Preface

The area of detection and estimation theory that we shall study in this book represents a combination of the classical techniques of statistical inference and the random process characterization of communication, radar, sonar, and other modern data processing systems. The two major areas of statistical inference are decision theory and estimation theory. In the first case we observe an output that has a random character and decide which of two possible causes produced it. This type of problem was studied in the middle of the eighteenth century by Thomas Bayes [1]. In the estimation theory case the output is related to the value of some parameter of interest, and we try to estimate the value of this parameter. Work in this area was published by Legendre [2] and Gauss [3] in the early nineteenth century. Significant contributions to the classical theory that we use as background were developed by Fisher [4] and Neyman and Pearson [5] more than 30 years ago. In 1941 and 1942 Kolmogoroff [6] and Wiener [7] applied statistical techniques to the solution of the optimum linear filtering problem. Since that time the application of statistical techniques to the synthesis and analysis of all types of systems has grown rapidly. The application of these techniques and the resulting implications are the subject of this book.

This book and the subsequent volume, Detection, Estimation, and Modulation Theory, Part II, are based on notes prepared for a course entitled "Detection, Estimation, and Modulation Theory," which is taught as a second-level graduate course at M.I.T. My original interest in the material grew out of my research activities in the area of analog modulation theory. A preliminary version of the material that deals with modulation theory was used as a text for a summer course presented at M.I.T. in 1964. It turned out that our viewpoint on modulation theory could best be understood by an audience with a clear understanding of modern detection and estimation theory. At that time there was no suitable text available to cover the material of interest and emphasize the points that I felt were
important, so I started writing notes. It was clear that in order to present the material to graduate students in a reasonable amount of time it would be necessary to develop a unified presentation of the three topics: detection, estimation, and modulation theory, and exploit the fundamental ideas that connected them. As the development proceeded, it grew in size until the material that was originally intended to be background for modulation theory occupies the entire contents of this book. The original material on modulation theory starts at the beginning of the second book. Collectively, the two books provide a unified coverage of the three topics and their application to many important physical problems.

For the last three years I have presented successively revised versions of the material in my course. The audience consists typically of 40 to 50 students who have completed a graduate course in random processes which covered most of the material in Davenport and Root [8]. In general, they have a good understanding of random process theory and a fair amount of practice with the routine manipulation required to solve problems. In addition, many of them are interested in doing research in this general area or closely related areas. This interest provides a great deal of motivation which I exploit by requiring them to develop many of the important ideas as problems. It is for this audience that the book is primarily intended. The appendix contains a detailed outline of the course.

On the other hand, many practicing engineers deal with systems that have been or should have been designed and analyzed with the techniques developed in this book. I have attempted to make the book useful to them. An earlier version was used successfully as a text for an in-plant course for graduate engineers.

From the standpoint of specific background little advanced material is required. A knowledge of elementary probability theory and second moment characterization of random processes is assumed. Some familiarity with matrix theory and linear algebra is helpful but certainly not necessary. The level of mathematical rigor is low, although in most sections the results could be rigorously proved by simply being more careful in our derivations. We have adopted this approach in order not to obscure the important ideas with a lot of detail and to make the material readable for the kind of engineering audience that will find it useful. Fortunately, in almost all cases we can verify that our answers are intuitively logical. It is worthwhile to observe that this ability to check our answers intuitively would be necessary even if our derivations were rigorous, because our ultimate objective is to obtain an answer that corresponds to some physical system of interest. It is easy to find physical problems in which a plausible mathematical model and correct mathematics lead to an unrealistic answer for the original problem.

We have several idiosyncrasies that it might be appropriate to mention. In general, we look at a problem in a fair amount of detail. Many times we look at the same problem in several different ways in order to gain a better understanding of the meaning of the result. Teaching students a number of ways of doing things helps them to be more flexible in their approach to new problems. A second feature is the necessity for the reader to solve problems to understand the material fully. Throughout the course and the book we emphasize the development of an ability to work problems. At the end of each chapter are problems that range from routine manipulations to significant extensions of the material in the text. In many cases they are equivalent to journal articles currently being published. Only by working a fair number of them is it possible to appreciate the significance and generality of the results. Solutions for an individual problem will be supplied on request, and a book containing solutions to about one third of the problems is available to faculty members teaching the course. We are continually generating new problems in conjunction with the course and will send them to anyone who is using the book as a course text. A third issue is the abundance of block diagrams, outlines, and pictures. The diagrams are included because most engineers (including myself) are more at home with these items than with the corresponding equations.

One problem always encountered is the amount of notation needed to cover the large range of subjects. We have tried to choose the notation in a logical manner and to make it mnemonic. All the notation is summarized in the glossary at the end of the book. We have tried to make our list of references as complete as possible and to acknowledge any ideas due to other people.

A number of people have contributed in many ways and it is a pleasure to acknowledge them. Professors W. B. Davenport and W. M. Siebert have provided continual encouragement and technical comments on the various chapters. Professors Estil Hoversten and Donald Snyder of the M.I.T. faculty and Lewis Collins, Arthur Baggeroer, and Michael Austin, three of my doctoral students, have carefully read and criticized the various chapters. Their suggestions have improved the manuscript appreciably. In addition, Baggeroer and Collins contributed a number of the problems in the various chapters and Baggeroer did the programming necessary for many of the graphical results. Lt. David Wright read and criticized Chapter 2. L. A. Frasco and H. D. Goldfein, two of my teaching assistants, worked all of the problems in the book. Dr. Howard Yudkin of Lincoln Laboratory read the entire manuscript and offered a number of important criticisms. In addition, various graduate students taking the course have made suggestions which have been incorporated. Most of the final draft was typed by Miss Aina Sils. Her patience with the innumerable changes is
3 *Representations of Random Processes*  

3.1 Introduction  
3.2 Deterministic Functions: Orthogonal Representations  
3.3 Random Process Characterization  
3.4 Homogeneous Integral Equations and Eigenfunctions  
3.5 Periodic Processes  
3.6 Infinite Time Interval: Spectral Decomposition  
3.7 Vector Random Processes  
3.8 Summary  
3.9 Problems  
References  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>166</td>
</tr>
</tbody>
</table>
| Signals with Unwanted Parameters: The Composite Hypothesis Problem  
| 4.5 | 174 |
| Multiple Channels  
   Formulation. Application. |
| 4.6 | 186 |
| Multiple Parameter Estimation  
   Additive White Gaussian Noise Channel. Extensions. |
| 4.7 | 209 |
| Summary and Omissions  
   Summary. Topics Omitted. |
| 4.8 | 212 |
| Problems  
   References |
| 4.9 | 220 |
| 4.10 | 224 |
| 4.11 | 226 |
| 5 | 237 |

5 *Estimation of Continuous Waveforms*  

5.1 Introduction  
5.2 Derivation of Estimator Equations  
   No-Memory Modulation Systems. Modulation Systems with Memory.  
5.3 A Lower Bound on the Mean-Square Estimation Error  
5.4 Multidimensional Waveform Estimation  
5.5 Nonrandom Waveform Estimation  
5.6 Summary  
5.7 Problems  
References  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4</td>
<td>333</td>
</tr>
</tbody>
</table>
| Signals with Unwanted Parameters: The Composite Hypothesis Problem  
| 6.5 | 366 |
| Multiple Channels  
   Formulation. Application. |
| 6.6 | 370 |
| Multiple Parameter Estimation  
   Additive White Gaussian Noise Channel. Extensions. |
| 6.7 | 374 |
| Summary and Omissions  
   Summary. Topics Omitted. |
| 6.8 | 377 |
| Problems  
   References |
| 6.9 | 418 |
| 5.1 | 423 |
| Estimation of Continuous Waveforms |
| 5.2 | 426 |
| Derivation of Estimator Equations  
   No-Memory Modulation Systems. Modulation Systems with Memory. |
| 5.3 | 437 |
| A Lower Bound on the Mean-Square Estimation Error |
| 5.4 | 446 |
| Multidimensional Waveform Estimation  
| 5.5 | 456 |
| Nonrandom Waveform Estimation |
| 5.6 | 459 |
| Summary |
| 5.7 | 460 |
| Problems  
   References |
| 6.1 | 467 |
| 6.2 | 468 |
| Properties of Optimum Processors  
   Realizable Linear Filters: Stationary Processes, Infinite Past: Wiener Filters |
| 6.3 | 481 |
Introduction

In these two books, we shall study three areas of statistical theory, which we have labeled detection theory, estimation theory, and modulation theory. The goal is to develop these theories in a common mathematical framework and to demonstrate how they can be used to solve a wealth of practical problems in many diverse physical situations.

