8

Introduction to Statistics

8.1 STATISTICAL DECISIONS

Suppose that the number of telephone calls made per day at a given exchange is known to have a Poisson distribution with parameter θ, but θ itself is unknown. In order to obtain some information about θ, we observe the number of calls over a certain period of time, and then try to come to a decision about θ. The nature of the decision will depend on the type of information desired. For example, it may be that extra equipment will be needed if $\theta > \theta_0$, but not if $\theta \leq \theta_0$. In this case we make one of two possible decisions: we decide either that $\theta > \theta_0$ or that $\theta \leq \theta_0$. Alternatively, we may want to estimate the actual value of θ in order to know how much equipment to install. In this case the decision results in a number $\hat{\theta}$, which we hope is as close to θ as possible. In general, an incorrect decision will result in a loss, which may be measurable in precise terms, as in the case of the cost of unnecessary equipment, but which also may have intangible components. For example, it may be difficult to assign a numerical value to losses due to customer complaints, unfavorable publicity, or government investigations.

Decision problems such as the one just discussed may be formulated mathematically by means of a *statistical decision model*. The ingredients of the model are as follows.

1. N, the set of *states of nature*.

2. A random variable (or random vector) R, the *observable*, whose distribution function F_θ depends on the particular $\theta \in N$. We may imagine that "nature" chooses the parameter $\theta \in N$ (without revealing the result to us); we then observe the value of a random variable R with distribution function F_θ. In the above example, N is the set of positive real numbers, and F_θ is the distribution function of a Poisson random variable with parameter θ.
3. A, the set of possible *actions*. In the above example, since we are trying to determine the value of θ, $A = N = (0, \infty)$.
4. A *loss function* (or *cost function*) $L(\theta, a)$, $\theta \in N$, $a \in A$; $L(\theta, a)$ represents our loss when the true state of nature is θ and we take action a.

The process by which we arrive at a decision may be described by means of a *decision function*, defined as follows.

Let E be the range of the observable R (e.g., E^1 if R is a random variable, E^n if R is an n-dimensional random vector). A *nonrandomized decision function* is a function φ from E to A. Thus, if R takes the value x, we take action $\varphi(x)$. φ is to be chosen so as to minimize the loss, in some sense.

Nonrandomized decision functions are not adequate to describe all aspects of the decision-making process. For example, under certain conditions we may flip a coin or use some other chance device to determine the appropriate action. (If you are a statistician employed by a company, it is best to do this out of sight of the customer.) The general concept of a *decision function* is that of a mapping assigning to each $x \in E$ a probability measure P_x on an appropriate sigma field of subsets of A. Thus $P_x(B)$ is the probability of taking an action in the set B when $R = x$ is observed. A nonrandomized decision function may be regarded as a decision function with each P_x concentrated on a single point; that is, for each x we have $P_x\{a\} = 1$ for some a $(= \varphi(x))$ in A.

We shall concentrate on the two most important special cases of the statistical decision problem, hypothesis testing and estimation.

A typical physical situation in which decisions of this type occur is the problem of signal detection. The input to a radar receiver at a particular instant of time may be regarded as a random variable R with density f_θ, where θ is related to the signal strength. In the simplest model, $R = \theta + R'$, where R' (the noise) is a random variable with a specified density, and θ is a fixed but unknown constant determined by the strength of the signal. We may be interested in distinguishing between two conditions: the absence of a target $(\theta = \theta_0)$ versus its presence $(\theta = \theta_1)$; this is an example of a hypothesis-testing problem. Alternatively, we may know that a signal is present and wish to estimate its strength. Thus, after observing R, we record a number that we hope is close to the true value of θ; this is an example of a problem in estimation.

As another example, suppose that θ is the (unknown) percentage of defective components produced on an assembly line. We inspect n components (i.e., we observe $R_1, \ldots, R_n$, where $R_i = 1$ if component i is defective, $R_i = 0$ if component i is accceptable) and then try to say something about θ. We may be trying to distinguish between the two conditions $\theta \leq \theta_0$ and $\theta > \theta_0$ (hypothesis testing), or we may be trying to come as close as possible to the true value of θ (estimation).

In working the specific examples in the chapter, the table of common density and probability functions and their properties given at the end of the book may be helpful.

8.2 HYPOTHESIS TESTING

Consider again the statistical decision model of the preceding section. Suppose that H_0 and H_1 are disjoint nonempty subsets of N whose union is N, and our objective is to determine whether the true state of nature θ belongs to H_0 or to H_1. (In the example on the telephone exchange, H_0 might correspond to $\theta \leq \theta_0$, and H_1 to $\theta > \theta_0$.) Thus our ultimate decision must be either "$\theta \in H_0$" or "$\theta \in H_1$," so that the action space A contains only two points, labeled 0 and 1 for convenience.

The above decision problem is called a *hypothesis-testing problem*; H_0 is called the *null hypothesis*, and H_1 the *alternative*. H_0 is said to be *simple* iff it contains only one element; otherwise H_0 is said to be *composite*, and similarly for H_1. To take action 1 is to *reject the null hypothesis* H_0; to take action 0 is to *accept* H_0.

We first consider the case of *simple hypothesis versus simple alternative*. Here H_0 and H_1 each contain one element, say θ_0 and θ_1. For the sake of definiteness, we assume that under H_0, R is absolutely continuous with density f_0, and under H_1, R is absolutely continuous with density f_1. (The results of this section will also apply to the discrete case upon replacing integrals by sums.) Thus the problem essentially comes down to deciding, after observing R, whether R has density f_0 or f_1.

A decision function may be specified by giving a (Borel measurable) function φ from E to $[0, 1]$, with $\varphi(x)$ interpreted as the probability of rejecting H_0 when x is observed. Thus, if $\varphi(x) = 1$, we reject H_0; if $\varphi(x) = 0$, we accept H_0; and if $\varphi(x) = a$, $0 < a < 1$, we toss a coin with probability a of heads: if the coin comes up heads, we reject H_0; if tails, we accept H_0. The set $\{x: \varphi(x) = 1\}$ is called the *rejection region* or the *critical region*; the function φ is called a *test*. The decision we arrive at may be in error in two possible ways. A *type 1 error* occurs if we reject H_0 when it is in fact true, and a *type 2 error* occurs if H_0 is accepted when it is false, that is, when H_1 is

true. Now if H_0 is true and we observe $R = x$, an error will be made if H_0 is rejected; this happens with probability $\varphi(x)$. Thus the probability of a type 1 error is

$$\alpha = \int_{-\infty}^{\infty} \varphi(x) f_0(x)\, dx \tag{8.2.1}$$

Similarly, the probability of a type 2 error is

$$\beta = \int_{-\infty}^{\infty} (1 - \varphi(x)) f_1(x)\, dx \tag{8.2.2}$$

Note that α is the expectation of $\varphi(R)$ under H_0, sometimes written $E_{\theta_0}\varphi$; similarly, $\beta = 1 - E_\theta\, \varphi$.

It would be desirable to choose φ so that both α and β will be small, but, as we shall see, a decrease in one of the two error probabilities usually results in an increase in the other. For example, if we ignore the observed data and always accept H_0, then $\alpha = 0$ but $\beta = 1$.

There is no unique answer to the question of what is a good test; we shall consider several possibilities First, suppose that there is a nonnegative cost c_i associated with a type i error, $i = 1, 2$. (For simplicity, assume that the cost of a correct decision is 0.) Suppose also that we know the probability p that the null hypothesis will be true. (p is called the *a priori probability* of H_0. In many situations it will be difficult to estimate; for example, in a radar reception problem, H_0 might correspond to no signal being present.)

Let φ be a test with error probabilities $\alpha(\varphi)$ and $\beta(\varphi)$ The over-all average cost associated with φ is

$$B(\varphi) = pc_1\alpha(\varphi) + (1 - p)c_2\beta(\varphi) \tag{8.2.3}$$

$B(\varphi)$ is called the *Bayes risk* associated with φ; a test that minimizes $B(\varphi)$ is called a *Bayes test* corresponding to the given p, c_1, c_2, f_0, and f_1.

The Bayes solution can be computed in a straightforward way. We have, from (8.2.1–8.2.3),

$$\begin{aligned} B(\varphi) &= \int_{-\infty}^{\infty} [pc_1\varphi(x)f_0(x) + (1 - p)c_2(1 - \varphi(x))f_1(x)]\, dx \\ &= \int_{-\infty}^{\infty} \varphi(x)[pc_1f_0(x) - (1 - p)c_2f_1(x)]\, dx + (1 - p)c_2 \end{aligned} \tag{8.2.4}$$

Now if we wish to minimize $\int_S \varphi(x)g(x)\, dx$ and $g(x) < 0$ on S, we can do no better than to take $\varphi(x) = 1$ for all x in S; if $g(x) > 0$ on S, we should take $\varphi(x) = 0$ for all x in S; if $g(x) = 0$ on S, $\varphi(x)$ may be chosen arbitrarily. In this case $g(x) = pc_1f_0(x) - (1 - p)c_2f_1(x)$, and the Bayes solution may

therefore be given as follows.

Let $L(x) = f_1(x)/f_0(x)$.

If $L(x) > pc_1/(1-p)c_2$, take $\varphi(x) = 1$; that is, reject H_0.

If $L(x) < pc_1/(1-p)c_2$, take $\varphi(x) = 0$; that is, accept H_0.

If $L(x) = pc_1/(1-p)c_2$, take $\varphi(x) =$ anything.

L is called the *likelihood ratio*, and a test φ such that for some constant λ, $0 \leq \lambda \leq \infty$, $\varphi(x) = 1$ when $L(x) > \lambda$ and $\varphi(x) = 0$ when $L(x) < \lambda$, is called a *likelihood ratio test*, abbreviated LRT.

To avoid ambiguity, if $f_1(x) > 0$ and $f_0(x) = 0$, we take $L(x) = \infty$. The set on which $f_1(x) = f_0(x) = 0$ may be ignored, since it will have probability 0 under both H_0 and H_1. Also, if we observe an x for which $f_1(x) > 0$ and $f_0(x) = 0$, it must be associated with H_1, so that we should take $\varphi(x) = 1$. It will be convenient to build this requirement into the definition of a likelihood ratio test: *if $L(x) = \infty$ we assume that $\varphi(x) = 1$.*

In fact, likelihood ratio tests are completely adequate to describe the problem of testing a simple hypothesis versus a simple alternative. This assertion will be justified by the sequence of theorems to follow.

From now on, the notation $P_\theta(B)$ will indicate the probability that the value of R will belong to the set B when the true state of nature is θ.

Theorem 1. *For any α, $0 \leq \alpha \leq 1$, there is a likelihood ratio test whose probability of type 1 error is α.*

PROOF. If $\alpha = 0$, the test given by $\varphi(x) = 1$ if $L(x) = \infty$; $\varphi(x) = 0$ if $L(x) < \infty$, is the desired LRT, so assume $\alpha > 0$. Now $G(y) = P_{\theta_0}\{x: L(x) \leq y\}$, $-\infty < y < \infty$, is a distribution function [of the random variable $L(R)$; notice that $L(R) \geq 0$, and $L(R)$ cannot be infinite under H_0]. Thus either we can find λ, $0 \leq \lambda < \infty$, such that $G(\lambda) = 1 - \alpha$, or else G jumps through $1 - \alpha$; that is, for some λ we have $G(\lambda^-) \leq 1 - \alpha \leq G(\lambda)$ (see Figure 8.2.1). Define

$$\begin{aligned} \varphi(x) &= 1 \quad \text{if } L(x) > \lambda \\ &= 0 \quad \text{if } L(x) < \lambda \\ &= a \quad \text{if } L(x) = \lambda \end{aligned}$$

where $a = [G(\lambda) - (1-\alpha)]/[G(\lambda) - G(\lambda^-)]$ if $G(\lambda) > G(\lambda^-)$, $a =$ an arbitrary number in $[0, 1]$ if $G(\lambda) = G(\lambda^-)$. Then the probability of a type 1 error is

$$P_{\theta_0}\{x: L(x) > \lambda\} + aP_{\theta_0}\{x: L(x) = \lambda\} = 1 - G(\lambda) + a[G(\lambda) - G(\lambda^-)] = \alpha$$

as desired.

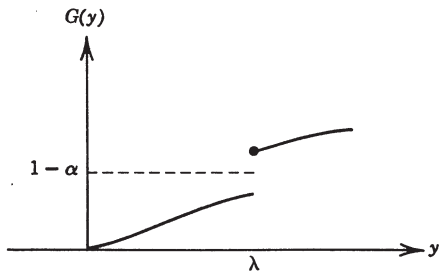

FIGURE 8.2.1

A test is said to be *at level* α_0 if its probability α of type 1 error is $\leq \alpha_0$. α itself is called the *size* of the test, and $1 - \beta$, the probability of rejecting the null hypothesis when it is false, is called the *power* of the test.

The following result, known as the *Neyman-Pearson lemma*, is the fundamental theorem of hypothesis testing.

Theorem 2. *Let φ_λ be α LRT with parameter λ and error probabilities α_λ and β_λ. Let φ be an arbitrary test with error probabilities α and β; if $\alpha \leq \alpha_\lambda$ then $\beta \geq \beta_\lambda$. In other words, the LRT has maximum power among all tests at level α_λ.*

We give two proofs

FIRST PROOF. Consider the Bayes problem with costs $c_1 = c_2 = 1$, and set $\lambda = pc_1/(1 - p)c_2 = p/(1 - p)$. Assuming first that $\lambda < \infty$, we have $p = \lambda/(1 + \lambda)$. Thus φ_λ is the Bayes solution when the a priori probability is $p = \lambda/(1 + \lambda)$.

If $\beta < \beta_\lambda$, we compute the Bayes risk [see (8.2.3)] for $p = \lambda/(1 + \lambda)$, using the test φ.

$$B(\varphi) = p\alpha + (1 - p)\beta$$

But $\alpha \leq \alpha_\lambda$ by hypothesis, while $\beta < \beta_\lambda$ and $p < 1$ by assumption. Thus $B(\varphi) < B(\varphi_\lambda)$, contradicting the fact that φ_λ is the Bayes solution.

It remains to consider the case $\lambda = \infty$. Then we must have $\varphi_\lambda(x) = 1$ if $L(x) = \infty$, $\varphi_\lambda(x) = 0$ if $L(x) < \infty$. Then $\alpha_\lambda = 0$, since $L(R)$ is never infinite under H_0; consequently $\alpha = 0$, so that, by (8.2.1), $\varphi(x)f_0(x) \equiv 0$ [strictly speaking, $\varphi(x)f_0(x) = 0$ except possibly on a set of Lebesgue measure 0]. By (8.2.2),

$$\beta = \int_{\{x:L(x)<\infty\}} (1 - \varphi(x))f_1(x)\,dx + \int_{\{x:L(x)=\infty\}} (1 - \varphi(x))f_1(x)\,dx$$

If $L(x) < \infty$, then $f_0(x) > 0$; hence $\varphi(x) = 0$. Thus, in order to minimize β, we must take $\varphi(x) = 1$ when $L(x) = \infty$. But this says that $\beta \geq \beta_\lambda$, completing the proof.

