermore, using (8.5.2) and the definition of moment-generating function in Section 3.4,

$$\frac{c^+(\mu)\exp(n\mu^2/2\sigma_0^2)}{\int_{-\infty}^{\infty} g^+(\bar{x})\,d\bar{x}} \tag{8.5.3}$$

is the moment-generating function of this distribution evaluated at $n\mu/\sigma_0^2$.

Similarly, we define

$$g^-(\bar{x}) = h^-(\bar{x})\frac{1}{\sqrt{2\pi}\sigma_0}\exp(-n\bar{x}^2/2\sigma_0^2)$$

so that

$$\frac{g^-(\bar{x})}{\int_{-\infty}^{\infty} g^-(\bar{x})\,d\bar{x}}$$

is a probability density of a distribution concentrated on $A^- = \{\bar{x} : h(\bar{x}) < 0\}$. Also,

$$\frac{c^-(\mu)\exp(n\mu^2/2\sigma_0^2)}{\int_{-\infty}^{\infty} g^-(\bar{x})\,d\bar{x}} \tag{8.5.4}$$

is the moment-generating function of this distribution evaluated at $n\mu/\sigma_0^2$.

Because $c^+(\mu) = c^-(\mu)$, we have that (setting $\mu = 0$)

$$\int_{-\infty}^{\infty} g^+(\bar{x})\,d\bar{x} = \int_{-\infty}^{\infty} g^-(\bar{x})\,d\bar{x}.$$

This implies that (8.5.3) equals (8.5.4) for every μ, and so the moment-generating functions of these two distributions are the same everywhere. By Theorem 3.4.6, these distributions must be the same. But this is impossible, as the distribution given by g^+ is concentrated on A^+ whereas the distribution given by g^- is concentrated on A^- and $A^+ \cap A^- = \phi$. Accordingly, we conclude that we cannot have $c^+(\mu) > 0$, and we are done.

The Proof of Theorem 8.2.1 (the Neyman–Pearson Theorem)

We want to prove that when $\Omega = \{\theta_0, \theta_1\}$, *and we want to test* $H_0 : \theta = \theta_0$, *then an exact size* α *test function* φ_0 *exists of the form*

$$\varphi_0(s) = \begin{cases} 1 & f_{\theta_1}(s)/f_{\theta_0}(s) > c_0 \\ \gamma & f_{\theta_1}(s)/f_{\theta_0}(s) = c_0 \\ 0 & f_{\theta_1}(s)/f_{\theta_0}(s) < c_0 \end{cases} \tag{8.5.5}$$

for some $\gamma \in [0, 1]$ *and* $c_0 \geq 0$, *and this test is UMP size* α.

We develop the proof of this result in the discrete case. The proof in the more general context is similar.

First, we note that $\{s : f_{\theta_0}(s) = f_{\theta_1}(s) = 0\}$ has P_θ measure equal to 0 for both $\theta = \theta_0$ and $\theta = \theta_1$. Accordingly, without loss we can remove this set from the sample space and assume hereafter that $f_{\theta_0}(s)$ and $f_{\theta_1}(s)$ cannot be simultaneously 0. Therefore, the ratio $f_{\theta_1}(s)/f_{\theta_0}(s)$ is always defined.

Suppose that $\alpha = 1$. Then setting $c = 0$ and $\gamma = 1$ in (8.5.5), we see that $\varphi_0(s) \equiv 1$, and so $E_{\theta_1}(\varphi_0) = 1$. Therefore, φ_0 is UMP size α, because no test can have power greater than 1.

Suppose that $\alpha = 0$. Setting $c_0 = \infty$ and $\gamma = 1$ in (8.5.5), we see that $\varphi_0(s) = 0$ if and only if $f_{\theta_0}(s) > 0$ (if $f_{\theta_0}(s) = 0$, then $f_{\theta_1}(s)/f_{\theta_0}(s) = \infty$ and conversely). So φ_0 is the indicator function for the set $A = \{s : f_{\theta_0}(s) = 0\}$, and therefore $E_{\theta_0}(\varphi_0) = 0$. Further, any size 0 test function φ must be 0 on A^c to have $E_{\theta_0}(\varphi) = 0$. On A we have that $0 \le \varphi(s) \le 1 = \varphi_0(s)$ and so $E_{\theta_1}(\varphi) \le E_{\theta_1}(\varphi_0)$. Therefore, φ_0 is UMP size α.

Now assume that $0 < \alpha < 1$. Consider the distribution function of the likelihood ratio when $\theta = \theta_0$, namely,

$$1 - \alpha^*(c) = P_{\theta_0}(f_{\theta_1}(s)/f_{\theta_0}(s) \le c).$$

So $1-\alpha^*(c)$ is a nondecreasing function of c with $1-\alpha^*(-\infty) = 0$ and $1-\alpha^*(\infty) = 1$.