In this chapter we present three outlines of the material. The first is a topical outline in which we develop a qualitative understanding of the three areas by examining some typical problems of interest. The second is a logical outline in which we explore the various methods of attacking the problems. The third is a chronological outline in which we explain the structure of the books.

1.1 TOPICAL OUTLINE

An easy way to explain what is meant by detection theory is to examine several physical situations that lead to detection theory problems.

A simple digital communication system is shown in Fig. 1.1. The source puts out a binary digit every $T$ seconds. Our object is to transmit this sequence of digits to some other location. The channel available for transmitting the sequence depends on the particular situation. Typically, it could be a telephone line, a radio link, or an acoustical channel. For
presents of illustration, we shall consider a radio link. In order to transmit
the information, we must put it into a form suitable for propagating over
the channel. A straightforward method would be to build a device that
generates a sine wave,

\[ s_1(t) = \sin \omega_1 t, \]  

for \( T \) seconds if the source generated a "one" in the preceding interval, and a sine wave of a different frequency,

\[ s_0(t) = \sin \omega_0 t, \]

for \( T \) seconds if the source generated a "zero" in the preceding interval. The frequencies are chosen so that the signals \( s_0(t) \) and \( s_1(t) \) will propagate over the particular radio link of concern. The output of the device is fed into an antenna and transmitted over the channel. Typical source and transmitted signal sequences are shown in Fig. 1.2. In the simplest kind of channel the signal sequence arrives at the receiving antenna attenuated but essentially undistorted. To process the received signal we pass it through the antenna and some stages of rf-amplification, in the course of which a thermal noise \( n(t) \) is added to the message sequence. Thus in any \( T \)-second interval we have available a waveform \( r(t) \) in which

\[ r(t) = s_1(t) + n(t), \quad 0 \leq t \leq T, \]  

if \( s_1(t) \) was transmitted, and

\[ r(t) = s_0(t) + n(t), \quad 0 \leq t \leq T, \]

if \( s_0(t) \) was transmitted. We are now faced with the problem of deciding which of the two possible signals was transmitted. We label the device that does this a decision device. It is simply a processor that observes \( r(t) \) and guesses whether \( s_1(t) \) or \( s_0(t) \) was sent according to some set of rules. This is equivalent to guessing what the source output was in the preceding interval. We refer to designing and evaluating the processor as a detection

\[ \begin{array}{c}
\text{Transmitter} \\
\text{Source output} \\
\text{Transmitted sequence}
\end{array} \]

\[ \begin{array}{cccccc}
1 & 0 & 0 & 1 & \\
T & T & T & T & \\
\text{Fig. 1.2} & \text{Typical sequences.}
\end{array} \]

\[ \begin{array}{ccccccc}
sin \omega_1 t & \sin \omega_1 t & \sin \omega_0 t & \sin \omega_0 t \\
\text{Transmitter} & \text{Source output} & \text{Transmitted sequence}
\end{array} \]

\[ \begin{array}{ccccccc}
\text{Fig. 1.3} & \text{Sequence with phase shifts.}
\end{array} \]

theory problem. In this particular case the only possible source of error in making a decision is the additive noise. If it were not present, the input would be completely known and we could make decisions without errors. We denote this type of problem as the known signal in noise problem. It corresponds to the lowest level (i.e., simplest) of the detection problems of interest.

An example of the next level of detection problem is shown in Fig. 1.3. The oscillators used to generate \( s_1(t) \) and \( s_0(t) \) in the preceding example have a phase drift. Therefore in a particular \( T \)-second interval the received signal corresponding to a "one" is

\[ r(t) = \sin (\omega_1 t + \theta_1) + n(t), \quad 0 \leq t \leq T, \]

and the received signal corresponding to a "zero" is

\[ r(t) = \sin (\omega_0 t + \theta_0) + n(t), \quad 0 \leq t \leq T, \]

where \( \theta_0 \) and \( \theta_1 \) are unknown constant phase angles. Thus even in the absence of noise the input waveform is not completely known. In a practical system the receiver may include auxiliary equipment to measure the oscillator phase. If the phase varies slowly enough, we shall see that essentially perfect measurement is possible. If this is true, the problem is the same as above. However, if the measurement is not perfect, we must incorporate the signal uncertainty in our model.

A corresponding problem arises in the radar and sonar areas. A conventional radar transmits a pulse at some frequency \( \omega_c \) with a rectangular envelope:

\[ s(t) = \sin \omega_c t, \quad 0 \leq t \leq T. \]

If a target is present, the pulse is reflected. Even the simplest target will introduce an attenuation and phase shift in the transmitted signal. Thus the signal available for processing in the interval of interest is

\[ r(t) = V_r \sin (\omega_c (t - \tau) + \theta_r) + n(t), \quad \tau \leq t \leq \tau + T, \]

\[ = n(t), \quad 0 \leq t < \tau, \tau + T < t < \infty. \]
In our examination of representative examples we have seen that detection theory problems are characterized by the fact that we must decide which of several alternatives is true. There were only two alternatives in the examples cited; therefore we refer to them as binary detection problems. Later we will encounter problems in which there are \( M \) alternatives available (the \( M \)-ary detection problem). Our hierarchy of detection problems is presented graphically in Fig. 1.4.

There is a parallel set of problems in the estimation theory area. A simple example is given in Fig. 1.5, in which the source puts out an analog message \( a(t) \) (Fig. 1.5a). To transmit the message we first sample it every \( T \) seconds. Then, every \( T \) seconds we transmit a signal that contains

<table>
<thead>
<tr>
<th>Detection theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1. Known signals in noise</td>
</tr>
<tr>
<td>1. Synchronous digital communication</td>
</tr>
<tr>
<td>2. Pattern recognition problems</td>
</tr>
<tr>
<td>Level 2. Signals with unknown parameters in noise</td>
</tr>
<tr>
<td>1. Conventional pulsed radar or sonar, target detection</td>
</tr>
<tr>
<td>2. Target classification (orientation of target unknown)</td>
</tr>
<tr>
<td>3. Digital communication systems without phase reference</td>
</tr>
<tr>
<td>4. Digital communication over slowly-fading channels</td>
</tr>
<tr>
<td>Level 3. Random signals in noise</td>
</tr>
<tr>
<td>1. Digital communication over scatter link, orbiting dipole channel, or chaff link</td>
</tr>
<tr>
<td>2. Passive sonar</td>
</tr>
<tr>
<td>3. Seismic detection system</td>
</tr>
<tr>
<td>4. Radio astronomy (detection of noise sources)</td>
</tr>
</tbody>
</table>

Fig. 1.4 Detection theory hierarchy.
a parameter which is uniquely related to the last sample value. In Fig. 1.5b the signal is a sinusoid whose amplitude depends on the last sample. Thus, if the sample at time \( nT \) is \( A_n \), the signal in the interval \( [nT, (n + 1)T] \) is

\[
s(t, A_n) = A_n \sin(\omega_n t), \quad nT \leq t \leq (n + 1)T.
\]

A system of this type is called a pulse amplitude modulation (PAM) system. In Fig. 1.5c the signal is a sinusoid whose frequency in the interval differs from the reference frequency \( \omega_c \) by an amount proportional to the preceding sample value,

\[
s(t, A_n) = \sin(\omega_c t + A_n t), \quad nT \leq t \leq (n + 1)T.
\]

A system of this type is called a pulse frequency modulation (PFM) system. Once again there is additive noise. The received waveform, given that \( A_n \) was the sample value, is

\[
r(t) = s(t, A_n) + n(t), \quad nT \leq t \leq (n + 1)T.
\]

During each interval the receiver tries to estimate \( A_n \). We denote these estimates as \( \hat{A}_n \). Over a period of time we obtain a sequence of estimates, as shown in Fig. 1.5d, which is passed into a device whose output is an estimate of the original message \( a(t) \). If \( a(t) \) is a band-limited signal, the device is just an ideal low-pass filter. For other cases it is more involved.