SECOND PROOF. First assume $\lambda < \infty$. We claim that $[\varphi_\lambda(x) - \varphi(x)] \times [f_1(x) - \lambda f_0(x)] \geq 0$ for all x. For if $f_1(x) > \lambda f_0(x)$, then $\varphi_\lambda(x) = 1 \geq \varphi(x)$, and if $f_1(x) < \lambda f_0(x)$, then $\varphi_\lambda(x) = 0 \leq \varphi(x)$. Thus

$$\int_{-\infty}^{\infty} [\varphi_\lambda(x) - \varphi(x)][f_1(x) - \lambda f_0(x)] \geq 0$$

By (8.2.1) and (8.2.2),

$$1 - \beta_\lambda - (1 - \beta) - \lambda\alpha_\lambda + \lambda\alpha \geq 0$$

or

$$\beta - \beta_\lambda \geq \lambda(\alpha_\lambda - \alpha) \geq 0$$

The case $\lambda = \infty$ is handled just as in the first proof.

If we wish to construct a test that is best at level α in the sense of maximum power, we find, by Theorem 1, a LRT of size α. By Theorem 2, the test has maximum power among all tests at level α. We shall illustrate the procedure with examples and problems later in the section.

Finally, we show that no matter what criterion the statistician adopts in defining a good test, he can restrict himself to the class of likelihood ratio tests.

A test φ with error probabilities α and β is said to be *inadmissible* iff there is a test φ' with error probabilities α' and β', with $\alpha' \leq \alpha$, $\beta' \leq \beta$, and either $\alpha' < \alpha$ or $\beta' < \beta$. (In this case we say that φ' is *better than* φ.) Of course, φ is *admissible* iff it is not inadmissible.

Theorem 3. *Every LRT is admissible.*

PROOF. Let φ_λ be a LRT with parameter λ and error probabilities α_λ and β_λ, and φ an arbitrary test with error probabilities α and β. We have seen that if $\alpha \leq \alpha_\lambda$, then $\beta \geq \beta_\lambda$. But *the Neyman-Pearson lemma is symmetric in* H_0 *and* H_1. In other words, if we relabel H_1 as the null hypothesis and H_0 as the alternative, Theorem 2 states that if $\beta \leq \beta_\lambda$, then $\alpha \geq \alpha_\lambda$; the result follows.

Thus no test can be better than a LRT. In fact, if φ is any test, then there is a LRT φ_λ that is *as good as* φ; that is, $\alpha_\lambda \leq \alpha$ and $\beta_\lambda \leq \beta$. For by Theorem 1 there is a LRT φ_λ with $\alpha_\lambda = \alpha$, and by Theorem 2 $\beta_\lambda \leq \beta$. This argument establishes the following result, essentially a converse to Theorem 3.

Theorem 4. *If φ is an admissible test, there is a LRT with exactly the same error probabilities.*

PROOF. As above, we find a LRT φ_λ with $\alpha_\lambda = \alpha$ and $\beta_\lambda \leq \beta$; since φ is admissible, we must have $\beta_\lambda = \beta$.

▶ **Example 1.** Suppose that under H_0, R is uniformly distributed between 0 and 1, and under H_1, R has density $3x^2$, $0 \leq x \leq 1$. For short we write

$$H_0: f_0(x) = 1, \qquad 0 \leq x \leq 1$$
$$H_1: f_1(x) = 3x^2, \qquad 0 \leq x \leq 1$$

We are going to find the *risk set* S, that is, the set of points $(\alpha(\varphi), \beta(\varphi))$ where φ ranges over all possible tests. [The individual points $(\alpha(\varphi), \beta(\varphi))$ are called *risk points*.] We are also going to find the set S_A of *admissible risk points*, that is, the set of risk points corresponding to admissible tests. By Theorems 3 and 4, S_A is the set of risk points corresponding to LRTs.

First we notice two general properties of S.

1. *S is convex*; that is, if Q_1 and Q_2 belong to S, so do all points on the line segment joining Q_1 to Q_2. In other words, $(1 - a)Q_1 + aQ_2 \in S$ for all $a \in [0, 1]$.

For if $Q_1 = (\alpha(\varphi_1), \beta(\varphi_1))$, $Q_2 = (\alpha(\varphi_2), \beta(\varphi_2))$ and $0 \leq a \leq 1$, let $\varphi = (1 - a)\varphi_1 + a\varphi_2$. Then φ is a test, and by (8.2.1) and (8.2.2), $\alpha(\varphi) = (1 - a)\alpha(\varphi_1) + a\alpha(\varphi_2)$, $\beta(\varphi) = (1 - a)\beta(\varphi_1) + a\beta(\varphi_2)$. If $Q = (\alpha(\varphi), \beta(\varphi))$, then $Q \in S$, since φ is a test, and $Q = (1 - a)Q_1 + aQ_2$.

2. *S is symmetric about (1/2, 1/2)*; that is, if $|\varepsilon|, |\delta| \leq 1/2$ and $(1/2 - \varepsilon, 1/2 - \delta) \in S$, then $(1/2 + \varepsilon, 1/2 + \delta) \in S$. Equivalently, $(\alpha, \beta) \in S$ implies $(1 - \alpha, 1 - \beta) \in S$.

For if $(\alpha(\varphi), \beta(\varphi)) \in S$, let $\varphi' = 1 - \varphi$; then φ' is a test and $\alpha(\varphi') = 1 - \alpha(\varphi)$, $\beta(\varphi') = 1 - \beta(\varphi)$.

To return to the present example, we have $L(x) = 3x^2$, $0 \leq x \leq 1$. Thus the error probabilities for a LRT with parameter $\lambda \leq 3$ are

$$\alpha = P_{\theta_0}\{x: L(x) > \lambda\} = P_{\theta_0}\left\{x: x > \left(\frac{\lambda}{3}\right)^{1/2}\right\} = 1 - \left(\frac{\lambda}{3}\right)^{1/2}$$

$$\beta = P_{\theta_1}\{x: L(x) < \lambda\} = P_{\theta_1}\left\{x: x < \left(\frac{\lambda}{3}\right)^{1/2}\right\}$$
$$= \int_0^{(\lambda/3)^{1/2}} 3x^2\,dx = \left(\frac{\lambda}{3}\right)^{3/2} = (1 - \alpha)^3$$

(If $\lambda > 3$, then $\alpha = 0$, $\beta = 1$.) Thus $S_A = \{(\alpha, (1 - \alpha)^3): 0 \leq \alpha \leq 1\}$. Since no test can be better than a LRT, S_A is the lower boundary of the

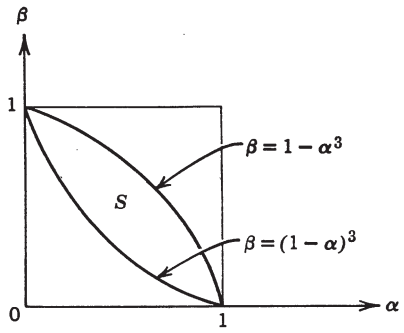

FIGURE 8.2.2

set S; hence, by symmetry, $\{(1-\alpha, 1-(1-\alpha)^3), 0 \le \alpha \le 1\} = \{(\alpha, 1-\alpha^3): 0 \le \alpha \le 1\}$ is the upper boundary of S. Thus S must be $\{(\alpha, \beta): 0 \le \alpha \le 1, (1-\alpha)^3 \le \beta \le 1-\alpha^3\}$ (see Figure 8.2.2).

Various tests may now be computed without difficulty. We give some typical illustrations.

(a) Find a most powerful test at level .15. Set $\alpha = .15 = 1 - (\lambda/3)^{1/2}$. Since $L(x) > \lambda$ iff $x > (\lambda/3)^{1/2}$, the test is given by

$$\begin{aligned} \varphi(x) &= 1 \quad \text{if } x > .85 \\ &= 0 \quad \text{if } x < .85 \\ &= \text{anything} \quad \text{if } x = .85 \end{aligned}$$

We have $\beta = (1-\alpha)^3 = (.85)^3 = .614$.

(b) Find a Bayes test corresponding to $c_1 = 3/2$, $c_2 = 3$, $p = 3/4$. This is a LRT with $\lambda = pc_1/(1-p)c_2 = 3/2$; that is,

$$\begin{aligned} \varphi(x) &= 1 \quad \text{if } x > \left(\frac{\lambda}{3}\right)^{1/2} = \frac{\sqrt{2}}{2} = .707 \\ &= 0 \quad \text{if } x < \frac{\sqrt{2}}{2} \\ &= \text{anything} \quad \text{if } x = \frac{\sqrt{2}}{2} \end{aligned}$$

Thus $\alpha = 1 - (\lambda/3)^{1/2} = .293$, $\beta = (1-\alpha)^3$, and the Bayes risk may be computed using (8.2.3).

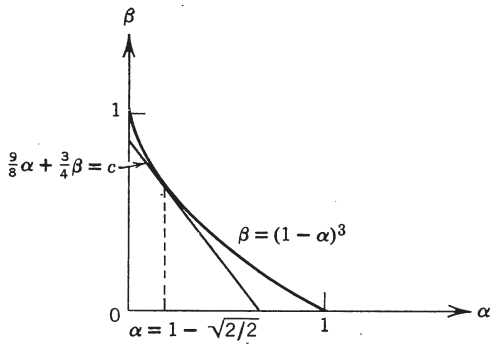

FIGURE 8.2.3 Geometric Interpretation of Bayes Solution.

The Bayes solution may be interpreted geometrically as follows. We are trying to find a test that minimizes the Bayes risk $pc_1\alpha + (1-p)c_2\beta = (9/8)\alpha + (3/4)\beta$. If we vary c until the line $(9/8)\alpha + (3/4)\beta = c$ intersects S_A, we find the desired test (see Figure 8.2.3).

Notice also that to find the Bayes solution we may differentiate $(9/8)\alpha + (3/4)(1-\alpha)^3$ and set the result equal to zero to obtain $\alpha = 1 - \sqrt{2}/2$, as before.

(c) Find a *minimax test*, that is, a test that minimizes max (α, β). It is immediate from the definition of admissibility that *an admissible test with constant risk* (*i.e.*, $\alpha = \beta$) *is minimax*. Thus we set $\alpha = \beta = (1-\alpha)^3$, which yields $\alpha = .318$ (approximately). Therefore $(\lambda/3)^{1/2} = 1 - \alpha = .682$, and so we reject H_0 if $x > .682$ and accept H_0 if $x < .682$. ◀

▶ **Example 2.** Let R be a discrete random variable taking on only the values 0, 1, 2, 3. Let the probability function of R under H_i be p_i, $i = 0, 1$, where the p_i are as follows.

x	0	1	2	3
$p_0(x)$	.1	.2	.3	.4
$p_1(x)$	.2	.1	.4	.3

The appropriate likelihood ratio here is $L(x) = p_1(x)/p_0(x)$. Arranging the values of $L(x)$ in increasing order, we have the following table.

x	1	3	2	0
$L(x)$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{4}{3}$	2

We may therefore describe the LRT with parameter λ as follows.

LRT	*Rejection Region*	*Acceptance Region*	α	β
$0 \le \lambda < \frac{1}{2}$	All x	Empty	1	0
$1/2 < \lambda < \frac{3}{4}$	$x = 0, 2, 3$	$x = 1$	.8	.1
$3/4 < \lambda < \frac{4}{3}$	$x = 0, 2$	$x = 1, 3$	.4	.4
$4/3 < \lambda < 2$	$x = 0$	$x = 1, 2, 3$	.1	.8
$2 < \lambda \le \infty$	Empty	All x	0	1

Now assume $\lambda = 3/4$. Then we reject H_0 if $x = 0$ or 2, accept H_0 if $x = 1$, and if $x = 3$ we *randomize*, that is, reject H_0 with probability a, $0 \le a \le 1$. Thus

$$\alpha = p_0(0) + p_0(2) + ap_0(3) = .4 + .4a$$
$$\beta = p_1(1) + (1 - a)p_1(3) = .1 + .3(1 - a)$$

As a ranges over $[0, 1]$, (α, β) traces out the line segment joining $(.4, .4)$ to $(.8, .1)$. In a similar fashion we calculate the error probabilities for $\lambda = 1/2$, 4/3, and 2. The admissible risk points are shown in Figure 8.2.4.

We compute several tests.

(a) Find a most powerful test at level .25. Since $.1 < .25 < .4$, we have $\lambda = 4/3$. Thus we reject H_0 if $x = 0$, accept H_0 if $x = 1$ or 3, and reject H_0 with probability a if $x = 2$, where $.1(1 - a) + .4a = .25$, so that $a = 1/2$. Notice that $\beta = .8(1 - a) + .4a = .6$.

(b) Find a Bayes test with $c_1 = c_2 = 1, p = .6$. We have $\lambda = pc_1/(1 - p)c_2 = 3/2$. Thus we reject H_0 if $x = 0$ and accept H_0 otherwise. The error probabilities are $\alpha = .1$, $\beta = .8$, and the Bayes risk is $pc_1\alpha + (1 - p)c_2\beta = .38$.

(c) Find a minimax test. The only admissible test with $\alpha = \beta$ has $\alpha = \beta = .4$, so that $3/4 < \lambda < 4/3$. We reject H_0 when $x = 0$ or 2 and accept H_0 if $x = 1$ or 3. ◀

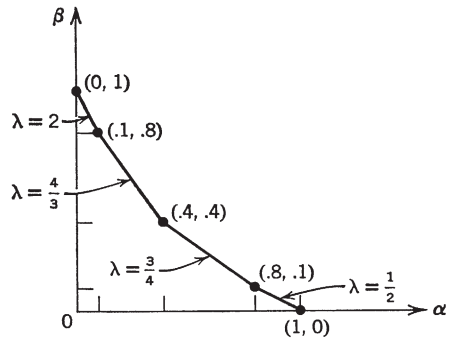

FIGURE 8.2.4 Admissible Risk Points When R is Discrete.

▶ **Example 3.** Let R be normally distributed with mean θ and variance σ^2, where σ^2 is known. We wish to test the null hypothesis that $\theta = \theta_0$ against the alternative that $\theta = \theta_1$, and the test is to be based on n independent observations $R_1, \ldots, R_n$ of R. (Assume $\theta_0 < \theta_1$.)

The appropriate likelihood ratio is

$$L(x_1, \ldots, x_n) = \frac{f_1(x_1, \ldots, x_n)}{f_0(x_1, \ldots, x_n)} = \frac{(2\pi\sigma^2)^{-n/2} \exp\left[-\sum_{k=1}^{n}(x_k - \theta_1)^2/2\sigma^2\right]}{(2\pi\sigma^2)^{-n/2} \exp\left[-\sum_{k=1}^{n}(x_k - \theta_0)^2/2\sigma^2\right]}$$

The condition $L(x) > \lambda$ is equivalent to $\ln L(x) > \ln \lambda$; that is,

$$\sum_{k=1}^{n} 2(\theta_1 - \theta_0)x_k + n(\theta_0^2 - \theta_1^2) > 2\sigma^2 \ln \lambda \tag{8.2.5}$$

This is of the form $\sum_{k=1}^{n} x_k > c$. Thus a LRT must be of the form

$$\begin{aligned} \varphi(x_1, \ldots, x_n) &= 1 \quad \text{if } \sum_{k=1}^{n} x_k > c \\ &= 0 \quad \text{if } \sum_{k=1}^{n} x_k < c \\ &= \text{anything} \quad \text{if } \sum_{k=1}^{n} x_k = c \end{aligned}$$

Now $R_1 + \cdots + R_n$ is normal with mean $n\theta$ and variance $n\sigma^2$, so that the error probabilities are

$$\begin{aligned} \alpha &= P_{\theta_0}\left\{(x_1, \ldots, x_n): \sum_{k=1}^{n} x_k > c\right\} \\ &= P_{\theta_0}\{R_1 + \cdots + R_n > c\} \\ &= P_{\theta_0}\left\{\frac{R_1 + \cdots + R_n - n\theta_0}{\sqrt{n}\,\sigma} > \frac{c - n\theta_0}{\sqrt{n}\,\sigma}\right\} \\ &= 1 - F^*\left(\frac{c - n\theta_0}{\sqrt{n}\,\sigma}\right) \quad \text{where } F^* \text{ is the normal } (0, 1) \text{ distribution function} \end{aligned}$$

$$\begin{aligned} \beta &= P_{\theta_1}\left\{(x_1, \ldots, x_n): \sum_{k=1}^{n} x_k < c\right\} \\ &= P_{\theta_1}\{R_1 + \cdots + R_n < c\} \\ &= F^*\left(\frac{c - n\theta_1}{\sqrt{n}\,\sigma}\right) \end{aligned}$$

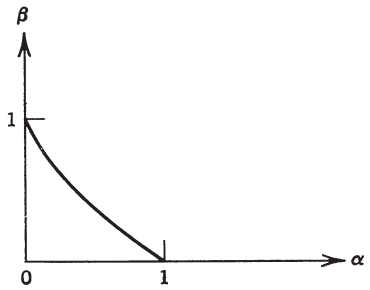

FIGURE 8.2.5 Admissible Risk Points When R is Normal.