Let c_0 be the smallest value of c such that $1 - \alpha \le 1 - \alpha^*(c)$ (recall that $1 - \alpha^*(c)$ is right continuous because it is a distribution function). Then we have that $1-\alpha^*(c_0 - 0) = 1 - \lim_{\varepsilon \searrow 0} \alpha^*(c_0 - \varepsilon) \le 1 - \alpha \le 1 - \alpha^*(c_0)$ and (using the fact that the jump in a distribution function at a point equals the probability of the point)

$$\begin{aligned} P_{\theta_0}(f_{\theta_1}(s)/f_{\theta_0}(s) = c_0) &= (1 - \alpha^*(c_0)) - (1 - \alpha^*(c_0 - 0)) \\ &= \alpha^*(c_0 - 0) - \alpha^*(c_0). \end{aligned}$$

Using this value of c_0 in (8.5.5), put

$$\gamma = \begin{cases} \frac{\alpha-\alpha^*(c_0)}{\alpha^*(c_0-0)-\alpha^*(c_0)} & \alpha^*(c_0 - 0) \ne \alpha^*(c_0) \\ \\ 0 & \text{otherwise,} \end{cases}$$

and note that $\gamma \in [0, 1]$. Then we have

$$\begin{aligned} E_{\theta_0}(\varphi_0) &= \gamma P_{\theta_0}(f_{\theta_1}(s)/f_{\theta_0}(s) = c_0) + P_{\theta_0}(f_{\theta_1}(s)/f_{\theta_0}(s) > c_0) \\ &= \alpha - \alpha^*(c_0) + \alpha^*(c_0) = \alpha, \end{aligned}$$

so φ_0 has exact size α.

Now suppose that φ is another size α test and $E_{\theta_1}(\varphi) \ge E_{\theta_1}(\varphi_0)$. We partition the sample space as $S = S_0 \cup S_1 \cup S_2$ where

$$\begin{aligned} S_0 &= \{s : \varphi_0(s) - \varphi(s) = 0\}, \\ S_1 &= \{s : \varphi_0(s) - \varphi(s) < 0\}, \\ S_2 &= \{s : \varphi_0(s) - \varphi(s) > 0\}. \end{aligned}$$

Note that

$$S_1 = \{s : \varphi_0(s) - \varphi(s) < 0,\ f_{\theta_1}(s)/f_{\theta_0}(s) \le c_0\}$$

because $f_{\theta_1}(s)/f_{\theta_0}(s) > c_0$ implies $\varphi_0(s) = 1$, which implies $\varphi_0(s) - \varphi(s) = 1 - \varphi(s) \ge 0$ as $0 \le \varphi(s) \le 1$. Also

$$S_2 = \{s : \varphi_0(s) - \varphi(s) > 0,\ f_{\theta_1}(s)/f_{\theta_0}(s) \ge c_0\}$$

because $f_{\theta_1}(s)/f_{\theta_0}(s) < c_0$ implies $\varphi_0(s) = 0$, which implies $\varphi_0(s)-\varphi(s) = -\varphi(s) \le 0$ as $0 \le \varphi(s) \le 1$.

Therefore,

$$\begin{aligned} 0 &\geq E_{\theta_1}(\varphi_0) - E_{\theta_1}(\varphi) = E_{\theta_1}(\varphi_0 - \varphi) \\ &= E_{\theta_1}(I_{S_1}(s)(\varphi_0(s) - \varphi(s))) + E_{\theta_1}(I_{S_2}(s)(\varphi_0(s) - \varphi(s))). \end{aligned}$$

Now note that

$$\begin{aligned} &E_{\theta_1}(I_{S_1}(s)(\varphi_0(s) - \varphi(s))) = \sum_{s \in S_1} (\varphi_0(s) - \varphi(s))\, f_{\theta_1}(s) \\ &\geq c_0 \sum_{s \in S_1} (\varphi_0(s) - \varphi(s))\, f_{\theta_0}(s) = c_0 E_{\theta_0}(I_{S_1}(s)(\varphi_0(s) - \varphi(s))) \end{aligned}$$

because $\varphi_0(s) - \varphi(s) < 0$ and $f_{\theta_1}(s)/f_{\theta_0}(s) \leq c_0$ when $s \in S_1$. Similarly, we have that

$$\begin{aligned} &E_{\theta_1}(I_{S_2}(s)(\varphi_0(s) - \varphi(s))) = \sum_{s \in S_2} (\varphi_0(s) - \varphi(s))\, f_{\theta_1}(s) \\ &\geq c_0 \sum_{s \in S_2} (\varphi_0(s) - \varphi(s))\, f_{\theta_0}(s) = c_0 E_{\theta_0}(I_{S_2}(s)(\varphi_0(s) - \varphi(s))) \end{aligned}$$

because $\varphi_0(s) - \varphi(s) > 0$ and $f_{\theta_1}(s)/f_{\theta_0}(s) \geq c_0$ when $s \in S_2$.

Combining these inequalities, we obtain

$$\begin{aligned} 0 &\geq E_{\theta_1}(\varphi_0) - E_{\theta_1}(\varphi) \geq c_0 E_{\theta_0}(\varphi_0 - \varphi) \\ &= c_0(E_{\theta_0}(\varphi_0) - E_{\theta_0}(\varphi)) = c_0(\alpha - E_{\theta_0}(\varphi)) \geq 0 \end{aligned}$$

because $E_{\theta_0}(\varphi) \leq 0$. Therefore, $E_{\theta_1}(\varphi_0) = E_{\theta_1}(\varphi)$, which proves that φ_0 is UMP among all size α tests. ∎