If, however, the parameters in this example were known and the noise was absent, the received signal would be completely known. We refer to problems in this category as known signal in noise problems. If we assume that the mapping from \( A_n \) to \( s(t, A_n) \) in the transmitter has an inverse, we see that if the noise were not present we could determine \( A_n \) unambiguously. (Clearly, if we were allowed to design the transmitter, we should always choose a mapping with an inverse.) The known signal in noise problem is the first level of the estimation problem hierarchy.

Returning to the area of radar, we consider a somewhat different problem. We assume that we know a target is present but do not know its range or velocity. Then the received signal is

\[
r(t) = V_r \sin(\omega_d + \omega_n) t + n(t), \quad 0 \leq t < \tau + T.
\]

where \( \omega_d \) denotes a Doppler shift caused by the target’s motion. We want to estimate \( \tau \) and \( \omega_d \). Now, even if the noise were absent and \( \tau \) and \( \omega_d \) were known, the signal would still contain the unknown parameters \( V_r \) and \( \theta \). This is a typical second-level estimation problem. As in detection theory, we refer to problems in this category as signal with unknown parameters in noise problems.

At the third level the signal component is a random process whose statistical characteristics contain parameters we want to estimate. The received signal is of the form

\[
r(t) = s_0(t, A) + n(t),
\]

where \( s_0(t, A) \) is a sample function from a random process. In a simple case it might be a stationary process with the narrow-band spectrum shown in Fig. 1.6. The shape of the spectrum is known but the center frequency
is not. The receiver must observe $r(t)$ and, using the statistical properties of $s(t, A)$ and $n(t)$, estimate the value of $A$. This particular example could arise in either radio astronomy or passive sonar. The general class of problem in which the signal containing the parameters is a sample function from a random process is referred to as the random signal in noise problem. The hierarchy of estimation theory problems is shown in Fig. 1.7.

We note that there appears to be considerable parallelism in the detection and estimation theory problems. We shall frequently exploit this parallelism to reduce the work, but there is a basic difference that should be emphasized. In binary detection the receiver is either “right” or “wrong.” In the estimation of a continuous parameter the receiver will seldom be exactly right, but it can try to be close most of the time. This difference will be reflected in the manner in which we judge system performance.

The third area of interest is frequently referred to as modulation theory. We shall see shortly that this term is too narrow for the actual problems. Once again a simple example is useful. In Fig. 1.8 we show an analog message source whose output might typically be music or speech. To convey the message over the channel, we transform it by using a modulation scheme to get it into a form suitable for propagation. The transmitted signal is a continuous waveform that depends on $a(t)$ in some deterministic manner. In Fig. 1.8 it is an amplitude modulated waveform:

$$s(t, a(t)) = [1 + ma(t)] \sin(\omega_c t).$$

(19)

(This is conventional double-sideband AM with modulation index $m$.) In Fig. 1.8c the transmitted signal is a frequency modulated (FM) waveform:

$$s(t, a(t)) = \sin\left[ \omega_c t + \int_{-\infty}^{t} a(u) \, du \right].$$

(20)

When noise is added the received signal is

$$r(t) = s(t, a(t)) + n(t).$$

(21)

Now the receiver must observe $r(t)$ and put out a continuous estimate of the message $a(t)$, as shown in Fig. 1.8. This particular example is a first-level modulation problem, for if $n(t)$ were absent and $a(t)$ were known the received signal would be completely known. Once again we describe it as a known signal in noise problem.

Another type of physical situation in which we want to estimate a continuous function is shown in Fig. 1.9. The channel is a time-invariant linear system whose impulse response $h(\tau)$ is unknown. To estimate the impulse response we transmit a known signal $x(t)$. The received signal is

$$r(t) = \int_{0}^{t} h(\tau) x(t - \tau) \, d\tau + n(t).$$

(22)
The receiver observes $r(t)$ and tries to estimate $h(t)$. This particular example could best be described as a continuous estimation problem. Many other problems of interest in which we shall try to estimate a continuous waveform will be encountered. For convenience, we shall use the term modulation theory for this category, even though the term continuous waveform estimation might be more descriptive.

The other levels of the modulation theory problem follow by direct analogy. In the amplitude modulation system shown in Fig. 1.8b the receiver frequently does not know the phase angle of the carrier. In this case a suitable model is

$$r(t) = (1 + ma(t)) \sin (\omega_c t + \theta) + n(t), \quad (23)$$

---

**Fig. 1.8** A modulation theory example: (a) analog transmission system; (b) amplitude modulated signal; (c) frequency modulated signal; (d) demodulator.

**Modulation Theory (Continuous waveform estimation)**

<table>
<thead>
<tr>
<th>1. Known signals in noise</th>
<th>1. Conventional communication systems such as AM (DSB-AM, SSB), FM, and PM with phase synchronization</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Optimum filter theory</td>
<td>2. Estimation of channel characteristics when phase of input signal is unknown</td>
</tr>
<tr>
<td>4. Channel measurement</td>
<td>4. Estimation of plant characteristics</td>
</tr>
<tr>
<td>5. Orbital estimation for satellites</td>
<td>5. Signal estimation in seismic and sonar classification systems</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Signals with unknown parameters in noise</th>
<th>1. Conventional communication systems without phase synchronization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. Estimation of channel characteristics when phase of input signal is unknown</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3. Random signals in noise</th>
<th>1. Analog communication over randomly varying channels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. Estimation of statistics of time-varying processes</td>
</tr>
<tr>
<td></td>
<td>3. Estimation of plant characteristics</td>
</tr>
</tbody>
</table>

**Fig. 1.9** Channel measurement.

**Fig. 1.10** Modulation theory hierarchy.
where \( \theta \) is an unknown parameter. This is an example of a signal with unknown parameter problem in the modulation theory area.

A simple example of a third-level problem (random signal in noise) is one in which we transmit a frequency-modulated signal over a radio link whose gain and phase characteristics are time-varying. We shall find that if we transmit the signal in (20) over this channel the received waveform will be

\[
    r(t) = V(t) \sin \left[ \omega_c t + \int_{-\infty}^t a(u) \, du + \theta(t) \right] + n(t),
\]

where \( V(t) \) and \( \theta(t) \) are sample functions from random processes. Thus, even if \( a(u) \) were known and the noise \( n(t) \) were absent, the received signal would still be a random process. An overall outline of the problems of interest to us appears in Fig. 1.10. Additional examples included in the table to indicate the breadth of the problems that fit into the outline are discussed in more detail in the text.

Now that we have outlined the areas of interest it is appropriate to determine how to go about solving them.

1.2 POSSIBLE APPROACHES

From the examples we have discussed it is obvious that an inherent feature of all the problems is randomness of source, channel, or noise (often all three). Thus our approach must be statistical in nature. Even assuming that we are using a statistical model, there are many different ways to approach the problem. We can divide the possible approaches into two categories, which we denote as “structured” and “nonstructured.”

Some simple examples will illustrate what we mean by a structured approach.