Thus we have parametric equations for α and β with c as parameter, $-\infty \leq c \leq \infty$. The admissible risk points are sketched in Figure 8.2.5.

Suppose that we want a LRT of size α. If N_α is the number such that $1 - F^*(N_\alpha) = \alpha$, then $(c - n\theta_0)/\sqrt{n}\,\sigma = N_\alpha$, so that $c = n\theta_0 + \sqrt{n}\,\sigma N_\alpha$.

We now apply the results to a problem in testing a simple hypothesis versus a composite alternative. Again let R be normal (θ, σ^2), and take $H_0: \theta = \theta_0$, $H_1: \theta > \theta_0$.

If we choose any particular $\theta_1 > \theta_0$ and test $\theta = \theta_0$ against $\theta = \theta_1$, the test described above is most powerful at level α. However, the test is completely specified by c, and c does not depend on θ_1. Thus, for any $\theta_1 > \theta_0$, the test has the highest power of any test at level α of $\theta = \theta_0$ versus $\theta = \theta_1$. Such a test is called a *uniformly most powerful* (UMP) level α test of $\theta = \theta_0$ versus $\theta > \theta_0$.

We expect intuitively that the larger the separation between θ_0 and θ_1, the better the performance of the test in distinguishing between the two possibilities. This may be verified by considering the *power function* Q, defined by

$$
\begin{aligned}
Q(\theta) &= E_\theta \varphi \\
&= \text{the probability of rejecting } H_0 \text{ when the true state of nature is } \theta \\
&= P_\theta\{R_1 + \cdots + R_n > c\} \\
&= 1 - F^*\left(\frac{c - n\theta}{\sqrt{n}\,c}\right)
\end{aligned}
$$

Thus $Q(\theta)$ increases with θ.

Now if $H_0: \theta = \theta_0$, $H_1: \theta = \theta_1$, where $\theta_1 < \theta_0$, the same technique as

above shows that a size α LRT is of the form

$$\begin{aligned}\varphi(x_1, \ldots, x_n) &= 1 \qquad \text{if } \sum_{k=1}^{n} x_k < c \\ &= 0 \qquad \text{if } \sum_{k=1}^{n} x_k > c \\ &= \text{anything} \qquad \text{if } \sum_{k=1}^{n} x_k = c\end{aligned}$$

where

$$\alpha = F^*\left(\frac{c - n\theta_0}{\sqrt{n}\,\sigma}\right)$$

$$\beta = 1 - F^*\left(\frac{c - n\theta_1}{\sqrt{n}\,\sigma}\right)$$

$$c = n\theta_0 + \sqrt{n}\,\sigma N_{1-\alpha}$$

Again, the test is UMP at level α for $\theta = \theta_0$ versus $\theta < \theta_0$, with power function

$$Q'(\theta) = F^*\left(\frac{c - n\theta}{\sqrt{n}\,\sigma}\right)$$

which increases as θ decreases (see Figure 8.2.6).

The above discussion suggests that there can be no UMP level α test of $\theta = \theta_0$ versus $\theta \neq \theta_0$. For any such test φ must have power function $Q(\theta)$ for $\theta > \theta_0$, and $Q'(\theta)$ for $\theta < \theta_0$. But the power function of φ is given by

$$E_\theta \varphi = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \varphi(x_1, \ldots, x_n) f_\theta(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

where f_θ is the joint density of n independent normal random variables with mean θ and variance σ^2. It can be shown that this is differentiable for all θ (the derivative can be taken under the integral sign). But a function that is $Q(\theta)$ for $\theta > \theta_0$ and $Q'(\theta)$ for $\theta < \theta_0$ cannot be differentiable at θ_0.

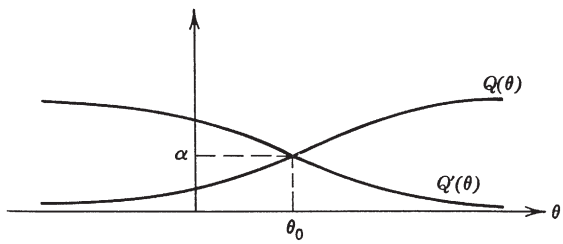

FIGURE 8.2.6 Power Functions.

In fact, the test φ with power function $Q(\theta)$ is UMP at level α for the composite hypothesis $H_0: \theta \leq \theta_0$ versus the composite alternative $H_1: \theta > \theta_0$. Let us explain what this means.

φ is said to be at level α for H_0 versus H_1 iff $E_\theta\varphi \leq \alpha$ for all $\theta \leq \theta_0$; φ is UMP at level α if for any test φ' at level α for H_0 versus H_1 we have $E_\theta\varphi' \leq E_\theta\varphi$ for all $\theta > \theta_0$.

In the present case $E_\theta\varphi = Q(\theta) \leq \alpha$ for $\theta \leq \theta_0$ by monotonicity of $Q(\theta)$, and $E_\theta\varphi' \leq E_\theta\varphi$ for $\theta > \theta_0$, since φ is UMP at level α for $\theta = \theta_0$ versus $\theta > \theta_0$.

The underlying reason for the existence of uniformly most powerful tests is the following. If $\theta < \theta'$, the likelihood ratio $f_{\theta'}(x)/f_\theta(x)$ can be expressed as a nondecreasing function of $t(x)$ [where, in this case, $t(x) = x_1 + \cdots + x_n$; see (8.2.5)]. Whenever this happens, the family of densities f_θ is said to have the *monotone likelihood ratio* (MLR) property.

Suppose that the f_θ have the MLR property. Consider the following test of $\theta = \theta_0$ versus $\theta = \theta_1$, $\theta_1 > \theta_0$.

$$\begin{aligned} \varphi(x) &= 1 \quad \text{if } t(x) > c \\ &= 0 \quad \text{if } t(x) < c \\ &= a \quad \text{if } t(x) = c \end{aligned}$$

where $P_{\theta_0}\{x: t(x) > c\} + aP_{\theta_0}\{x: t(x) = c\} = \alpha$ (notice that c does not depend on θ_1). Let λ be the value of the likelihood ratio when $t(x) = c$; then $L(x) > \lambda$ implies $t(x) > c$; hence $\varphi(x) = 1$. Also $L(x) < \lambda$ implies $t(x) < c$, so that $\varphi(x) = 0$. Thus φ is a LRT and hence is most powerful at level α. We may make the following observations.

1. φ is UMP at level α for $\theta = \theta_0$ versus $\theta > \theta_0$.

This is immediate from the Neyman-Pearson lemma and the fact that c does not depend on the particular $\theta > \theta_0$.

2. If $\theta_1 < \theta_2$, φ is the most powerful test at level $\alpha_1 = E_{\theta_1}\varphi$ for $\theta = \theta_1$ versus $\theta = \theta_2$.

Since φ is a LRT, the Neyman-Pearson lemma yields this result immediately.

3. If $\theta_1 < \theta_2$, then $E_{\theta_1}\varphi \leq E_{\theta_2}\varphi$; that is, φ has a monotone nondecreasing power function. It follows, as in the earlier discussion, that φ is UMP at level α for $\theta \leq \theta_0$ versus $\theta > \theta_0$.

By property 2, φ is most powerful at level $\alpha_1 = E_{\theta_1}\varphi$ for $\theta = \theta_1$ versus $\theta = \theta_2$. But the test $\varphi'(x) \equiv \alpha_1$ is also at level α_1; hence $E_{\theta_2}\varphi' \leq E_{\theta_2}\varphi$, that is, $\alpha_1 = E_{\theta_1}\varphi \leq E_{\theta_2}\varphi$.

REMARK. Since the Neyman-Pearson lemma is symmetric in H_0 and H_1, if $\theta_1 < \theta_2$, then for all tests φ' with $\beta(\varphi') \leq \beta(\varphi)$, we have $E_{\theta_1}\varphi \leq E_{\theta_1}\varphi'$.

We might say that φ is *uniformly least powerful* for $\theta < \theta_0$ among all tests whose type 2 error is $\leq \beta$ whenever $\theta \geq \theta_0$. ◀

PROBLEMS

1. Let $H_0: f_0(x) = e^{-x}$, $x \geq 0$; $H_1: f_1(x) = 2e^{-2x}$, $x \geq 0$.
(a) Find the risk set and the admissible risk points.
(b) Find a most powerful test at level .05.
(c) Find a minimax test.

2. Show that the following families have the MLR property, and thus UMP tests may be constructed as in the discussion of Example 3.
(a) p_θ = the joint probability function of n independent random variables, each Poisson with parameter θ.
(b) p_θ = the joint probability function of n independent random variables R_i, where R_i is *Bernoulli with parameter* θ; that is, $P\{R_i = 1\} = \theta$, $P\{R_i = 0\} = 1 - \theta$, $0 \leq \theta \leq 1$; notice that $R_1 + \cdots + R_n$ has the binomial distribution with parameters n and θ.
(c) Suppose that of N objects, θ are defective. If n objects are drawn without replacement, the probability that exactly x defective objects will be found in the sample is

$$p_\theta(x) = \frac{\binom{\theta}{x}\binom{N-\theta}{n-x}}{\binom{N}{n}}, \qquad x = 0, 1, \ldots, \theta \ (\theta = 0, 1, \ldots, N)$$

This is the hypergeometric probability function; see Problem 7, Section 1.5.
(d) f_θ = the joint density of n independent normally distributed random variables with mean 0 and variance $\theta > 0$.

3. It is desired to test the null hypothesis that a die is unbiased versus the alternative that the die is loaded, with faces 1 and 2 having probability 1/4 and faces 3, 4, 5, and 6 having probability 1/8.
(a) Sketch the set of admissible risk points.
(b) Find a most powerful test at level .1.
(c) Find a Bayes solution if the cost of a type 1 error is c_1, the cost of a type 2 error is $2c_1$, and the null hypothesis has probability 3/4.

4. It is desired to test the null hypothesis that R is normal with mean θ_0 and known variance σ^2 versus the alternative that R is normal with mean $\theta_1 = \theta_0 + \sigma$ and variance σ^2, on the basis of n independent observations of R. Find the minimum value of n such that $\alpha \leq .05$ and $\beta \leq .03$.

5. Consider the problem of testing the null hypothesis that R is normal $(0, \theta_0)$ versus the alternative that R is normal $(0, \theta_1)$, $\theta_1 > \theta_0$ (notice that in this case a UMP test of $\theta \leq \theta_0$ versus $\theta > \theta_0$ exists; see Problem 2d). Describe a most

powerful test at level α and indicate how to find the minimum number of independent observations of R necessary to reduce the probability of a type 2 error below a given figure.

6. Let $R_1, \ldots, R_n$ be independent random variables, each uniformly distributed between 0 and θ, $\theta > 0$. Show that the following test is UMP at level α for $H_0\colon \theta = \theta_0$ versus $H_1\colon \theta \neq \theta_0$.

$$\varphi(x_1, \ldots, x_n) = 1 \quad \text{if } \max_{1\le i\le n} x_i \le \theta_0\alpha^{1/n} \quad \text{or if } \max_{1\le i\le n} x_i > \theta_0$$
$$= 0 \quad \text{otherwise}$$

Find the sketch the power function of the test.

7. Consider the test of Problem 6 with $H_0\colon \theta = 1$; $H_1\colon \theta = 2$. Find the risk set and the set of admissible risk points.

8. Let R_1, R_2, and R_3 be independent, each Bernoulli with parameter θ, $0 \le \theta \le 1$. Find the UMP test of size $\alpha = .1$ of $\theta \le 1/4$ versus $\theta > 1/4$, and find the power function of the test.

9. Show that every admissible test is a Bayes test for some choice of costs c_1 and c_2 and a priori probability p. Conversely, show that every Bayes test with $c_1 > 0$, $c_2 > 0$, $0 < p < 1$ is admissible. Give an example of an inadmissible Bayes test with $c_1 > 0$, $c_2 > 0$.

10. If φ is most powerful at level α_0 and $\beta(\varphi) > 0$, show that φ is actually of size α_0. Give a counterexample to the assertion if $\beta(\varphi) = 0$.

***11.** Let φ be a most powerful test at level α. Show that for some constant λ we have $\varphi(x) = 1$ if $x > \lambda$; $\varphi(x) = 0$ if $x < \lambda$, except possibly for x in a set of Lebesgue measure 0.

12. A class C of tests is said to be *essentially complete* iff for any test φ_1 there is a test $\varphi_2 \in C$ such that φ_2 is as good as φ_1. Show that the following classes are essentially complete.

(a) The likelihood ratio tests.
(b) The admissible tests.
(c) The Bayes tests (i.e., considering all possible c_1, c_2, and p).

13. Give an example of tests φ_1 and φ_2 such that the statements "φ_1 is as good as φ_2" and "φ_2 is as good as φ_1" are both false.

14. Let $R_1, R_2, \ldots$ be independent random variables, each with density h_θ, and let $H_0\colon \theta = \theta_0$, $H_1\colon \theta = \theta_1$.

(a) If φ_n is a test based on n observations that minimizes the sum of the error probabilities, show that $\varphi_n(x) = 1$ if $g_n(x) = \prod_{i=1}^n [h_{\theta_1}(x_i)/h_{\theta_0}(x_i)] > 1$, $\varphi_n(x) = 0$ if $g_n(x) < 1$. Thus

$$\alpha_n + \beta_n = P_{\theta_0}\{x\colon g_n(x) > 1\} + P_{\theta_1}\left\{x\colon \frac{1}{g_n(x)} > 1\right\}$$

(b) Let $t(x_i) = [h_{\theta_1}(x_i)/h_{\theta_0}(x_i)]^{1/2}$. Show that

$$P_{\theta_0}\{x\colon g_n(x) > 1\} \le \prod_{i=1}^n E_{\theta_0} t(R_i) = [E_{\theta_0} t(R_1)]^n$$

(c) Show that $E_{\theta_0} t(R_1) < 1$; hence $\alpha_n \to 0$ as $n \to \infty$. A similar argument with θ_0 and θ_1 interchanged shows that $\beta_n \to 0$ as $n \to \infty$, so that if enough observations are taken, both error probabilities can be made arbitrarily small.

8.3 ESTIMATION

Consider the statistical decision model of Section 8.1. Suppose that γ is a real-valued function on the set N of states of nature, and we wish to estimate $\gamma(\theta)$. If we observe $R = x$ we must produce a number $\psi(x)$ that we hope will be close to $\gamma(\theta)$. Thus the action space A is the set of reals E^1, and a decision function may be specified by giving a (Borel measurable) function ψ from the range of R to E^1; such a ψ is called an *estimate*, and the above decision problem is called a problem of *point estimation of a real parameter*.

Although the estimate ψ appears intrinsically nonrandomized, it is possible to introduce randomization without an essential change in the model. If R_1 is the observable, we let R_2 be a random variable independent of R_1 and θ, with an arbitrary distribution function F. Formally, assume $P_\theta\{R_1 \in B_1, R_2 \in B_2\} = P_\theta\{R_1 \in B_1\}P\{R_2 \in B_2\}$, where $P\{R_2 \in B_2\}$ is determined by the distribution function F and is unaffected by θ. If $R_1 = x$ and $R_2 = y$, we estimate $\gamma(\theta)$ by a number $\psi(x, y)$. Thus we introduce randomization by enlarging the observable.

There is no unique way of specifying a good estimate; we shall discuss several classes of estimates that have desirable properties.

We first consider *maximum likelihood* estimates. Let f_θ be the density (or probability) function corresponding to the state of nature θ, and assume for simplicity that $\gamma(\theta) = \theta$. If $R = x$, the maximum likelihood estimate of θ is given by $\gamma(x) = \hat{\theta} =$ the value of θ that maximizes $f_\theta(x)$. Thus (at least in the discrete case) the estimate is the state of nature that makes the particular observation most likely. In many cases the maximum likelihood estimate is easily computable.

▶ **Example 1.** Let R have the binomial distribution with parameters n and θ, $0 \leq \theta \leq 1$, so that $p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, $x = 0, 1, \ldots, n$. To find $\hat{\theta}$ we may set

$$\frac{\partial}{\partial\theta} \ln p_\theta(x) = 0$$

to obtain

$$\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \qquad \text{or } \hat{\theta} = \frac{x}{n}$$

Notice that R may be regarded as a sum of independent random variables

$R_1, \ldots, R_n$, where R_i is 1 with probability θ and 0 with probability $1 - \theta$. In terms of the R_i we have $\hat{\theta}(R) = (R_1 + \cdots + R_n)/n$, which converges in probability to $E(R_i) = \theta$ by the weak law of large numbers. Convergence in probability of the maximum likelihood estimate to the true parameter can be established under rather general conditions. ◀

▶ **Example 2.** Let $R_1, \ldots, R_n$ be independent, normally distributed random variables with mean μ and variance σ^2. Find the maximum likelihood estimate of $\theta = (\mu, \sigma^2)$. (Here θ is a point in E^2 rather than a real number, but the maximum likelihood estimate is defined as before.)

We have

$$f_\theta(x) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

so that

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

Thus

$$\frac{\partial}{\partial\mu}\ln f_\theta(x) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{n}{\sigma^2}(\bar{x} - \mu)$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$\frac{\partial}{\partial\sigma}\ln f_\theta(x) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{n}{\sigma^3}\left(-\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

Setting the partial derivatives equal to zero, we obtain

$$\hat{\theta} = (\bar{x}, s^2)$$

where

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

(A standard calculus argument shows that this is actually a maximum.) In terms of the R_i, we have

$$\hat{\theta}(R_1, \ldots, R_n) = (\bar{R}, V^2)$$

where $\bar{R}$ is the *sample mean* $(R_1 + \cdots + R_n)/n$ and V^2 is the *sample variance* $(1/n)\sum_{i=1}^{n}(R_i - \bar{R})^2$.

If the problem is changed so that $\theta = \mu$ (i.e., σ^2 is known), we obtain $\hat{\theta} = \bar{R}$ as above. However, if $\theta = \sigma^2$, then we find $\hat{\theta} = (1/n)\sum_{i=1}^{n}(x_i - \mu)^2$, since the equation $\partial \ln f_\theta(x)/\partial\mu = 0$ is no longer present. ◀

We now discuss *Bayes* estimates. For the sake of definiteness we consider the absolutely continuous case. Assume $N = E^1$, and let f_θ be the density of R when the state of nature is θ. Assume that there is an *a priori density* g for θ; that is, the probability that the state of nature will lie in the set B is given by $\int_B g(\theta)\, d\theta$. Finally, assume that we are given a (nonnegative) loss function $L(\gamma(\theta), a)$, $\theta \in N$, $a \in A$; $L(\gamma(\theta), a)$ is the cost when our estimate of $\gamma(\theta)$ turns out to be a. If ψ is an estimate, the over-all average cost associated with ψ is

$$B(\psi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\theta) f_\theta(x) L(\gamma(\theta), \psi(x))\, d\theta\, dx$$

$B(\psi)$ is called the *Bayes risk* of ψ, and an estimate that minimizes $B(\psi)$ is called a *Bayes estimate*. If we write

$$B(\psi) = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(\theta) f_\theta(x) L(\gamma(\theta), \psi(x))\, d\theta\right] dx \tag{8.3.1}$$

it follows that in order to minimize $B(\psi)$ it is sufficient to minimize the expression in brackets for each x.

Often this is computationally feasible. In particular, let $L(\gamma(\theta), a) = (\gamma(\theta) - a)^2$. Thus we are trying to minimize

$$\int_{-\infty}^{\infty} g(\theta) f_\theta(x)(\gamma(\theta) - \psi(x))^2\, d\theta$$

This is of the form $A\psi^2(x) - 2B\psi(x) + C$, which is a minimum when $\psi(x) = B/A$; that is,

$$\psi(x) = \frac{\int_{-\infty}^{\infty} g(\theta) f_\theta(x)\gamma(\theta)\, d\theta}{\int_{-\infty}^{\infty} g(\theta) f_\theta(x)\, d\theta} \tag{8.3.2}$$

But the conditional density of θ given $R = x$ is $g(\theta)f_\theta(x)/\int_{-\infty}^{\infty} g(\theta)f_\theta(x)\, d\theta$, so that $\psi(x)$ is simply the conditional expectation of $\gamma(\theta)$ given $R = x$.

To summarize: To find a Bayes estimate with quadratic loss function, set $\psi(x)$ = the conditional expectation of the parameter to be estimated, given that the observable takes the value x.

▶ **Example 3.** Let R have the binomial distribution with parameters n and θ, $0 \le \theta \le 1$, and let $\gamma(\theta) = \theta$. Take g as the *beta density* with parameters r and s; that is,

$$g(\theta) = \frac{\theta^{r-1}(1-\theta)^{s-1}}{\beta(r, s)}, \qquad 0 \le \theta \le 1,\ r, s > 0$$

where $\beta(r, s)$ is the beta function (see Section 2 of Chapter 4). First we find a Bayes estimate of θ with quadratic loss function.

The discussion leading to (8.3.2) applies, with $f_\theta(x)$ replaced by $p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, $x = 0, 1, \ldots, n$. Thus

$$\psi(x) = \frac{\int_0^1 \binom{n}{x}\theta^{r-1+x+1}(1-\theta)^{s-1+n-x}\,d\theta}{\int_0^1 \binom{n}{x}\theta^{r-1+x}(1-\theta)^{s-1+n-x}\,d\theta}$$

$$= \frac{\beta(r+x+1, n-x+s)}{\beta(r+x, n-x+s)}$$

$$= \frac{\Gamma(r+x+1)\Gamma(n-x+s)}{\Gamma(r+x)\Gamma(n-x+s)}\,\frac{\Gamma(r+s+n)}{\Gamma(r+s+n+1)}$$

$$= \frac{r+x}{r+s+n}$$

Now, for a given θ, the average loss $\rho_\psi(\theta)$, using ψ, may be computed as follows.

$$\rho_\psi(\theta) = E_\theta\left[\left(\frac{r+R}{r+s+n} - \theta\right)^2\right]$$

$$= \frac{1}{(r+s+n)^2}E_\theta[(R - n\theta + r - r\theta - s\theta)^2]$$

Since $E_\theta[(R-n\theta)^2] = \operatorname{Var}_\theta R = n\theta(1-\theta)$ and $E_\theta R = n\theta$, we have

$$\rho_\psi(\theta) = \frac{1}{(r+s+n)^2}[n\theta(1-\theta) + (r - r\theta - s\theta)^2]$$

$$= \frac{1}{(r+s+n)^2}[((r+s)^2 - n)\theta^2 + (n - 2r(r+s))\theta + r^2]$$

ρ_ψ is called the *risk function* of ψ; notice that

$$B(\psi) = \int_0^1 g(\theta)\rho_\psi(\theta)\,d\theta \tag{8.3.3}$$

It is possible to choose r and s so that ρ_ψ will be constant for all θ. For this to happen,

$$n = (r+s)^2 = 2r(r+s)$$

which is satisfied if $r = s = \sqrt{n}/2$. We then have

$$\psi(x) = \frac{x + \sqrt{n}/2}{n + \sqrt{n}} = \frac{\sqrt{n}}{1 + \sqrt{n}} \frac{x}{n} + \frac{1/2}{1 + \sqrt{n}}$$

$$\rho_\psi(\theta) = \frac{n/4}{(n + \sqrt{n})^2} = \frac{1}{4(1 + \sqrt{n})^2} = B(\psi)$$

Thus in this case ψ is a *Bayes estimate with constant risk*; we claim that ψ must be *minimax*, that is, ψ minimizes $\max_\theta \rho_\psi(\theta)$. For if ψ' had a maximum risk smaller than that of ψ, (8.3.3) shows that $B(\psi') < B(\psi)$, contradicting the fact that ψ is Bayes.

Notice that if $\hat{\theta}(x) = x/n$ is the maximum likelihood estimate, then $\psi(x) = a_n\hat{\theta}(x) + b_n$, where $a_n \to 1$, $b_n \to 0$ as $n \to \infty$. ◀

We have not yet discussed randomized estimates; in fact, in a wide variety of situations, including the case of quadratic loss functions, randomization can be ignored. In order to justify this, we first consider a basic theorem concerning convex functions.

A function f from the reals to the reals is said to be *convex* iff $f[(1 - a)x + ay] \leq (1 - a)f(x) + af(y)$ for all real x, y and all $a \in [0, 1]$. A sufficient condition for f to be convex is that it have a nonnegative second derivative ("concave upward" is the phrase used in calculus books). The geometric interpretation is that f lies on or above any of its tangents.

***Theorem 1* (*Jensen's Inequality*).** *If R is a random variable, f is a convex function, and $E(R)$ is finite, then $E[f(R)] \geq f[E(R)]$. (For example, $E[R^{2n}] \geq [E(R)]^{2n}$, $n = 1, 2, \ldots$.)*

PROOF. Consider a tangent to f at the point $E(R)$ (see Figure 8.3.1); let the equation of the tangent be $y = ax + b$. Since f is convex, $f(x) \geq ax + b$ for all x; hence $f(R) \geq aR + b$. Thus $E[f(R)] \geq aE(R) + b = f(E(R))$.

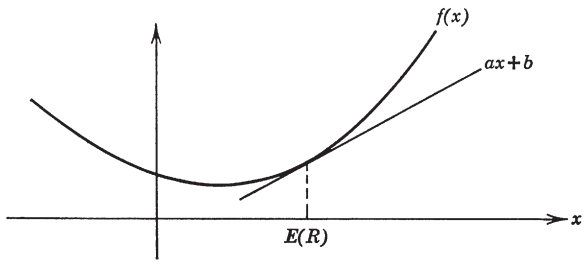

FIGURE 8.3.1 Proof of Jensen's Inequality.

We may now prove the theorem that allows us to ignore randomized estimates.

***Theorem 2* (Rao-Blackwell).** *Let R_1 be an observable, and let R_2 be independent of R_1 and θ, as indicated in the discussion of randomized estimates at the beginning of this section. Let $\psi = \psi(x, y)$ be any estimate of $\gamma(\theta)$ based on observation of R_1 and R_2. Assume that the loss function $L(\gamma(\theta), a)$ is a convex function of a for each θ (this includes the case of quadratic loss). Define*

$$\begin{aligned}\psi^*(x) &= E_\theta[\psi(R_1, R_2) \mid R_1 = x] \\ &= E[\psi(x, R_2)] \\ &\quad (E_\theta\psi(R_1, R_2) \text{ is assumed finite.})\end{aligned}$$

Let ρ_ψ be the risk function of ψ, defined by $\rho_\psi(\theta) = E_\theta[L(\gamma(\theta), \psi(R_1, R_2))] =$ the average loss, using ψ, when the state of nature is θ. Similarly, let $\rho_{\psi^*}(\theta) = E_\theta[L(\gamma(\theta), \psi^*(R_1))]$. Then $\rho_{\psi^*}(\theta) \leq \rho_\psi(\theta)$ for all θ; hence the nonrandomized estimate ψ^* is at least as good as the randomized estimate ψ.

PROOF.

$$L(\gamma(\theta), E_\theta[\psi(R_1, R_2) \mid R_1 = x]) \leq E_\theta[L(\gamma(\theta), \psi(R_1, R_2)) \mid R_1 = x]$$

by the argument of Jensen's inequality applied to conditional expectations. Therefore

$$L(\gamma(\theta), \psi^*(R_1)) \leq E_\theta[L(\gamma(\theta), \psi(R_1, R_2)) \mid R_1]$$

Take expectations on both sides to obtain

$$\rho_{\psi*}(\theta) \leq E_\theta[L(\gamma(\theta), \psi(R_1, R_2))] = \rho_\psi(\theta)$$

as desired.

PROBLEMS

1. Let $R_1, \ldots, R_n$ be independent random variables, all having the same density h_θ; thus $f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n h_\theta(x_i)$. In each case find the maximum likelihood estimate of θ.

(a) $h_\theta(x) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \theta > 0$

(b) $h_\theta(x) = \dfrac{1}{\theta} e^{-x/\theta}, \quad x \geq 0, \theta > 0$

(c) $h_\theta(x) = \dfrac{1}{\theta}, \quad 0 \leq x \leq \theta, \theta > 0$

2. Let R have the Cauchy density with parameter θ; that is,

$$f_\theta(x) = \frac{\theta}{\pi(x^2 + \theta^2)}, \qquad \theta > 0$$

Find the maximum likelihood estimate of θ.

3. Let R have the negative binomial distribution; that is, (see Problem 6, Section 6.4),

$$p_\theta(x) = P\{R = x\} = \binom{x-1}{r-1}\theta^r(1 - \theta)^{x-r}, \qquad x = r, r + 1, \ldots, 0 < \theta \leq 1$$

Find the maximum likelihood estimate of θ.

4. Find the risk function in Example 3, using the maximum likelihood estimate $\hat{\theta} = x/n$.

5. In Example 3, find the Bayes estimate if θ is uniformly distributed between 0 and 1.

6. In Example 3, change the loss function to $L(\theta, a) = (\theta - a)^2/\theta(1 - \theta)$, and let θ be uniformly distributed between 0 and 1. Find the Bayes estimate and show that it has constant risk and is therefore minimax.