**Example 1.** The input to a linear time-invariant system is \( r(t) \):

\[
    r(t) = s(t) + w(t) \quad 0 \leq t \leq T, \quad = 0, \quad \text{elsewhere,}
\]

The impulse response of the system is \( h(\tau) \). The signal \( s(t) \) is a known function with energy \( E_s \).

\[
    E_s = \int_0^T s^2(t) \, dt,
\]

and \( w(t) \) is a sample function from a zero-mean random process with a covariance function:

\[
    K_w(t, u) = \frac{N_0}{2} \delta(t - u).
\]

We are concerned with the output of the system at time \( T \). The output due to the signal is a deterministic quantity:

\[
    s_d(T) = \int_0^T h(\tau) s(T - \tau) \, d\tau.
\]

The output due to the noise is a random variable:

\[
    n_d(T) = \int_0^T h(\tau) n(T - \tau) \, d\tau.
\]

We can define the output signal-to-noise ratio at time \( T \) as

\[
    \frac{S}{N} = \frac{s^2_d(T)}{E[n^2_d(T)]},
\]

where \( E(\cdot) \) denotes expectation.

Substituting (28) and (29) into (30), we obtain

\[
    \frac{S}{N} = \frac{\left[ \int_0^T h(\tau) s(T - \tau) \, d\tau \right]^2}{E \left[ \int_0^T h(\tau) n(T - \tau) n(T - u) \, d\tau \, du \right]}.
\]

By bringing the expectation inside the integral, using (27), and performing the integration with respect to \( u \), we have

\[
    \frac{S}{N} = \frac{\left[ \int_0^T h(\tau) s(T - \tau) \, d\tau \right]^2}{N_0/2 \left[ \int_0^T h^2(\tau) \, d\tau \right]}.
\]

The problem of interest is to choose \( h(\tau) \) to maximize the signal-to-noise ratio. The solution follows easily, but it is not important for our present discussion. (See Problem 3.3.1.)

This example illustrates the three essential features of the structured approach to a statistical optimization problem:

**Structure.** The processor was required to be a linear time-invariant filter. We wanted to choose the best system in this class. Systems that were not in this class (e.g., nonlinear or time-varying) were not allowed.

**Criterion.** In this case we wanted to maximize a quantity that we called the signal-to-noise ratio.

**Information.** To write the expression for \( S/N \) we had to know the signal shape and the covariance function of the noise process.

If we knew more about the process (e.g., its first-order probability density), we could not use it, and if we knew less, we could not solve the problem. Clearly, if we changed the criterion, the information required might be different. For example, to maximize

\[
    x = -\frac{s^2_d(T)}{E[n^2_d(T)]},
\]

the covariance function of the noise process would not be adequate. Alternatively, if we changed the structure, the information required might...
change. Thus the three ideas of structure, criterion, and information are closely related. It is important to emphasize that the structured approach does not imply a linear system, as illustrated by Example 2.

**Example 2.** The input to the nonlinear no-memory device shown in Fig. 1.11 is \( r(t) \), where

\[
    r(t) = x(t) + n(t), \quad -\infty < t < \infty. \tag{34}
\]

At any time \( t \), \( x(t) \) is the value of a random variable \( s \) with known probability density \( p_s(S) \). Similarly, \( n(t) \) is the value of a statistically independent random variable \( n \) with known density \( p_n(N) \). The output of the device is \( y(t) \), where

\[
    y(t) = a_0 + a_1[r(t)] + a_2[r(t)]^2 \tag{35}
\]

is a quadratic no-memory function of \( r(t) \). The adjective no-memory emphasizes that the value of \( y(t) \) depends only on \( r(t) \). We want to choose the coefficients \( a_0, a_1, \) and \( a_2 \) so that \( s(t) \) is the minimum mean-square error estimate of \( x(t) \). The mean-square error is

\[
    \xi(t) = E[(y(t) - s(t))^2] = E[(a_0 + a_1[r(t)] + a_2[r(t)]^2) - s(t))^2] \tag{36}
\]

and \( a_0, a_1, \) and \( a_2 \) are chosen to minimize \( \xi(t) \). The solution to this particular problem is given in Chapter 3.

The technique for solving structured problems is conceptually straightforward. We allow the structure to vary within the allowed class and choose the particular system that maximizes (or minimizes) the criterion of interest.

An obvious advantage to the structured approach is that it usually requires only a partial characterization of the processes. This is important because, in practice, we must measure or calculate the process properties needed.

An obvious disadvantage is that it is often impossible to tell if the structure chosen is correct. In Example 1 a simple nonlinear system might be far superior to the best linear system. Similarly, in Example 2 some other nonlinear system might be far superior to the quadratic system. Once a class of structure is chosen we are committed. A number of trivial examples demonstrate the effect of choosing the wrong structure. We shall encounter an important practical example when we study frequency modulation in Chapter II-2.

At first glance it appears that one way to get around the problem of choosing the proper structure is to let the structure be an arbitrary nonlinear time-varying system. In other words, the class of structure is chosen to be so large that every possible system will be included in it. The difficulty is that there is no convenient tool, such as the convolution integral, to express the output of a nonlinear system in terms of its input. This means that there is no convenient way to investigate all possible systems by using a structured approach.

The alternative to the structured approach is a nonstructured approach. Here we refuse to make any a priori guesses about what structure the processor should have. We establish a criterion, solve the problem, and implement whatever processing procedure is indicated.

A simple example of the nonstructured approach can be obtained by modifying Example 2. Instead of assigning characteristics to the device, we denote the estimate by \( y(t) \). Letting

\[
    \xi(t) = E[(y(t) - s(t))^2], \tag{37}
\]

we solve for the \( y(t) that is obtained from \( r(t) \) in any manner to minimize \( \xi \). The obvious advantage is that if we can solve the problem we know that our answer, is with respect to the chosen criterion, the best processor of all possible processors. The obvious disadvantage is that we must completely characterize all the signals, channels, and noises that enter into the problem. Fortunately, it turns out that there are a large number of problems of practical importance in which this complete characterization is possible. Throughout both books we shall emphasize the nonstructured approach.

Our discussion up to this point has developed the topical and logical basis of these books. We now discuss the actual organization.

### 1.3 ORGANIZATION

The material covered in this book and Volume II can be divided into five parts. The first can be labeled Background and consists of Chapters 2 and 3. In Chapter 2 we develop in detail a topic that we call Classical Detection and Estimation Theory. Here we deal with problems in which...
the observations are sets of random variables instead of random waveforms. The theory needed to solve problems of this type has been studied by statisticians for many years. We therefore use the adjective classical to describe it. The purpose of the chapter is twofold: first, to derive all the basic statistical results we need in the remainder of the chapters; second, to provide a general background in detection and estimation theory that can be extended into various areas that we do not discuss in detail.

To accomplish the second purpose we keep the discussion as general as possible. We consider in detail the binary and M-ary hypothesis testing problem, the problem of estimating random and nonrandom variables, and the composite hypothesis testing problem. Two more specialized topics, the general Gaussian problem and performance bounds on binary tests, are developed as background for specific problems we shall encounter later.

The next step is to bridge the gap between the classical case and the waveform problems discussed in Section 1.1. Chapter 3 develops the necessary techniques. The key to the transition is a suitable method for characterizing random processes. When the observation interval is finite, the most useful characterization is by a series expansion of the random process which is a generalization of the conventional Fourier series. When the observation interval is infinite, a transform characterization, which is a generalization of the usual Fourier transform, is needed. In the process of developing these characterizations, we encounter integral equations and we digress briefly to develop methods of solution. Just as in Chapter 2, our discussion is general and provides background for other areas of application.

With these two chapters in the first part as background, we are prepared to work our way through the hierarchy of problems outlined in Figs. 1.4, 1.7, and 1.10. The second part of the book (Chapter 4) can be labeled Elementary Detection and Estimation Theory. Here we develop the first two levels described in Section 1.1. (This material corresponds to the upper two levels in Figs. 1.4 and 1.7.) We begin by looking at the simple binary digital communication system described in Fig. 1.1 and then proceed to more complicated problems in the communications, radar, and sonar area involving M-ary communication, random phase channels, random amplitude and phase channels, and colored noise interference. By exploiting the parallel nature of the estimation problem, results are obtained easily for the estimation problem outlined in Fig. 1.5 and other more complex systems. The extension of the results to include the multiple channel (e.g., frequency diversity systems or arrays) and multiple parameter (e.g., range and Doppler) problems completes our discussion. The results in this chapter are fundamental to the understanding of modern communication and radar/sonar systems.

The third part, which can be labeled Modulation Theory or Continuous Estimation Theory, consists of Chapters 5 and 6 and Chapter 2 of Volume II. In Chapter 5 we formulate a quantitative model for the first two levels of the continuous waveform estimation problem and derive a set of integral equations whose solution is the optimum estimate of the message. We also derive equations that give bounds on the performance of the estimators. In order to study solution techniques, we divide the estimation problem into two categories, linear and nonlinear.

In Chapter 6 we study linear estimation problems in detail. In the first section of the chapter we discuss the relationships between various criteria, process characteristics, and the structure of the processor. In the next section we discuss the special case in which the processes are stationary and the infinite past is available. This case, the Wiener problem, leads to straightforward solution techniques. The original work of Wiener is extended to obtain some important closed-form error expressions. In the next section we discuss the case in which the processes can be characterized by using state-variable techniques. This case, the Kalman-Bucy problem, enables us to deal with nonstationary, finite-interval problems and adds considerable insight to the results of the preceding section.

The material in Chapters 1 through 6 has two characteristics:

1. In almost all cases we can obtain explicit, exact solutions to the problems that we formulate.

2. Most of the topics discussed are of such fundamental interest that everyone concerned with the statistical design of communication, radar, or sonar systems should be familiar with them.

As soon as we try to solve the nonlinear estimation problem, we see a sharp departure. To obtain useful results we must resort to approximate solution techniques. To decide what approximations are valid, however, we must consider specific nonlinear modulation systems. Thus the precise quantitative results are only applicable to the specific system. In view of this departure, we pause briefly in our logical development and summarize our results in Chapter 7.

After a brief introduction we return to the nonlinear modulation problem in Chapter 2 of Volume II and consider angle modulation systems in great detail. After an approximation to the optimum processor is developed, its performance and possible design modifications are analyzed both theoretically and experimentally. More advanced techniques from Markov process theory and information theory are used to obtain significant results.

In the fourth part we revisit the problems of detection, estimation, and modulation theory at the third level of the hierarchy described in Section 1.1. Looking at the bottom boxes in Figs. 1.4, 1.7, and 1.10, we see that this is the Random Signals in Noise problem. Chapter II-3 studies it in
2

Classical Detection and Estimation Theory

2.1 INTRODUCTION

In this chapter we develop in detail the basic ideas of classical detection and estimation theory. The first step is to define the various terms.

The basic components of a simple decision-theory problem are shown in Fig. 2.1. The first is a source that generates an output. In the simplest case this output is one of two choices. We refer to them as hypotheses and label them \( H_0 \) and \( H_1 \) in the two-choice case. More generally, the output might be one of \( M \) hypotheses, which we label \( H_0, H_1, \ldots, H_{M-1} \). Some typical source mechanisms are the following:

1. A digital communication system transmits information by sending ones and zeros. When “one” is sent, we call it \( H_1 \), and when “zero” is sent, we call it \( H_0 \).
2. In a radar system we look at a particular range and azimuth and try
to decide whether a target is present; \( H_1 \) corresponds to the presence of a target and \( H_0 \) corresponds to no target.

3. In a medical diagnosis problem we examine an electrocardiogram. Here \( H_1 \) could correspond to the patient having had a heart attack and \( H_0 \) to the absence of one.

4. In a speaker classification problem we know the speaker is German, British, or American and either male or female. There are six possible hypotheses.

In the cases of interest to us we do not know which hypothesis is true. The second component of the problem is a **probabilistic transition mechanism**; the third is an observation space. The transition mechanism can be viewed as a device that knows which hypothesis is true. Based on this knowledge, it generates a point in the observation space according to some probability law.

A simple example is given in Fig. 2.2. When \( H_1 \) is true, the source generates \(+1\). When \( H_0 \) is true, the source generates \(-1\).

An independent discrete random variable \( n \) whose probability density is shown in Fig. 2.2b is added to the source output. The sum of the source output and \( n \) is the observed variable \( r \).

Under the two hypotheses, we have

\[
H_1: r = 1 + n, \\
H_0: r = -1 + n
\]

The probability densities of \( r \) on the two hypotheses are shown in Fig. 2.2b. The observation space is one-dimensional, for any output can be plotted on a line.

A related example is shown in Fig. 2.3a in which the source generates two numbers in sequence. A random variable \( n_1 \) is added to the first number and an independent random variable \( n_2 \) is added to the second.

Thus

\[
H_1: r_1 = 1 + n_1, \\
r_2 = 1 + n_2, \\
H_0: r_1 = -1 + n_1, \\
r_2 = -1 + n_2.
\]

The joint probability density of \( r_1 \) and \( r_2 \) when \( H_1 \) is true is shown in Fig. 2.3b. The observation space is two-dimensional and any observation can be represented as a point in a plane.

In this chapter we confine our discussion to problems in which the observation space is finite-dimensional. In other words, the observations consist of a set of \( N \) numbers and can be represented as a point in an \( N \)-dimensional space. This is the class of problem that statisticians have treated for many years. For this reason we refer to it as the classical decision problem.

The fourth component of the detection problem is a decision rule. After observing the outcome in the observation space we shall guess which hypothesis was true, and to accomplish this we develop a decision rule that assigns each point to one of the hypotheses. Suitable choices for decision rules will depend on several factors which we discuss in detail later. Our study will demonstrate how these four components fit together to form the total decision (or hypothesis-testing) problem.

The classical estimation problem is closely related to the detection problem. We describe it in detail later.
test will work. In Section 2.7 we develop bounds and approximate expressions for the performance that will be necessary for some of the later chapters.

Finally, in Section 2.8 we summarize our results and indicate some of the topics that we have omitted.

2.2 SIMPLE BINARY HYPOTHESIS TESTS

As a starting point we consider the decision problem in which each of two source outputs corresponds to a hypothesis. Each hypothesis maps into a point in the observation space. We assume that the observation space corresponds to a set of $N$ observations: $r_1, r_2, r_3, \ldots, r_N$. Thus each set can be thought of as a point in an $N$-dimensional space and can be denoted by a vector $r$:

\[ r = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix} \tag{3} \]

The probabilistic transition mechanism generates points in accord with the two known conditional probability densities $p_{r|H_1}(r|H_1)$ and $p_{r|H_0}(r|H_0)$. The object is to use this information to develop a suitable decision rule. To do this we must look at various criteria for making decisions.

2.2.1 Decision Criteria

In the binary hypothesis problem we know that either $H_0$ or $H_1$ is true. We shall confine our discussion to decision rules that are required to make a choice. (An alternative procedure would be to allow decision rules with three outputs (a) $H_0$ true, (b) $H_1$ true, (c) don't know.) Thus each time the experiment is conducted one of the four things can happen:

1. $H_0$ true; choose $H_0$.
2. $H_0$ true; choose $H_1$.
3. $H_1$ true; choose $H_1$.
4. $H_1$ true; choose $H_0$.

The first and third alternatives correspond to correct choices. The second and fourth alternatives correspond to errors. The purpose of a decision criterion is to attach some relative importance to the four possible courses of action. It might be expected that the method of processing the received
data (r) would depend on the decision criterion we select. In this section we show that for the two criteria of most interest, the Bayes and the Neyman–Pearson, the operations on r are identical.