7. Let R have the Poisson distribution with parameter $\theta > 0$. Find the Bayes estimate ψ of θ with quadratic loss function if the a priori density is $g(\theta) = e^{-\theta}$. Compute the risk function and the Bayes risk using ψ, and compare with the results using the maximum likelihood estimate.

8.4 SUFFICIENT STATISTICS

In many situations the statistician is concerned with reduction of data. For example, if a sequence of observations results in numbers $x_1, \ldots, x_n$, it is easier to store the single number $x_1 + \cdots + x_n$ than to record the entire set of observations. Under certain conditions no essential information is lost in reducing the data; let us illustrate this by an example.

Let $R_1, \ldots, R_n$ be independent, Bernoulli random variables with parameter θ; that is, $P\{R_i = 1\} = \theta$, $P\{R_i = 0\} = 1 - \theta$, $0 \leq \theta \leq 1$. Let $T = t(R_1, \ldots, R_n) = R_1 + \cdots + R_n$, which has the binomial distribution with parameters n and θ. We claim that $P_\theta\{R_1 = x_1, \ldots, R_n = x_n \mid T = y\}$ actually does not depend on θ. We compute, for $x_i = 0$ or 1, $i = 1, \ldots, n$,

$$P_\theta\{R_1 = x_1, \ldots, R_n = x_n \mid T = y\} = \frac{P_\theta\{R_1 = x_1, \ldots, R_n = x_n, T = y\}}{P_\theta\{T = y\}}$$

This is 0 unless $y = x_1 + \cdots + x_n$, in which case we obtain

$$\frac{P_\theta\{R_1 = x_1, \ldots, R_n = x_n\}}{P_\theta\{T = y\}} = \frac{\theta^y(1 - \theta)^{n-y}}{\binom{n}{y}\theta^y(1 - \theta)^{n-y}} = \frac{1}{\binom{n}{y}}$$

The significance of this result is that for the purpose of making a statistical decision based on observation of $R_1, \ldots, R_n$, we may ignore the individual R_i and base the decision entirely on $R_1 + \cdots + R_n$. To justify this, consider two statisticians, A and B. Statistician A observes $R_1, \ldots, R_n$ and then makes his decision. Statistician B, on the other hand, is only given $T = R_1 + \cdots + R_n$. He then constructs random variables $R_1', \ldots, R_n'$ as follows. If $T = y$, let $R_1', \ldots, R_n'$ be chosen according to the conditional probability function of $R_1, \ldots, R_n$ given $T = y$. Explicitly,

$$P\{R_1' = x_1, \ldots, R_n' = x_n \mid T = y\} = \frac{1}{\binom{n}{y}}$$

where $x_i = 0$ or 1, $i = 1, \ldots, n$, $x_1 + \cdots + x_n = y$. B then follows A's decision procedure, using $R_1', \ldots, R_n'$. Note that since the conditional probability function of $R_1, \ldots, R_n$ given $T = y$ does not depend on the unknown parameter θ, B's procedure is sensible. Now if $x_1 + \cdots + x_n = y$,

$$\begin{aligned} P_\theta\{R_1' = x_1, \ldots, R_n' = x_n\} &= P_\theta\{R_1' = x_1, \ldots, R_n' = x_n, T = y\} \\ &= P_\theta\{T = y\} P_\theta\{R_1' = x_1, \ldots, R_n' = x_n \mid T = y\} \\ &= \binom{n}{y} \theta^y (1 - \theta)^{n-y} \frac{1}{\binom{n}{y}} \\ &= \theta^y (1 - \theta)^{n-y} \\ &= P_\theta\{R_1 = x_1, \ldots, R_n = x_n\} \end{aligned}$$

Thus $(R_1', \ldots, R_n')$ has exactly the same probability function as $(R_1, \ldots, R_n)$, so that the procedures of A and B are equivalent. In other words, anything A can do, B can do at least as well, even though B starts with less information.

We now give the formal definitions. For simplicity, we restrict ourselves to the discrete case. However, the definition of sufficiency in the absolutely continuous case is the same, with probability functions replaced by densities. Also, the basic factorization theorem, to be proved below, holds in the absolutely continuous case (admittedly with a more difficult proof).

Let R be a discrete random variable (or random vector) whose probability function under the state of nature θ is p_θ. Let T be a *statistic* for R, that is, a function of R that is also a random variable. T is said to be *sufficient* for R (or for the family p_θ, $\theta \in N$) iff the conditional probability function of R given T does not depend on θ.

The definition is often unwieldy, and the following criterion for sufficiency is useful.

***Theorem 1* (*Factorization Theorem*).** *Let $T = t(R)$ be a statistic for R. T is sufficient for R if and only if the probability function p_θ can be factored in the form $p_\theta(x) = g(\theta, t(x))h(x)$.*

PROOF. Assume a factorization of this form. Then

$$P_\theta\{R = x \mid T = y\} = \frac{p_\theta\{R = x, T = y\}}{P_\theta\{T = y\}}$$

This is 0 unless $t(x) = y$, in which case we obtain

$$\begin{aligned}\frac{P_\theta\{R = x\}}{P_\theta\{T = y\}} &= \frac{g(\theta, t(x))h(x)}{\sum_{\{z:t(z)=y\}} g(\theta, t(z))h(z)} \\ &= \frac{g(\theta, y)h(x)}{\sum_{\{z:t(z)=y\}} g(\theta, y)h(z)} \\ &= \frac{h(x)}{\sum_{\{z:t(z)=y\}} h(z)}, \qquad \text{which is free of } \theta\end{aligned}$$

Conversely, if T is sufficient, then

$$\begin{aligned}p_\theta(x) &= P_\theta\{R = x\} = P_\theta\{R = x, T = t(x)\} \\ &= P_\theta\{T = t(x)\}P_\theta\{R = x \mid T = t(x)\} \\ &= g(\theta, t(x))h(x) \qquad \text{by definition of sufficiency}\end{aligned}$$

▶ **Example 1.** Let $R_1, \ldots, R_n$ be independent, each Bernoulli with parameter θ. Show that $R_1 + \cdots + R_n$ is sufficient for $(R_1, \ldots, R_n)$.

We have done this in the introductory discussion, using the definition of sufficiency. Let us check the result using the factorization theorem. If $x = x_1 + \cdots + x_n$, $x_i = 0, 1$, and $t(x) = x_1 + \cdots + x_n$, then

$$p_\theta(x_1, \ldots, x_n) = \theta^{t(x)}(1 - \theta)^{n-t(x)}$$

which is of the form specified in the factorization theorem [with $h(x) = 1$]. ◀

▶ **Example 2.** Let $R_1, \ldots, R_n$ be independent, each Poisson with parameter θ. Again $R_1 + \cdots + R_n$ is sufficient for $(R_1, \ldots, R_n)$. (Notice that $R_1 + \cdots + R_n$ is Poisson with parameter $n\theta$.)

For

$$\begin{aligned}p_\theta(x_1, \ldots, x_n) &= P_\theta\{R_1 = x_1, \ldots, R_n = x_n\}, x_1, \ldots, x_n = 0, 1, \ldots \\ &= \prod_{i=1}^{n} P_\theta\{R_i = x_i\} \\ &= \frac{e^{-n\theta}\theta^{x_1+\cdots+x_n}}{x_1! \cdots x_n!}\end{aligned}$$

The factorization theorem applies, with $g(\theta, t(x)) = e^{-n\theta}\theta^{t(x)}$, $h(x) = 1/x_1! \ldots x_n!$, $t(x) = x_1 + \cdots + x_n$. ◀

▶ **Example 3.** Let $R_1, \ldots, R_n$ be independent, each normally distributed with mean μ and variance σ^2. Find a sufficient statistic for $(R_1, \ldots, R_n)$ assuming

(a) μ and σ^2 both unknown; that is, $\theta = (\mu, \sigma^2)$.
(b) σ^2 known; that is, $\theta = \mu$.
(c) μ known; that is, $\theta = \sigma^2$.

[Of course $(R_1, \ldots, R_n)$ is always sufficient for itself, but we hope to reduce the data a bit more.] We compute

$$f_\theta(x) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \tag{8.4.1}$$

Let

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Since $x_i - \bar{x} = x_i - \mu - (\bar{x} - \mu)$, we have

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

Thus

$$f_\theta(x) = (2\pi\sigma^2)^{-n/2} e^{-ns^2/2\sigma^2} e^{-n(\bar{x}-\mu)^2/2\sigma^2} \tag{8.4.2}$$

By (8.4.2), if μ and σ^2 are unknown, then [take $h(x) = 1$] $(\bar{R}, V^2)$ is sufficient, where $\bar{R}$ is the sample mean $(1/n)\sum_{i=1}^{n} R_i$ and V^2 is the sample variance $(1/n)\sum_{i=1}^{n}(R_i - \bar{R})^2$. If σ^2 is known, then the term $(2\pi\sigma^2)^{-n/2}e^{-ns^2/2\sigma^2}$ can be taken as $h(x)$ in the factorization theorem; hence $\bar{R}$ is sufficient. If μ is known, then, by (8.4.1), $\sum_{i=1}^{n}(R_i - \mu)^2$ is sufficient. ◀

PROBLEMS

1. Let $R_1, \ldots, R_n$ be independent, each uniformly distributed on the interval $[\theta_1, \theta_2]$. Find a sufficient statistic for $(R_1, \ldots, R_n)$, assuming
(a) θ_1, θ_2 both unknown
(b) θ_1 known
(c) θ_2 known

2. Repeat Problem 1 if each R_i has the gamma density with parameters θ_1 and θ_2, that is,

$$f(x) = \frac{x^{\theta_1 - 1}e^{-x/\theta_2}}{\Gamma(\theta_1)\theta_2^{\theta_1}}, \qquad x \geq 0,\ \theta_1, \theta_2 > 0$$

3. Repeat Problem 1 if each R_i has the beta density with parameters θ_1 and θ_2, that is,

$$f(x) = \frac{x^{\theta_1-1}(1-x)^{\theta_2-1}}{\beta(\theta_1, \theta_2)}, \qquad 0 \le x \le 1,\ \theta_1, \theta_2 > 0$$

4. Let R_1 and R_2 be independent, with R_1 normal (θ, σ^2), R_2 normal (θ, τ^2), where σ^2 and τ^2 are known. Show that $R_1/\sigma^2 + R_2/\tau^2$ is sufficient for (R_1, R_2).

5. An *exponential family* of densities is a family of the form

$$f_\theta(x) = a(\theta)b(x)\exp\left[\sum_{i=1}^{k} c_i(\theta)t_i(x)\right], \qquad x \text{ real},\ \theta \in N$$

(a) Verify that the following density (or probability) functions can be put into the above form.

(i) Binomial (n, θ): $p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, $\quad x = 0, 1, \ldots, n,\ 0 < \theta < 1$

(ii) Poisson (θ): $p_\theta(x) = \dfrac{e^{-\theta}\theta^x}{x!}$, $\quad x = 0, 1, \ldots,\ \theta > 0$

(iii) Normal (μ, σ^2): $f_\theta(x) = \dfrac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-u)^2/2\theta^2}$, $\quad \theta = (\mu, \sigma^2)$

(iv) Gamma (θ_1, θ_2): $f_\theta(x) = \dfrac{x^{\theta_1-1}e^{-x/\theta_2}}{\Gamma(\theta_1)\theta_2^{\theta_1}}$, $\quad x > 0,\ \theta = (\theta_1, \theta_2),\ \theta_1, \theta_2 > 0$

(v) Beta (θ_1, θ_2): $f_\theta(x) = \dfrac{x^{\theta_1-1}(1-x)^{\theta_2-1}}{\beta(\theta_1, \theta_2)}$, $\quad 0 < x < 1,\ \theta = (\theta_1, \theta_2),\ \theta_1, \theta_2 > 0$

(vi) Negative binomial (r, θ): $p_\theta(x) = \binom{x-1}{r-1}\theta^r(1-\theta)^{x-r}$, $\quad x = r, r+1, \ldots,\ 0 < \theta < 1$, r a known positive integer

(b) If $R_1, \ldots, R_n$ are independent, each R_i having the density f_θ of part (a), find a sufficient statistic for $(R_1, \ldots, R_n)$.

6. Let T be sufficient for the family of densities f_θ, $\theta \in N$. Consider the problem of testing the null hypothesis that $\theta \in H_0$ versus the alternative that $\theta \in H_1$. Show that all possible risk points can be obtained from tests based on T [i.e., $\varphi(x)$ expressible as a function of $t(x)$].

8.5 UNBIASED ESTIMATES BASED ON A COMPLETE SUFFICIENT STATISTIC

In this section we require our estimates ψ of $\gamma(\theta)$ to be *unbiased*; that is, $E_\theta\psi(R) = \gamma(\theta)$ for all $\theta \in N$. Our objective is to show that in a wide class of situations it is possible to construct unbiased estimates ψ that have *uniformly minimum risk*; that is, if ψ' is any unbiased estimate of $\gamma(\theta)$, then $\rho_\psi(\theta) \le \rho_{\psi'}(\theta)$ for all θ. We need a technical definition first. If T is a statistic for R,

T is said to be *complete* iff there are no unbiased estimates of 0 based on T, that is, iff whenever $E_\theta g(T) = 0$ for all $\theta \in N$, we have $P_\theta\{g(T) = 0\} = 1$ for all $\theta \in N$.

Theorem 1. *Let $T = t(R)$ be a complete sufficient statistic for R, and let ψ be an unbiased estimate of $\gamma(\theta)$ based on T [i.e., $\psi(x)$ can be expressed as a function of $t(x)$]. Assume that the loss function $L(\gamma(\theta), a)$ is convex in a for each fixed θ. Then ψ has uniformly minimum risk among all unbiased estimates of $\gamma(\theta)$.*

PROOF. Let ψ' be any unbiased estimate of $\gamma(\theta)$, and define $\psi''(x) = E[\psi'(R) \mid T = t(x)]$. (Since T is sufficient, ψ'' does not depend on θ and hence is a legitimate estimate.) ψ'' is an unbiased estimate of $\gamma(\theta)$ based on T, and so is ψ, and therefore $E_\theta[\psi''(R) - \psi(R)] = 0$ for all θ. But $\psi''(R)$ and $\psi(R)$ can be expressed as functions of T; hence, by completeness,

$$P_\theta\{\psi''(R) = \psi(R)\} = 1 \qquad \text{for all } \theta$$

It follows that $\rho_{\psi''}(\theta) = \rho_\psi(\theta)$ for all θ. But the proof of the Rao-Blackwell theorem, with R_1 replaced by T, x by $t(x)$, $\psi(R_1, R_2)$ by $\psi'(R)$, and $\psi^*(R_1)$ by $\psi''(R)$, shows that $\rho_{\psi''}(\theta) \leq \rho_{\psi'}(\theta)$ for all θ, as desired.

If $L(\gamma(\theta), a) = (\gamma(\theta) - a)^2$, then $\rho_\psi(\theta) = E_\theta[(\gamma(\theta) - \psi(R))^2] = \operatorname{Var}_\theta \psi(R)$. Thus ψ has the smallest variance of all unbiased estimates of $\gamma(\theta)$, regardless of the state of nature. In this case ψ is said to be a *uniformly minimum variance unbiased estimate* (UMVUE).

▶ **Example 1.** Let $R_1, \ldots, R_n$ be independent, each Bernoulli with parameter θ, $0 \leq \theta \leq 1$. By Example 1, Section 8.4, $T = R_1 + \cdots + R_n$ is sufficient for $(R_1, \ldots, R_n)$; let us show that it is complete.

Now T is binomial with parameters n and θ; hence

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

$$= \left[\sum_{k=0}^{n} g(k)\binom{n}{k}\left(\frac{\theta}{1-\theta}\right)^k\right](1-\theta)^n \qquad \text{if } \theta < 1$$

If $E_\theta g(T) = 0$ for all $\theta \in [0, 1]$, then $\sum_{k=0}^{n} g(k)\binom{n}{k}z^k = 0$ for all $z \in [0, \infty)$; hence $g(k) = 0$ for $k = 0, 1, \ldots, n$.

We now look for unbiased estimates of $\gamma(\theta)$ based on T. If $\psi(x_1, \ldots, x_n) = g(t(x_1, \ldots, x_n))$, $t(x_1, \ldots, x_n) = x_1 + \cdots + x_n$, is such an estimate, the above argument shows that $E_\theta\psi(R_1, \ldots, R_n) = E_\theta g(T)$ is a polynomial in

θ of degree $\leq n$. Thus $\gamma(\theta)$ must be of the form $a_0 + a_1\theta + \cdots + a_n\theta^n$; furthermore, an unbiased estimate of such an expression is easily found. If $T^{(r)} = T(T-1)\cdots(T-r+1)$, $n^{(r)} = (n-1)\cdots(n-r+1)$, then

$$E\left[\frac{T^{(r)}}{n^{(r)}}\right] = \sum_{k=0}^{n} \frac{k(k-1)\cdots(k-r+1)}{n(n-1)\cdots(n-r+1)} \frac{n!}{k!\,(n-k)!} \theta^k(1-\theta)^{n-k}$$

$$= \theta^r \sum_{k=r}^{n} \frac{(n-r)!}{(k-r)!\,(n-k)!} \theta^{k-r}(1-\theta)^{n-k} = \theta^r(\theta + 1 - \theta)^{n-r}$$

$$= \theta^r$$

Thus $\sum_{k=0}^{n} a_k[T^{(k)}/n^{(k)}]$ is a UMVUE of $\gamma(\theta) = \sum_{k=0}^{n} a_k\theta^k$; in particular the sample mean T/n is a UMVUE of θ. ◀

▶ **Example 2.** Let $R_1, \ldots, R_n$ be independent, each Poisson with parameter θ. By Example 2, Section 8.4, $T = R_1 + \cdots + R_n$ is sufficient for $(R_1, \ldots, R_n)$; T is also complete. For T is Poisson with parameter $n\theta$; hence

$$E_\theta g(T) = \sum_{k=0}^{\infty} g(k)e^{-n\theta} \frac{(n\theta)^k}{k!}$$

If $E_\theta g(T) = 0$ for all $\theta > 0$, then

$$\sum_{k=0}^{\infty} \left[\frac{g(k)n^k}{k!}\right] \theta^k = 0 \qquad \text{for all } \theta > 0$$

Since this is a power series in θ, we must have $g \equiv 0$.

If $\psi(x_1, \ldots, x_n) = g(t(x_1, \ldots, x_n))$ is an unbiased estimate of $\gamma(\theta)$, then

$$\gamma(\theta) = E_\theta \psi(R_1, \ldots, R_n) = E_\theta g(T) = e^{-n\theta} \sum_{k=0}^{\infty} g(k) \frac{(n\theta)^k}{k!}$$

Thus $\gamma(\theta)$ must be expressible as a power series in θ. If $\gamma(\theta) = \sum_{k=0}^{\infty} a_k\theta^k$, then

$$\gamma(\theta)e^{n\theta} = \sum_{j=0}^{\infty} a_j\theta^j \sum_{k=0}^{\infty} \frac{(n\theta)^k}{k!}$$

$$= \sum_{k=0}^{\infty} c_k\theta^k \qquad \text{where } c_k = \sum_{i=0}^{k} \frac{n^i}{i!} a_{k-i}$$

But

$$\gamma(\theta)e^{n\theta} = \sum_{k=0}^{\infty} \frac{g(k)n^k}{k!} \theta^k$$

hence

$$g(k) = \frac{k!\, c_k}{n^k}$$

$$= \sum_{i=0}^{k} \frac{k!}{i!} \frac{a_{k-i}}{n^{k-i}}$$

We conclude that

$$\sum_{i=0}^{T} \frac{T!}{i!} \frac{a_{T-i}}{n^{T-i}} \quad \text{is a UMVUE of} \quad \gamma(\theta) = \sum_{k=0}^{\infty} a_k \theta^k$$

For example, if $\gamma(\theta) = \theta^r$, $r = 1, 2, \ldots$, the UMVUE is

$$\frac{T!}{(T-r)!} \frac{1}{n^r} = \frac{T^{(r)}}{n^r} \left(= \frac{T}{n} = \text{the sample mean when } r = 1 \right)$$

[In this particular case the above computation could have been avoided, since we know that $E_\theta(T^{(r)}) = (n\theta)^r$ (Problem 8, Section 3.2). Since $T^{(r)}/n^r$ is an unbiased estimate of θ^r based on T, it is a UMVUE.]

As another example, a UMVUE of $1/(1-\theta) = \sum_{k=0}^{\infty} \theta^k$, $0 < \theta < 1$, is

$$\sum_{i=0}^{T} \frac{T!}{i!} \frac{1}{n^{T-i}} \blacktriangleleft$$

PROBLEMS

1. Find a UMVUE of $e^{-\theta}$ in Example 2.
2. Let $R_1, \ldots, R_n$ be independent, each uniformly distributed between 0 and $\theta > 0$. By Problem 1, Section 8.4, $T = \max R_i$ is sufficient for $(R_1, \ldots, R_n)$.
   (a) Show that T is complete.
   (b) Find a UMVUE of $\gamma(\theta)$, assuming that γ extends to a function with a continuous derivative on $[0, \infty)$, and $\theta^n \gamma(\theta) \to 0$ as $\theta \to 0$. [In part (a), use without proof the fact that if $\int_0^\theta h(y)\, dy = 0$ for all $\theta > 0$, then $h(y) = 0$ except on a set of Lebesgue measure 0. Notice that if it is known that h is continuous, then $h \equiv 0$ by the fundamental theorem of calculus.]
3. Let $R_1, \ldots, R_n$ be independent, each normal with mean θ and known variance σ^2.
   (a) Show that the sample mean $\bar{R}$ is a UMVUE of θ.
   (b) Show that $(\bar{R})^2 - (\sigma^2/n)$ is a UMVUE of θ^2.
   [Use without proof the fact that if $\int_{-\infty}^{\infty} h(y) e^{\theta y}\, dy = 0$ for all $\theta > 0$, then $h(y) = 0$ except on a set of Lebesgue measure 0.]

4. Let $R_1, \ldots, R_n$ be independent, with $P\{R_i = k\} = 1/N$, $k = 1, \ldots, N$; take $\theta = N$, $N = 1, 2, \ldots$.
(a) Show that $\max_{1 \le i \le n} R_i$ is a complete sufficient statistic.
(b) Find a UMVUE of $\gamma(N)$.

5. Let R have the negative binomial distribution:

$$P\{R = k\} = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \qquad k = r, r+1, \ldots, 0 < p \le 1$$

Take $\theta = 1 - p$, $0 \le \theta < 1$. Show that $\gamma(\theta)$ has a UMVUE if and only if it is expressible as a power series in θ; find the form of the UMVUE.

6. Let

$$P\{R = k\} = \frac{e^{-\theta}}{1 - e^{-\theta}} \frac{\theta^k}{k!}, \qquad k = 1, 2, \ldots, \theta > 0$$

(This is the conditional probability function of a Poisson random variable R', given that $R' \ge 1$.) R is clearly sufficient for itself, and is complete by an argument similar to that of Example 2.
(a) Find a UMVUE of $e^{-\theta}$.
(b) Show that (assuming quadratic loss function) the estimate ψ found in part (a) is *inadmissible*; that is, there is another estimate ψ' such that $\rho_{\psi'}(\theta) \le \rho_{\psi}(\theta)$ for all θ, and $\rho_{\psi'}(\theta) < \rho_{\psi}(\theta)$ for some θ. This shows that unbiased estimates, while often easy to find, are not necessarily desirable.

7. The following is another method for obtaining a UMVUE. Let $R_1, \ldots, R_n$ be independent, each Bernoulli with parameter θ, $0 \le \theta \le 1$, as in Example 1. If $j = 1, \ldots, n$, then

$$E\left[\prod_{i=1}^{j} R_i\right] = P\{R_1 = \cdots = R_j = 1\} = \theta^j$$

Thus $R_1 R_2 \cdots R_j$ is an unbiased estimate of θ^j. But then $\psi(k) = E[R_1 \cdots R_j \mid \sum_{i=1}^n R_i = k\}$ is an unbiased estimate of θ^j based on the complete sufficient statistic $\sum_{i=1}^n R_i$, so that ψ is a UMVUE. Compute ψ directly and show that the result agrees with Example 1.

8. Let $R_1, \ldots, R_n$ be independent, each Poisson with parameter $\theta > 0$. Show, using the analysis in Problem 7, that

$$E\left(R_1 R_2 \mid \sum_{i=1}^{n} R_i = k\right) = \frac{k(k-1)}{n^2}, \qquad k = 0, 1, \ldots$$

9. Let $R_1, \ldots, R_n$ be independent, each uniformly distributed between 0 and θ; if $T = \max R_i$, then $[(n+1)/n]T$ is a UMVUE of θ (see Problem 2). Compare the risk function $E_\theta[((1 + 1/n)T - \theta)^2]$ using $[(n+1)/n]T$ with the risk function $E_\theta[((2/n)\sum_{i=1}^n R_i - \theta)^2]$ using the unbiased estimate $(2/n)\sum_{i=1}^n R_i$.

10. Let $R_1, \ldots, R_n$ be independent, each Bernoulli with parameter $\theta \in [0, 1]$. Show that (assuming quadratic loss function) there is no best estimate of θ based on $R_1, \ldots, R_n$; that is, there is no estimate ψ such that $\rho_\psi(\theta) \le \rho_{\psi'}(\theta)$ for all θ and all estimates ψ' of θ.

11. If $E_\theta\psi_1(R) = E_\theta\psi_2(R) = \gamma(\theta)$ and ψ_1, ψ_2 both minimize $\rho_\psi(\theta) = E_\theta[(\psi(R) - \gamma(\theta))^2]$, θ fixed, show that $P_\theta\{\psi_1(R) = \psi_2(R)\} = 1$. Consequently, if ψ_1 and ψ_2 are UMVUEs of $\gamma(\theta)$, then, for each θ, $\psi_1(R) = \psi_2(R)$ with probability 1.

12. Let f_θ, $\theta \in N$ = an open interval of reals, be a family of densities. Assume that $\partial f_\theta(x)/\partial\theta$ exists and is continuous everywhere, and that $\int_{-\infty}^{\infty} f_\theta(x)\,dx$ can be differentiated under the integral sign with respect to θ.

(a) If R has density f_θ when the state of nature is θ, show that

$$E_\theta\left[\frac{\partial}{\partial\theta}\ln f_\theta(R)\right] = 0$$

(b) If $E_\theta\psi(R) = \gamma(\theta)$ and $\int_{-\infty}^{\infty}\psi(x)f_\theta(x)\,dx$ can be differentiated under the integral sign with respect to θ, show that

$$\frac{d}{d\theta}\gamma(\theta) = E_\theta\left[\psi(R)\frac{\partial}{\partial\theta}\ln f_\theta(R)\right]$$

(c) Under the assumptions of part (b), show that

$$\operatorname{Var}_\theta \psi(R) \geq \frac{[(d/d\theta)\,\gamma(\theta)]^2}{[E_\theta(\partial \ln f_\theta(R)/\partial\theta)^2]}$$

if the denominator is >0. In particular, if $f_\theta(x) = f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n h_\theta(x_i)$, then

$$\begin{aligned} E_\theta\left(\frac{\partial}{\partial\theta}\ln f_\theta(R)\right)^2 &= \operatorname{Var}_\theta \frac{\partial}{\partial\theta}\ln f_\theta(R) \\ &= n\operatorname{Var}_\theta \frac{\partial}{\partial\theta}\ln h_\theta(R_i) \\ &= nE_\theta\left(\frac{\partial}{\partial\theta}\ln h_\theta(R_i)\right)^2 \end{aligned}$$

where $R = (R_1, \ldots, R_n)$.

The above result is called the *Cramer-Rao inequality* (an analogous theorem may be proved with densities replaced by probability functions). If ψ is an estimate that satisfies the Cramer-Rao lower bound with equality for all θ, then ψ is a UMVUE of $\gamma(\theta)$. This idea may be used to give an alternative proof that the sample mean is a UMVUE of the true mean in the Bernoulli, Poisson, and normal cases (see Examples 1 and 2 and Problem 3 of this section).

13. If $R_1, \ldots, R_n$ are independent, each with mean μ and variance σ^2, and V^2 is the sample variance, show that V^2 is a biased estimate of σ^2; specifically,

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

8.6 SAMPLING FROM A NORMAL POPULATION

If $R_1, \ldots, R_n$ are independent, each normally distributed with mean μ and variance σ^2, we have seen that $(\bar{R}, V^2)$, where $\bar{R}$ is the sample mean $(1/n)(R_1 + \cdots + R_n)$ and $V^2 = (1/n)\sum_{i=1}^{n}(R_i - \bar{R})^2$ is the sample variance, is a sufficient statistic for $(R_1, \ldots, R_n)$. $\bar{R}$ and V^2 have some special properties that are often useful. First, $\bar{R}$ is a sum of the independent normal random variables R_i/n, each of which has mean μ/n and variance σ^2/n^2; hence $\bar{R}$ *is normal with mean* μ *and variance* σ^2/n. *We* now prove that $\bar{R}$ and V^2 are independent.

Theorem 1. *If* $R_1, \ldots, R_n$ *are independent, each normal* (μ, σ^2), *the associated sample mean and variance are independent random variables.*

*PROOF. Define random variables $W_1, \ldots, W_n$ by

$$W_1 = \frac{1}{\sqrt{n}} R_1 + \cdots + \frac{1}{\sqrt{n}} R_n$$
$$W_2 = c_{21}R_1 + \cdots + c_{2n}R_n$$
$$\vdots$$
$$W_n = c_{n1}R_1 + \cdots + c_{nn}R_n$$

where the c_{ij} are chosen so as to make the transformation orthogonal. [This may be accomplished by extending the vector $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$ to an orthonormal basis for E^n.] The Jacobian J of the transformation is the determinant of the orthogonal matrix $A = [c_{ij}]$ (with $c_{1j} = 1/\sqrt{n}$, $j = 1, \ldots, n$), namely, ± 1. Thus (see Problem 12, Section 2.8) the density of $(W_1, \ldots, W_n)$ is given by

$$f^*(y_1, \ldots, y_n) = \frac{f(x_1, \ldots, x_n)}{|J|}$$
$$= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

where

$$\begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = A^{-1} \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}$$

Since $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$ by orthogonality, and $\sum_{i=1}^n x_i = \sqrt{n}\, y_1$,

$$f^*(y_1, \ldots, y_n) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - 2\mu\sqrt{n}\, y_1 + n\mu^2\right)\right]$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(y_1 - \sqrt{n}\,\mu)^2}{2\sigma^2}\right] \prod_{i=2}^n \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{y_i^2}{2\sigma^2}\right)$$

It follows that $W_1, \ldots, W_n$ are independent, with $W_2, \ldots, W_n$ each normal $(0, \sigma^2)$ and W_1 normal $(\sqrt{n}\,\mu, \sigma^2)$. But

$$nV^2 = \sum_{i=1}^n (R_i - \bar{R})^2 = \sum_{i=1}^n R_i^2 - 2\bar{R}\sum_{i=1}^n R_i + n(\bar{R})^2$$

$$= \sum_{i=1}^n R_i^2 - n(\bar{R})^2$$

$$= \sum_{i=1}^n W_i^2 - \left(\sum_{i=1}^n \frac{R_i}{\sqrt{n}}\right)^2$$

$$= \sum_{i=1}^n W_i^2 - W_1^2$$

$$= \sum_{i=2}^n W_i^2$$

Since $\sqrt{n}\,\bar{R} = W_1$, it follows that $\bar{R}$ and V^2 are independent, completing the proof.