**Bayes Criterion.** A Bayes test is based on two assumptions. The first is that the source outputs are governed by probability assignments, which are denoted by \( P_t \) and \( P_o \), respectively, and called the a priori probabilities. These probabilities represent the observer's information about the source before the experiment is conducted. The second assumption is that a cost is assigned to each possible course of action. We denote the cost for the four courses of action as \( C_{00}, C_{10}, C_{11}, C_{01} \), respectively. The first subscript indicates the hypothesis chosen and the second, the hypothesis that was true. Each time the experiment is conducted a certain cost will be incurred. We should like to design our decision rule so that on the average the cost will be as small as possible. To do this we first write an expression for the expected value of the cost. We see that there are two probabilities that we must average over; the a priori probability and the probability that a particular course of action will be taken. Denoting the expected value of the cost as the risk \( \mathcal{R} \), we have:

\[
\mathcal{R} = C_{00}P_0 \Pr(\text{say } H_0|H_0 \text{ is true}) \\
+ C_{10}P_0 \Pr(\text{say } H_1|H_0 \text{ is true}) \\
+ C_{11}P_1 \Pr(\text{say } H_1|H_1 \text{ is true}) \\
+ C_{01}P_1 \Pr(\text{say } H_0|H_1 \text{ is true}).
\]

Because we have assumed that the decision rule must say either \( H_1 \) or \( H_0 \), we can view it as a rule for dividing the total observation space \( Z \) into two parts, \( Z_0 \) and \( Z_1 \), as shown in Fig. 2.4. Whenever an observation falls in \( Z_0 \) we say \( H_0 \), and whenever an observation falls in \( Z_1 \) we say \( H_1 \).

We can now write the expression for the risk in terms of the transition probabilities and the decision regions:

\[
\mathcal{R} = C_{00}P_0 \int_{Z_0} p_{r|H_0}(r|H_0) \, dr \\
+ C_{10}P_0 \int_{Z_1} p_{r|H_0}(r|H_0) \, dr \\
+ C_{11}P_1 \int_{Z_1} p_{r|H_1}(r|H_1) \, dr \\
+ C_{01}P_1 \int_{Z_0} p_{r|H_1}(r|H_1) \, dr.
\]

For an \( N \)-dimensional observation space the integrals in (5) are \( N \)-fold integrals.

We shall assume throughout our work that the cost of a wrong decision is higher than the cost of a correct decision. In other words,

\[
C_{10} > C_{00}, \\
C_{01} > C_{11}.
\]

Now, to find the Bayes test we must choose the decision regions \( Z_0 \) and \( Z_1 \) in such a manner that the risk will be minimized. Because we require that a decision be made, this means that we must assign each point \( r \) in the observation space \( Z \) to \( Z_0 \) or \( Z_1 \).

Thus

\[
Z = Z_0 \cup Z_1.
\]

Rewriting (5), we have

\[
\mathcal{R} = P_0C_{00} \int_{Z_0} p_{r|H_0}(r|H_0) \, dr + P_0C_{10} \int_{Z_1 - Z_0} p_{r|H_0}(r|H_0) \, dr \\
+ P_1C_{01} \int_{Z_0} p_{r|H_1}(r|H_1) \, dr + P_1C_{11} \int_{Z_1 - Z_0} p_{r|H_1}(r|H_1) \, dr.
\]

Observing that

\[
\int_{Z} p_{r|H_0}(r|H_0) \, dr = \int_{Z} p_{r|H_1}(r|H_1) \, dr = 1,
\]

(8) reduces to

\[
\mathcal{R} = P_0C_{00} + P_1C_{01} \\
+ \int_{Z_0} ([P_1(C_{01} - C_{11})p_{r|H_1}(r|H_1)] \\
- [P_0(C_{10} - C_{00})p_{r|H_0}(r|H_0)]) \, dr.
\]
The first two terms represent the fixed cost. The integral represents the cost controlled by those points \( R \) that we assign to \( Z_0 \). The assumption in (6) implies that the two terms inside the brackets are positive. Therefore all values of \( R \) where the second term is larger than the first should be included in \( Z_0 \) because they contribute a negative amount to the integral. Similarly, all values of \( R \) where the first term is larger than the second should be excluded from \( Z_0 \) (assigned to \( Z_1 \)) because they would contribute a positive amount to the integral. Values of \( R \) where the two terms are equal have no effect on the cost and may be assigned arbitrarily. We shall assume that these points are assigned to \( H_1 \) and ignore them in our subsequent discussion. Thus the decision regions are defined by the statement:

If

\[
P_e(C_{01} - C_{11})P_{e|H_1}(R|H_1) \geq P_e(C_{10} - C_{00})P_{e|H_0}(R|H_0),
\]

assign \( R \) to \( Z_1 \) and consequently say that \( H_1 \) is true. Otherwise assign \( R \) to \( Z_0 \) and say \( H_0 \) is true.

Alternately, we may write

\[
P_{e|H_1}(R|H_1) \geq P_{e|H_0}(R|H_0)
\]

The quantity on the left is called the likelihood ratio and denoted by \( \Lambda(R) \)

\[
\Lambda(R) \triangleq \frac{P_{e|H_1}(R|H_1)}{P_{e|H_0}(R|H_0)}
\]

Because it is the ratio of two functions of a random variable, it is a random variable. We see that regardless of the dimensionality of \( R \), \( \Lambda(R) \) is a one-dimensional variable.

The quantity on the right of (12) is the threshold of the test and is denoted by \( \eta \):

\[
\eta \triangleq \frac{P_e(C_{10} - C_{00})}{P_e(C_{01} - C_{11})}
\]

Thus Bayes criterion leads us to a likelihood ratio test (LRT)

\[
\Lambda(R) \overset{H_1}{\gtrless} \overset{H_0}{\eta}.
\]

We see that all the data processing is involved in computing \( \Lambda(R) \) and is not affected by a priori probabilities or cost assignments. This invariance of the data processing is of considerable practical importance. Frequently the costs and a priori probabilities are merely educated guesses. The result in (15) enables us to build the entire processor and leave \( \eta \) as a variable threshold to accommodate changes in our estimates of a priori probabilities and costs.

Because the natural logarithm is a monotonic function, and both sides of (15) are positive, an equivalent test is

\[
\ln \Lambda(R) \overset{H_1}{\gtrless} \ln \eta.
\]

Two forms of a processor to implement a likelihood ratio test are shown in Fig. 2.5.

Before proceeding to other criteria, we consider three simple examples.

**Example 1.** We assume that under \( H_1 \) the source output is a constant voltage \( m \). Under \( H_0 \) the source output is zero. Before observation the voltage is corrupted by an additive noise. We sample the output waveform each second and obtain \( N \) samples. Each noise sample is a zero-mean Gaussian random variable \( n \) with variance \( \sigma^2 \). The noise samples at various instants are independent random variables and are independent of the source output. Looking at Fig. 2.6, we see that the observations under the two hypotheses are

\[
H_1: r_i = m + n_i, \quad i = 1, 2, \ldots, N,
\]

\[
H_0: r_i = n_i, \quad i = 1, 2, \ldots, N,
\]

and

\[
p_n(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{X^2}{2\sigma^2} \right).
\]

because the noise samples are Gaussian.

The probability density of \( r_i \) under each hypothesis follows easily:

\[
p_{n|H_1}(r_i|H_1) = p_n(r_i - m) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{(r_i - m)^2}{2\sigma^2} \right)
\]

and

\[
p_{n|H_0}(r_i|H_0) = p_n(r_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{r_i^2}{2\sigma^2} \right).
\]
In this example the only way the data appear in the likelihood ratio test is in a sum. This is an example of a sufficient statistic, which we denote by $l(R)$ (or simply $l$ when the argument is obvious). It is just a function of the received data which has the property that $\Lambda(R)$ can be written as a function of $l$. In other words, when making a decision, knowing the value of the sufficient statistic is just as good as knowing $R$. In Example 1, $l$ is a linear function of the $R_i$. A case in which this is not true is illustrated in Example 2.