The above argument also gives us the distribution of the sample variance. For

$$\frac{nV^2}{\sigma^2} = \sum_{i=2}^n \left(\frac{W_i}{\sigma}\right)^2$$

where the W_i/σ are independent, each normal $(0, 1)$. Thus nV^2/σ^2 *has the chi-square distribution with* $n - 1$ *degrees of freedom*; that is, the density of nV^2/σ^2 is

$$\frac{1}{2^{(n-1)/2}\Gamma((n-1)/2)}\, x^{(n-3)/2} e^{-x/2}, \qquad x \geq 0$$

(see Problem 3, Section 5.2).

Now since $\bar{R}$ is normal $(\mu, \sigma^2/n)$, $\sqrt{n}\,(\bar{R} - \mu)/\sigma$ is normal $(0, 1)$; hence

$$P\left\{-b \leq \sqrt{n}\,\frac{(\bar{R} - \mu)}{\sigma} \leq b\right\} = F^*(b) - F^*(-b)$$

$$= 2F^*(b) - 1$$

where F^* is the normal (0, 1) distribution function. If b is chosen so that $2F^*(b) - 1 = 1 - \alpha$, that is, $F^*(b) = 1 - \alpha/2$ ($b = N_{\alpha/2}$ in the terminology of Example 3, Section 8.2),

$$P\left\{-N_{\alpha/2} \leq \sqrt{n}\,\frac{(\bar{R} - \mu)}{\sigma} \leq N_{\alpha/2}\right\} = 1 - \alpha$$

or

$$P\left\{\bar{R} - \frac{\sigma N_{\alpha/2}}{\sqrt{n}} \leq \mu \leq \bar{R} + \frac{\sigma N_{\alpha/2}}{\sqrt{n}}\right\} = 1 - \alpha$$

Thus, with probability $1 - \alpha$, the true mean μ lies in the random interval

$$I = \left[\bar{R} - \frac{\sigma N_{\alpha/2}}{\sqrt{n}}, \qquad \bar{R} + \frac{\sigma N_{\alpha/2}}{\sqrt{n}}\right]$$

I is called a *confidence interval* for μ with *confidence coefficient* $1 - \alpha$.

The interval I is computable from the given observations of $R_1, \ldots, R_n$, provided that σ^2 is known. If σ^2 is unknown, it is natural to replace the true variance σ^2 by the sample variance V^2. However, we then must know something about the random variable $(\bar{R} - \mu)/V$. In order to provide the necessary information, we do the following computation.

Let R_1 be normal (0, 1), and let R_2 have the chi square distribution with m degrees of freedom; assume that R_1 and R_2 are independent. We compute the density of $\sqrt{m}\, R_1/\sqrt{R_2}$, as follows.

Let

$$W_1 = \frac{\sqrt{m}\, R_1}{\sqrt{R_2}}$$

$$W_2 = R_2$$

so that

$$R_1 = \frac{W_1\sqrt{W_2}}{\sqrt{m}}$$

$$R_2 = W_2$$

Thus we have a transformation of the form

$$(y_1, y_2) = g(x_1, x_2) = \left(\frac{\sqrt{m}\, x_1}{\sqrt{x_2}}, x_2\right)$$

with inverse given by

$$(x_1, x_2) = h(y_1, y_2) = \left(\frac{y_1\sqrt{y_2}}{\sqrt{m}}, y_2\right)$$

g is defined on $\{(x_1, x_2) \in E^2 : x_2 > 0\}$ and h on $\{(y_1, y_2) \in E^2 : y_2 > 0\}$.

By Problem 12, Section 2.8, the density of (W_1, W_2) is given by

$$f^*_{12}(y_1, y_2) = f_{12}(h(y_1, y_2))\,|J_h(y_1, y_2)|, \qquad y_2 > 0$$

where

$$J_h(y_1, y_2) = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} \dfrac{\sqrt{y_2}}{\sqrt{m}} & \dfrac{y_1}{2\sqrt{my_2}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{y_2}{m}}$$

Thus

$$f^*_{12}(y_1, y_2) = \frac{1}{\sqrt{2\pi}}\, e^{-y_1^2 y_2/2m}\, \frac{1}{2^{m/2}\Gamma(m/2)}\, y_2^{m/2-1} e^{-y_2/2} \sqrt{\frac{y_2}{m}}$$

Therefore the density of W_1 is

$$f_{W_1}(y_1) = \int_0^\infty f^*_{12}(y_1, y_2)\, dy_2$$

But

$$\int_0^\infty y_2^{(m+1)/2-1} e^{-(1+y_1^2/m)y_2/2} dy_2 = \frac{\Gamma((m+1)/2)2^{(m+1)/2}}{(1 + y_1^2/m)^{(m+1)/2}}$$

Hence

$$f_{W_1}(y_1) = \frac{\Gamma((m+1)/2)}{\sqrt{m\pi}\,\Gamma(m/2)} \frac{1}{(1 + y_1^2/m)^{(m+1)/2}}$$

A random variable with this density is said to have the *t distribution* with m degrees of freedom.

An application of Stirling's formula shows that the t density approaches the normal (0, 1) density as $m \to \infty$.

Now we know that $\sqrt{n}\,(\bar{R} - \mu)/\sigma$ is normal (0, 1), and nV^2/σ^2 has the chi-square distribution with $n - 1$ degrees of freedom. Thus

$$\frac{\sqrt{n-1}\,\sqrt{n}\,(\bar{R} - \mu)/\sigma}{\sqrt{n}\,V/\sigma} = \sqrt{n-1}\,\frac{(\bar{R} - \mu)}{V} = \frac{\bar{R} - \mu}{\left[(1/n(n-1))\sum_{i=1}^{n}(R_i - \bar{R})^2\right]^{1/2}}$$

has the t distribution with $n - 1$ degrees of freedom.

If $t_{\beta,m}$ is such that $\int_{t_{\beta,m}}^\infty h_m(t)\, dt = \beta$, where h_m is the t density with m degrees of freedom, then

$$P\left\{-t_{\alpha/2,n-1} \le \sqrt{n-1}\,\frac{(\bar{R} - \mu)}{V} \le t_{\alpha/2,n-1}\right\} = 1 - \alpha$$

Thus

$$\left[\bar{R} - \frac{Vt_{\alpha/2,n-1}}{\sqrt{n-1}}, \qquad \bar{R} + \frac{Vt_{\alpha/2,n-1}}{\sqrt{n-1}}\right]$$

is a confidence interval for μ with confidence coefficient $1 - \alpha$.

PROBLEMS

1. Let R_1 and R_2 be independent, chi-square random variables with m and n degrees of freedom, respectively. Show that $(R_1/m)/(R_2/n)$ has density

$$f_{mn}(x) = \frac{(m/n)^{m/2}}{\beta(m/2, n/2)} \frac{x^{(m/2)-1}}{(1 + mx/n)^{(m+n)/2}}, \qquad x \geq 0$$

$(R_1/m)/(R_2/n)$ is said to have the *F distribution with m and n degrees of freedom*, abbreviated $F(m, n)$.

2. Calculate the mean and variance of the chi-square, t, and F distributions.

3. (a) If T has the t distribution with n degrees of freedom, show that T^2 has the $F(1, n)$ distribution.
(b) If R has the $F(m, n)$ distribution, show that $1/R$ has the $F(n, m)$ distribution.
(c) If R_1 is chi-square (m) and R_2 is chi-square (n), show that $R_1 + R_2$ is chi-square $(m + n)$.

4. Discuss the problem of obtaining confidence intervals for the variance σ^2 of a normally distributed random variable, assuming that
(a) The mean μ is known
(b) μ is unknown

5. (A two-sample problem) Let $R_{11}, R_{12}, \ldots, R_{1n_1}, R_{21}, R_{22}, \ldots, R_{2n_2}$ be independent, with the R_{1j} normal (μ_1, σ^2) and the R_{2j} normal (μ_2, σ^2) (μ_1, μ_2 and σ^2 unknown). Thus we are taking independent samples from two different normal populations. Show that if $\bar{R}_i$ and V_i^2, $i = 1, 2$, are the sample mean and variance of the two samples, and

$$k = (n_1 V_1^2 + n_2 V_2^2)^{1/2} \left[\frac{n_1 + n_2}{n_1 n_2 (n_1 + n_2 - 2)}\right]^{1/2}$$

then $[\bar{R}_1 - \bar{R}_2 - kt_{\alpha/2,\, n_1+n_2-2},\ \bar{R}_1 - \bar{R}_2 + kt_{\alpha/2, n_1+n_2-2}]$ is a confidence interval for $\mu_1 - \mu_2$ with confidence coefficient $1 - \alpha$.

6. In Problem 5, assume that the samples have different variances σ_1^2 and σ_2^2. Discuss the problem of obtaining confidence intervals for the ratio σ_1^2/σ_2^2.

7. (a) Suppose that $C(R)$ is a *confidence set* for $\gamma(\theta)$ with confidence coefficient $\geq 1 - \alpha$; that is,

$$P_\theta\{\gamma(\theta) \in C(R)\} \geq 1 - \alpha \qquad \text{for all } \theta \in N$$

Consider the hypothesis-testing problem

$$H_0: \gamma(\theta) = k$$
$$H_1: \gamma(\theta) \neq k$$

and the following test.

$$\varphi_k(x) = 1 \qquad \text{if } k \notin C(x)$$
$$= 0 \qquad \text{if } k \in C(x)$$

[Thus $C(x)$ is the *acceptance region* of φ_k.] Show that φ_k is a test at level α.

(b) Suppose that for all k in the range of γ there is a nonrandomized test φ_k [i.e., $\varphi_k(x) = 0$ or 1 for all x] at level α for H_0: $\gamma(\theta) = k$ versus H_1: $\gamma(\theta) \neq k$. Let $C(x)$ be the set $\{k: \varphi_k(x) = 0\}$. Show that $C(R)$ is a confidence set for $\gamma(\theta)$ with confidence coefficient $\geq 1 - \alpha$.

This result allows the confidence interval examples in this section to be translated into the language of hypothesis testing.

*8.7 THE MULTIDIMENSIONAL GAUSSIAN DISTRIBUTION

If $R_1', \ldots, R_n'$ are independent, normally distributed random variables and we define random variables $R_1, \ldots, R_n$ by $R_i = \sum_{j=1}^n a_{ij}R_j' + b_j$, $i = 1, \ldots, n$, the R_i have a distribution of considerable importance in many aspects of probability and statistics. In this section we examine the properties of this distribution and make an application to the problem of prediction.

Let $R = (R_1, \ldots, R_n)$ be a random vector. The *characteristic function* of R (or the *joint characteristic function* of $R_1, \ldots, R_n$) is defined by

$$M(u_1, \ldots, u_n) = E[i(u_1R_1 + \cdots + u_nR_n)], \qquad u_1, \ldots, u_n \text{ real}$$
$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(i \sum_{k=1}^n u_k x_k\right) dF(x_1, \ldots, x_n)$$

where F is the distribution function of R. It will be convenient to use a vector-matrix notation. If $u = (u_1, \ldots, u_n) \in E^n$, $\mathbf{u}$ will denote the column vector with components $u_1, \ldots, u_n$. Similarly we write $\mathbf{x}$ for col $(x_1, \ldots, x_n)$ and $\mathbf{R}$ for col $(R_1, \ldots, R_n)$. A superscript t will indicate the transpose of a matrix.

Just as in one dimension, it can be shown that the characteristic function determines the distribution function uniquely.

DEFINITION. The random vector $R = (R_1, \ldots, R_n)$ is said to be *Gaussian* (or $R_1, \ldots, R_n$ are said to be *jointly Gaussian*) iff the characteristic function of R is

$$M(u_1, \ldots, u_n) = \exp[i\mathbf{u}^t\mathbf{b}] \exp[-\tfrac{1}{2}\mathbf{u}^tK\mathbf{u}]$$
$$= \exp\left[i \sum_{r=1}^n u_r b_r - \frac{1}{2} \sum_{r,s=1}^n u_r K_{rs} u_s\right] \tag{8.7.1}$$

where $b_1, \ldots, b_n$ are arbitrary real numbers and K is an arbitrary real symmetric nonnegative definite n by n matrix. (Nonnegative definite means that $\sum_{r,s=1}^n a_r K_{rs} a_s$ is real and ≥ 0 for all real numbers $a_1, \ldots, a_n$.)

We must show that there is a random vector with this characteristic function. We shall do this in the proof of the next theorem.

Theorem 1. *Let R be a random n-vector. R is Gaussian iff $\mathbf{R}$ can be expressed as $W\mathbf{R}' + \mathbf{b}$, where $b = (b_1, \ldots, b_n) \in E^n$, W is an n by n matrix, and $R_1', \ldots, R_n'$ are independent normal random variables with 0 mean.*

The matrix K of (8.7.1) is given by WDW^t, where $D = \text{diag}\ (\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix with entries $\lambda_j = Var\ R_j'$, $j = 1, \ldots, n$. (To avoid having to treat the case $\lambda_j = 0$ separately, we agree that normal with expectation m and variance 0 will mean degenerate at m.)

Furthermore, the matrix W can be taken as orthogonal.

PROOF. If $\mathbf{R} = W\mathbf{R}' + \mathbf{b}$, then

$$E[\exp\ (i\mathbf{u}^t\mathbf{R})] = \exp\ [i\mathbf{u}^t\mathbf{b}]\ E[\exp\ (i\mathbf{u}^tW\mathbf{R}')]$$

But

$$\begin{aligned} E[\exp\ (i\mathbf{v}^t\mathbf{R}')] &= E\left[\prod_{k=1}^{n} \exp\ (iv_kR_k')\right] \\ &= \prod_{k=1}^{n} E[\exp\ (iv_kR_k')] = \exp\left[-\frac{1}{2}\sum_{k=1}^{n}\lambda_k{v_k}^2\right] \\ &= \exp\ [-\tfrac{1}{2}\mathbf{v}^tD\mathbf{v}] \end{aligned}$$

Set $\mathbf{v} = W^t\mathbf{u}$ to obtain

$$E[\exp\ (i\mathbf{u}^t\mathbf{R})] = \exp\ [i\mathbf{u}^t\mathbf{b} - \tfrac{1}{2}\mathbf{u}^tK\mathbf{u}]$$

where $K = WDW^t$. K is clearly symmetric, and is also nonnegative definite, since $\mathbf{u}^tK\mathbf{u} = \mathbf{v}^tD\mathbf{v} = \sum_{k=1}^{n}\lambda_k{v_k}^2 \geq 0$, where $\mathbf{v} = W^t\mathbf{u}$. Thus R is Gaussian. (Notice also that if K is symmetric and nonnegative definite, there is an orthogonal matrix W such that $W^tKW = D$, where D is the diagonal matrix of eigenvalues of K. Thus $K = WDW^t$, so that it is always possible to construct a Gaussian random vector corresponding to a prescribed K and b.)