Example 2. Several different physical situations lead to the mathematical model of interest in this example. The observations consist of a set of $N$ values: $r_1, r_2, \ldots, r_N$. Under both hypotheses, the $r_i$ are independent, identically distributed, zero-mean Gaussian random variables. Under $H_1$, each $r_i$ has a variance $\sigma_1^2$. Under $H_0$, each $r_i$ has a variance $\sigma_0^2$. Because the variables are independent, the joint density is simply the product of the individual densities. Therefore

$$p_{r_1|H_1}(r_1|H_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi} \sigma_1} \exp \left( - \frac{(r_i - m_1)^2}{2\sigma_1^2} \right)$$

and

$$p_{r_1|H_0}(r_1|H_0) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi} \sigma_0} \exp \left( - \frac{r_i^2}{2\sigma_0^2} \right).$$

Substituting (27) and (28) into (13) and taking the logarithm, we have

$$l(R) = \sum_{i=1}^{N} (\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}) R_i^2 + N \ln \frac{\sigma_0^2}{\sigma_1^2} \ln \eta.$$  \hspace{1cm} (29)

In this case the sufficient statistic is the sum of the squares of the observations

$$l(R) = \sum_{i=1}^{N} R_i^2,$$  \hspace{1cm} (30)

and an equivalent test for $\sigma_1^2 > \sigma_0^2$ is

$$l(R) \geq \frac{2\sigma_0^2 \sigma_1^2}{\sigma_0^2 - \sigma_1^2} \left( \ln \eta - N \ln \frac{\sigma_0^2}{\sigma_1^2} \right) \triangleq l_0.$$  \hspace{1cm} (31)

For $\sigma_1^2 < \sigma_0^2$ the inequality is reversed because we are multiplying by a negative number:

$$l(R) \leq \frac{2\sigma_0^2 \sigma_1^2}{\sigma_0^2 - \sigma_1^2} \left( N \ln \frac{\sigma_0^2}{\sigma_1^2} - \ln \eta \right) \triangleq l_1; \quad (\sigma_1^2 < \sigma_0^2).$$  \hspace{1cm} (32)

These two examples have emphasized Gaussian variables. In the next example we consider a different type of distribution.

Example 3. The Poisson distribution of events is encountered frequently as a model of shot noise and other diverse phenomena (e.g., [11] or [21]). Each time the experiment is conducted a certain number of events occur. Our observation is just this number which ranges from 0 to $\infty$ and obeys a Poisson distribution on both hypotheses; that is,

$$p_{r_1|H_1}(r_1|H_1) = \frac{m_1^r e^{-m_1}}{r!}, \quad r = 0, 1, 2, \ldots,$$  \hspace{1cm} (33)

where $m_1$ is the parameter that specifies the average number of events:

$$E(n) = m_1.$$  \hspace{1cm} (34)
It is this parameter \( m_i \) that is different in the two hypotheses. Rewriting (33) to emphasize this point, we have for the two Poisson distributions

\[
H_i : \text{Pr}(n \text{ events}) = \frac{m_i^n}{n!} e^{-m_i}, \quad n = 0, 1, 2, \ldots \tag{35}
\]

\[
H_0 : \text{Pr}(n \text{ events}) = \frac{m_0^n}{n!} e^{-m_0}, \quad n = 0, 1, 2, \ldots \tag{36}
\]

Then the likelihood ratio test is

\[
\Lambda(n) = \left( \frac{m_1}{m_0} \right)^n \exp \left[ -(m_1 - m_0) \frac{n}{m_0} \right], \tag{37}
\]

or, equivalently,

\[
\sum_{n=0}^{n_1} \ln \frac{n!}{n_0!} \ln \frac{m_1}{m_0}, \quad \text{if } m_1 > m_0,
\]

\[
\sum_{n=n_1}^{n_0} \ln \frac{n!}{n_0!} \ln \frac{m_1}{m_0}, \quad \text{if } m_0 > m_1. \tag{38}
\]

This example illustrates how the likelihood ratio test which we originally wrote in terms of probability densities can be simply adapted to accommodate observations that are discrete random variables. We now return to our general discussion of Bayes tests.

There are several special kinds of Bayes test which are frequently used and which should be mentioned explicitly.

If we assume that \( C_{00} \) and \( C_{11} \) are zero and \( C_{01} = C_{10} = 1 \), the expression for the risk in (8) reduces to

\[
\mathcal{R} = P_0 \int_{z_0} p_{r|H_0}(r|H_0) \, dr + P_1 \int_{z_0} p_{r|H_1}(r|H_1) \, dr. \tag{39}
\]

We see that (39) is just the total probability of making an error. Therefore for this cost assignment the Bayes test is minimizing the total probability of error. The test is

\[
\ln \Lambda(r) = \ln \frac{P_0}{P_1} = \ln P_0 - \ln (1 - P_0). \tag{40}
\]

When the two hypotheses are equally likely, the threshold is zero. This assumption is normally true in digital communication systems. These processors are commonly referred to as minimum probability of error receivers.

A second special case of interest arises when the a priori probabilities are unknown. To investigate this case we look at (8) again. We observe that once the decision regions \( Z_0 \) and \( Z_1 \) are chosen, the values of the integrals are determined. We denote these values in the following manner:

\[
P_F = \int_{z_0} p_{r|H_0}(r|H_0) \, dr,
\]

\[
P_D = \int_{z_0} p_{r|H_1}(r|H_1) \, dr,
\]

\[
P_M = \int_{z_0} p_{r|H_1}(r|H_1) \, dr = 1 - P_D.
\]

We see that these quantities are conditional probabilities. The subscripts are mnemonic and chosen from the radar problem in which hypothesis \( H_1 \) corresponds to the presence of a target and hypothesis \( H_0 \) corresponds to its absence. \( P_F \) is the probability of a false alarm (i.e., we say the target is present when it is not); \( P_D \) is the probability of detection (i.e., we say the target is present when it is); \( P_M \) is the probability of a miss (we say the target is absent when it is present). Although we are interested in a much larger class of problems than this notation implies, we shall use it for convenience.

For any choice of decision regions the risk expression in (8) can be written in the notation of (41):

\[
\mathcal{R} = P_0 C_{10} + P_1 C_{11} + P_1(C_{01} - C_{11})P_M
\]

\[
- P_0(C_{10} - C_{00})(1 - P_F). \tag{42}
\]

Because

\[
P_0 = 1 - P_1, \tag{43}
\]

(42) becomes

\[
\mathcal{R}(P_1) = C_{00}(1 - P_F) + C_{10}P_F
\]

\[
+ P_1[(C_{11} - C_{00}) + (C_{01} - C_{11})P_M - (C_{10} - C_{00})P_F]. \tag{44}
\]

Now, if all the costs and the priori probabilities are known, we can find a Bayes test. In Fig. 2.7a we plot the Bayes risk, \( \mathcal{R}_P(P_1) \), as a function of \( P_1 \). Observe that as \( P_1 \) changes the decision regions for the Bayes test change and therefore \( P_F \) and \( P_M \) change.