Conversely, let R have characteristic function $\exp\ [i\mathbf{u}^t\mathbf{b} - (1/2)(\mathbf{u}^tK\mathbf{u})]$, where K is symmetric and nonnegative definite. Let W be an orthogonal matrix such that $W^tKW = D = \text{diag}\ (\lambda_1, \ldots, \lambda_n)$, where the λ_j are the eigenvalues of K. Let $\mathbf{R}' = W^t(\mathbf{R} - \mathbf{b})$. Then

$$\begin{aligned} E[\exp\ (i\mathbf{u}^t\mathbf{R}')] &= \exp\ (-i\mathbf{u}^tW^t\mathbf{b})E[\exp\ (i\mathbf{u}^tW^t\mathbf{R})] \\ &= \exp\ [-\tfrac{1}{2}\mathbf{v}^tK\mathbf{v}] \qquad \text{where} \qquad \mathbf{v} = W\mathbf{u} \\ &= \exp\ [-\tfrac{1}{2}\mathbf{u}^tD\mathbf{u}] = \exp\left[-\frac{1}{2}\sum_{k=1}^{n}\lambda_k{u_k}^2\right] \end{aligned}$$

It follows that $R_1', \ldots, R_n'$ are independent, with R_j normal $(0, \lambda_j)$. Since W is orthogonal, $W^t = W^{-1}$; hence $\mathbf{R} = W\mathbf{R}' + \mathbf{b}$.

The matrix K has probabilistic significance, as follows.

Theorem 2. *In Theorem 1 we have $E(\mathbf{R}) = \mathbf{b}$, that is, $E(R_j) = b_j$, $j = 1, \ldots, n$, and K is the covariance matrix of the R_j, that is, $K_{rs} = Cov(R_r, R_s)$, $r, s = 1, \ldots, n$.*

PROOF. Since the R_j' have finite second moments, so do the R_j. $E(\mathbf{R}) = \mathbf{b}$ follows immediately by linearity of the expectation. Now the covariance matrix of the R_j is

$$[\text{Cov}\,(R_r, R_s)] = [E((R_r - b_r)(R_s - b_s))] = E[(\mathbf{R} - \mathbf{b})(\mathbf{R} - \mathbf{b})^t]$$

where $E(A)$ for a matrix A means the matrix $[E(A_{rs})]$. Thus the covariance matrix is

$$E[W\mathbf{R}'(W\mathbf{R}')^t] = WE(\mathbf{R}'\mathbf{R}'^t)W^t = WDW^t = K$$

since D is the covariance matrix of the R_j'.

The representation of Theorem 1 yields many useful properties of Gaussian vectors.

Theorem 3. *Let R be Gaussian with representation $\boldsymbol{R} = W\boldsymbol{R}' + \boldsymbol{b}$, W orthogonal, as in Theorem 1.*

1. If K is nonsingular, then the random variables $R_j^* = R_j - b_j$ are linearly independent; that is, if $\sum_{j=1}^n a_j R_j^* = 0$ with probability 1, then all $a_j = 0$. In this case R has a density given by

$$f(x) = (2\pi)^{-n/2}(\det K)^{-1/2} \exp\,[-\tfrac{1}{2}(\mathbf{x} - \mathbf{b})^t K^{-1}(\mathbf{x} - \mathbf{b})]$$

2. If K is singular, the R_j^* are linearly dependent. If, say, $\{R_1^*, \ldots, R_r^*\}$ is a maximal linearly independent subset of $\{R_1^*, \ldots, R_n^*\}$, then $(R_1, \ldots, R_r)$ has a density of the above form, with K replaced by K_r = the first r rows and columns of K. $R_{r+1}^*, \ldots, R_n^*$ can be expressed (with probability 1) as linear combinations of $R_1^*, \ldots, R_r^*$.

PROOF.

1. If K is nonsingular, all λ_j are > 0; hence R' has density

$$f'(y) = (2\pi)^{-n/2}(\lambda_1 \cdots \lambda_n)^{-1/2} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}\frac{y_k^2}{\lambda_k}\right]$$

$$= (2\pi)^{-n/2}(\det K)^{-1/2} \exp\,[-\tfrac{1}{2}\mathbf{y}^t D^{-1}\mathbf{y}]$$

The Jacobian of the transformation $\mathbf{x} = W\mathbf{y} + \mathbf{b}$ is $\det W = \pm 1$; hence $\boldsymbol{R}$ has density

$$\begin{aligned} f(\mathbf{x}) &= f'(W^t(\mathbf{x} - \mathbf{b})) \\ &= (2\pi)^{-n/2} (\det K)^{-1/2} \exp\left[-\tfrac{1}{2}(\mathbf{x} - \mathbf{b})^t W D^{-1} W^t (\mathbf{x} - \mathbf{b})\right] \end{aligned}$$

Since $K = WDW^t$, we have $K^{-1} = WD^{-1}W^t$, which yields the desired expression for the density.

Now if $\sum_{j=1}^n a_j R_j^* = 0$ with probability 1,

$$\begin{aligned} 0 = E\left[\left|\sum_{j=1}^n a_j R_j^*\right|^2\right] &= \sum_{r,s=1}^n a_r E(R_r^* R_s^*) a_s \\ &= \sum_{r,s=1}^n a_r K_{rs} a_s \end{aligned}$$

Since K is nonsingular, it is positive rather than merely nonnegative definite, and thus all $a_r = 0$.

2. If K is singular, then $\sum_{r,s=1}^n a_r K_{rs} a_s$ will be 0 for some $a_1, \ldots, a_n$, not all 0. (This follows since $\mathbf{u}^t K \mathbf{u} = \sum_{k=1}^n \lambda_k v_k^2$, where $\mathbf{v} = W^t\mathbf{u}$; if K is singular, then some λ_j is 0.) But by the analysis of case 1, $E[|\sum_{j=1}^n a_j R_j^*|^2] = 0$; hence $\sum_{j=1}^n a_j R_j^* = 0$ with probability 1, proving linear dependence. The remaining statements of 2 follow from 1.

REMARK. The result that K is singular iff the R_j^* are linearly dependent is true for arbitrary random variables with finite second moments, as the above argument shows.

▶ **Example 1.** Let (R_1, R_2) be Gaussian. Then

$$K = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

where $\sigma_1^2 = \operatorname{Var} R_1$, $\sigma_2^2 = \operatorname{Var} R_2$, $\sigma_{12} = \operatorname{Cov}(R_1, R_2)$. Also, $\det K = \sigma_1^2\sigma_2^2(1 - \rho_{12}^2)$, where ρ_{12} is the correlation coefficient between R_1 and R_2. Thus K is singular iff $|\rho_{12}| = 1$. In the nonsingular case we have

$$K^{-1} = (\sigma_1^2\sigma_2^2(1 - \rho_{12}^2))^{-1} \begin{bmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{bmatrix}$$

and the density is given by

$$\begin{aligned} f(x, y) = &\frac{1}{2\pi\sigma_1\sigma_2(1 - \rho_{12}^2)^{1/2}} \\ &\times \exp\left[-\frac{\sigma_2^2(x - a)^2 - 2\sigma_{12}(x - a)(y - b) + \sigma_1^2(y - b)^2}{2\sigma_1^2\sigma_2^2(1 - \rho_{12}^2)}\right] \end{aligned}$$

where $a = E(R_1)$, $b = E(R_2)$. The characteristic function of (R_1, R_2) is $M(u_1, u_2) = \exp[i(au_1 + bu_2)] \exp[-\frac{1}{2}(\sigma_1^2 u_1^2 + 2\sigma_{12}u_1u_2 + \sigma_2^2 u_2^2)]$. Notice that if $n = 1$, the multidimensional Gaussian distribution reduces to the ordinary Gaussian distribution. For in this case we have

$$K = [\sigma^2], \qquad M(u) = e^{ium}e^{-u^2\sigma^2/2}, \quad f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-m)^2/2\sigma^2}$$

where $m = E(R)$. ◀

Theorem 4. *If R_1 is a Gaussian n-vector and $\mathbf{R}_2 = A\mathbf{R}_1$, where A is an m by n matrix, then R_2 is a Gaussian m-vector.*

PROOF. Let $\mathbf{R}_1 = W\mathbf{R}' + \mathbf{b}$ as in Theorem 1. Then $\mathbf{R}_2 = AW\mathbf{R}' + A\mathbf{b}$, and hence R_2 is Gaussian by Theorem 1.

COROLLARY.

(a) If $R_1, \ldots, R_n$ are jointly Gaussian, so are $R_1, \ldots, R_m$, $m \leq n$.

(b) If $R_1, \ldots, R_n$ are jointly Gaussian, then $a_1R_1 + \cdots + a_nR_n$ is a Gaussian random variable.

PROOF. For (a) take $A = [I \;\; 0]$, where I is an m by m identity matrix. For (b) take $A = [a_1a_2 \ldots a_n]$.

Thus we see that if $R_1, \ldots, R_n$ are jointly Gaussian, then the R_i are (individually) Gaussian. The converse is not true, however. It is possible to find Gaussian random variables R_1, R_2 such that (R_1, R_2) is not Gaussian, and in addition $R_1 + R_2$ is not Gaussian.

For example, let R_1 be normal (0, 1) and define R_2 as follows. Let R_3 be independent of R_1, with $P\{R_3 = 0\} = P\{R_3 = 1\} = 1/2$. If $R_3 = 0$, let $R_2 = R_1$; if $R_3 = 1$, let $R_2 = -R_1$. Then $P\{R_2 \leq y\} = (1/2)P\{R_1 \leq y\} + (1/2)\,P\{-R_1 \leq y\} = P\{R_1 \leq y\}$, so that R_2 is normal (0, 1). But if $R_3 = 0$, then $R_1 + R_2 = 2R_1$, and if $R_3 = 1$, then $R_1 + R_2 = 0$. Therefore $P\{R_1 + R_2 = 0\} = 1/2$; hence $R_1 + R_2$ is not Gaussian. By corollary (b) to Theorem 4, (R_1, R_2) is not Gaussian.

Notice that if $R_1, \ldots, R_n$ are *independent* and each R_i is Gaussian, then the R_i are jointly Gaussian (with $K =$ the diagonal matrix of variances of the R_i).

Theorem 5. *If $R_1, \ldots, R_n$ are jointly Gaussian and uncorrelated, that is, if $K_{ij} = 0$ for $i \neq j$, they are independent.*

PROOF. Let $\sigma_j^2 = \text{Var } R_j$. We may assume all $\sigma_j^2 > 0$; if $\sigma_j^2 = 0$, then R_j is constant with probability 1 and may be deleted. Now $K = \text{diag } (\sigma_1^2, \ldots, \sigma_n^2)$; hence $K^{-1} = \text{diag } (1/\sigma_1^2, \ldots, 1/\sigma_n^2)$, and so, by Theorem 3, $R_1, \ldots, R_n$ have a joint density given by

$$f(x_1, \ldots, x_n) = (2\pi)^{-n/2}(\sigma_1 \cdots \sigma_n)^{-1} \exp\left[-\frac{1}{2}\sum_{j=1}^{n}\frac{(x_j - b_j)^2}{\sigma_j^2}\right]$$

Thus $R_1, \ldots, R_n$ are independent, with R_j normal (b_j, σ_j^2).

We now consider the following prediction problem. Let $R_1, \ldots, R_{n+1}$ be jointly Gaussian. We observe $R_1 = x_1, \ldots, R_n = x_n$ and then try to predict the value of R_{n+1}. If the predicted value is $\psi(x_1, \ldots, x_n)$ and the actual value is x_{n+1}, we assume a quadratic loss $(x_{n+1} - \psi(x_1, \ldots, x_n))^2$. In other words, we are trying to minimize the mean square difference between the true value and the predicted value of R_{n+1}. This is simply a problem of Bayes estimation with quadratic loss function, as considered in Section 8.3; in this case R_{n+1} plays the role of the state of nature and $(R_1, \ldots, R_n)$ the observable. It follows that the best estimate is

$$\psi(x_1, \ldots, x_n) = E(R_{n+1} \mid R_1 = x_1, \ldots, R_n = x_n)$$

We now show that in the jointly Gaussian case ψ is a linear function of $x_1, \ldots, x_n$. Thus the optimum predictor assumes a particularly simple form.

Say $\{R_1, \ldots, R_r\}$ is a maximal linearly independent subset of $\{R_1, \ldots, R_n\}$. If $R_1, \ldots, R_r, R_{n+1}$ are linearly dependent, there is nothing to prove; if $R_1, \ldots, R_r, R_{n+1}$ are linearly independent, we may replace $R_1, \ldots, R_n$ by $R_1, \ldots, R_r$ in the problem. Thus we may as well assume $R_1, \ldots, R_{n+1}$ linearly independent. Then $(R_1, \ldots, R_{n+1})$ has a density, and the conditional density of R_{n+1} given $R_1 = x_1, \ldots, R_n = x_n$ is

$$h(x_{n+1} \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_{n+1})}{\displaystyle\int_{-\infty}^{\infty} f(x_1, \ldots, x_{n+1})\, dx_{n+1}}$$

$$= \frac{(2\pi)^{-(n+1)/2}(\det K)^{-1/2} \exp\left[-\frac{1}{2}\displaystyle\sum_{r,s=1}^{n+1} x_r q_{rs} x_s\right]}{(2\pi)^{-(n+1)/2}(\det K)^{-1/2}\displaystyle\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\sum_{r,s=1}^{n+1} x_r q_{rs} x_s\right] dx_{n+1}}$$

where K is the covariance matrix of $R_1, \ldots, R_{n+1}$ and $Q = [q_{rs}] = K^{-1}$.

Thus

$$h(x_{n+1} \mid x_1, \ldots, x_n) = \frac{1}{B(x_1, \ldots, x_n)} \times \exp\left[-\frac{1}{2}\left(\sum_{r,s=1}^{n} x_r q_{rs} x_s + x_{n+1}\sum_{s=1}^{n} q_{n+1,s}x_s + x_{n+1}\sum_{r=1}^{n} x_r q_{r,n+1} + q_{n+1,n+1}x_{n+1}^2\right)\right]$$

$$= \frac{A(x_1, \ldots, x_n)}{B(x_1, \ldots, x_n)} \exp\left[-(Cx_{n+1}^2 + Dx_{n+1})\right]$$

where

$$C = \tfrac{1}{2}q_{n+1,n+1}, \qquad D = D(x_1, \ldots, x_n) = \sum_{r=1}^{n} q_{n+1,r}x_r$$

Therefore the conditional density can be expressed as

$$\frac{A}{B}\exp\left[\frac{D^2}{4C}\right]\exp\left[-C\left(x_{n+1} + \frac{D}{2C}\right)^2\right]$$

Thus, given $R_1 = x_1, \ldots, R_n = x_n$, R_{n+1} is normal with mean $-D/2C$ and variance $1/2C = 1/q_{n+1,n+1}$. Hence

$$E(R_{n+1} \mid R_1 = x_1, \ldots, R_n = x_n) = -\frac{1}{q_{n+1,n+1}}\sum_{r=1}^{n} q_{n+1,r}x_r$$