Now consider the situation in which a certain \( P_1 \) (say \( P_1 = P_1^* \)) is assumed and the corresponding Bayes test designed. We now fix the threshold and assume that \( P_1 \) is allowed to change. We denote the risk for this fixed threshold test as \( \mathcal{R}_P(P_1^*; P_1) \). Because the threshold is fixed, \( P_F \) and \( P_M \) are fixed, and (44) is just a straight line. Because it is a Bayes test for \( P_1 = P_1^* \), it touches the \( \mathcal{R}_P(P_1) \) curve at that point. Looking at (14), we see that the threshold changes continuously with \( P_1 \). Therefore, whenever \( P_1 \neq P_1^* \), the threshold in the Bayes test will be different. Because the Bayes test minimizes the risk,

\[
\mathcal{R}_P(P_1^*, P_1) \geq \mathcal{R}_P(P_1). \tag{45}
\]
If \( A \) is a continuous random variable with a probability distribution function that is strictly monotonic, then changing \( \eta \) always changes the risk \( \mathcal{R}_\eta(P_1) \) strictly concave downward and the inequality in (45) is strict. This case, which is one of particular interest to us, is illustrated in Fig. 2.7a. We see that \( \mathcal{R}_\eta(P_1^*) \) is tangent to \( \mathcal{R}_0(P_1) \) at \( P_1 = P_1^* \). These curves demonstrate the effect of incorrect knowledge of the a priori probabilities.

An interesting problem is encountered if we assume that the a priori probabilities are chosen to make our performance as bad as possible. In other words, \( P_1 \) is chosen to maximize our risk \( \mathcal{R}_0(P_1^*, P_1) \). Three possible examples are given in Figs. 2.7b, c, and d. In Fig. 2.7b the maximum of \( \mathcal{R}_0(P_1) \) occurs at \( P_1 = 0 \). To minimize the maximum risk we use a Bayes test designed assuming \( P_1 = 0 \). In Fig. 2.7c the maximum of \( \mathcal{R}_0(P_1) \) occurs at \( P_1 = 1 \). To minimize the maximum risk we use a Bayes test designed assuming \( P_1 = 1 \). In Fig. 2.7d the maximum occurs inside the interval \([0, 1]\), and we choose \( \mathcal{R}_F \) to be the horizontal line. This implies that the coefficient of \( P_1 \) in (44) must be zero:

\[
(C_{11} - C_{10}) + (C_{01} - C_{00})P_M - (C_{10} - C_{00})P_F = 0.
\]

A Bayes test designed to minimize the maximum possible risk is called a minimax test. Equation 46 is referred to as the minimax equation and is useful whenever the maximum of \( \mathcal{R}_\eta(P_1) \) is interior to the interval.

A special cost assignment that is frequently logical is

\[
C_{00} = C_{11} = 0
\]

(This guarantees the maximum is interior.)

Denoting,

\[
C_{01} = C_M,
C_{10} = C_F,
\]

the risk is,

\[
\mathcal{R}_F = C_F P_F + P_1(C_M P_M - C_F P_F)
= P_0 C_F P_F + P_1 C_M P_M
\]

and the minimax equation is

\[
C_M P_M = C_F P_F.
\]

Before continuing our discussion of likelihood ratio tests we shall discuss a second criterion and prove that it also leads to a likelihood ratio test.

**Neyman-Pearson Tests.** In many physical situations it is difficult to assign realistic costs or a priori probabilities. A simple procedure to bypass this difficulty is to work with the conditional probabilities \( P_F \) and \( P_B \). In general, we should like to make \( P_F \) as small as possible and \( P_B \) as large as possible. For most problems of practical importance these are conflicting objectives. An obvious criterion is to constrain one of the probabilities and maximize (or minimize) the other. A specific statement of this criterion is the following:

**Neyman-Pearson Criterion.** Constrain \( P_F = \alpha' \leq \alpha \) and design a test to maximize \( P_B \) (or minimize \( P_M \)) under this constraint.

The solution is obtained easily by using Lagrange multipliers. We construct the function \( F \),

\[
F = P_M + \lambda \{P_F - \alpha'\},
\]

or

\[
F = \int_{z_0} \rho r_1(H_1|H_0) dR + \lambda \left[ \int_{z_1} \rho r_1(H_0|H_0) dR - \alpha' \right].
\]

Clearly, if \( P_F = \alpha' \), then minimizing \( F \) minimizes \( P_M \).
or

\[ F = \lambda(1 - \alpha') + \int_{Z_0} [p_{R|H_1}(R|H_1) - \lambda p_{R|H_0}(R|H_0)] d\lambda. \quad (53) \]

Now observe that for any positive value of \( \lambda \) an LRT will minimize \( F \).

(A negative value of \( \lambda \) gives an LRT with the inequalities reversed.)

This follows directly, because to minimize \( F \) we assign a point \( R \) to \( Z_0 \) only when the term in the bracket is negative. This is equivalent to the test

\[ \frac{p_{R|H_1}(R|H_1)}{p_{R|H_0}(R|H_0)} < \lambda, \quad \text{assign point to } Z_0 \text{ or say } H_0. \quad (54) \]

The quantity on the left is just the likelihood ratio. Thus \( F \) is minimized by the likelihood ratio test

\[ \Lambda(R) \cong \frac{H_1}{H_0}. \quad (55) \]

To satisfy the constraint we choose \( \lambda \) so that \( P_r = \alpha' \). If we denote the density of \( \Lambda \) when \( H_0 \) is true as \( p_{\Lambda|H_0}(\Lambda|H_0) \), then we require

\[ P_r = \int_{\lambda}^{\infty} p_{\Lambda|H_0}(\Lambda|H_0) d\Lambda = \alpha'. \quad (56) \]

Solving (56) for \( \lambda \) gives the threshold. The value of \( \lambda \) given by (56) will be non-negative because \( p_{\Lambda|H_0}(\Lambda|H_0) \) is zero for negative values of \( \lambda \). Observe that decreasing \( \lambda \) is equivalent to increasing \( Z_1 \), the region where we say \( H_1 \). Thus \( P_r \) increases as \( \lambda \) decreases. Therefore we decrease \( \lambda \) until we obtain the largest possible \( \alpha' \leq \alpha \). In most cases of interest to us \( P_r \) is a continuous function of \( \lambda \) and we have \( P_r = \alpha \). We shall assume this continuity in all subsequent discussions. Under this assumption the Neyman–Pearson criterion leads to a likelihood ratio test. On p. 41 we shall see the effect of the continuity assumption not being valid.

**Summary.** In this section we have developed two ideas of fundamental importance in hypothesis testing. The first result is the demonstration that for a Bayes or a Neyman–Pearson criterion the optimum test consists of processing the observation \( R \) to find the likelihood ratio \( \Lambda(R) \) and then comparing \( \Lambda(R) \) to a threshold in order to make a decision. Thus, regardless of the dimensionality of the observation space, the decision space is one-dimensional.

The second idea is that of a sufficient statistic \( \mathcal{I}(R) \). The idea of a sufficient statistic originated when we constructed the likelihood ratio and saw that it depended explicitly only on \( \mathcal{I}(R) \). If we actually construct \( \Lambda(R) \) and then recognize \( \mathcal{I}(R) \), the notion of a sufficient statistic is perhaps of secondary value. A more important case is when we can recognize \( \mathcal{I}(R) \) directly. An easy way to do this is to examine the geometric interpretation of a sufficient

statistic. We considered the observations \( r_1, r_2, \ldots, r_N \) as a point \( r \) in an \( N \)-dimensional space, and one way to describe this point is to use these coordinates. When we choose a sufficient statistic, we are simply describing the point in a coordinate system that is more useful for the decision problem. We denote the first coordinate in this system by \( l \), the sufficient statistic, and the remaining \( N - 1 \) coordinates which will not affect our decision by the \((N - 1)\)-dimensional vector \( y \). Thus

\[ \Lambda(R) = \Lambda(L, Y) = \frac{p_{R|H_1}(L, Y|H_1)}{p_{R|H_0}(L, Y|H_0)}, \quad (57) \]

Now the expression on the right can be written as

\[ \Lambda(L, Y) = \frac{p_{R|H_1}(L|H_1)p_{Y|H_1}(Y|L, H_1)}{p_{R|H_0}(L|H_0)p_{Y|H_0}(Y|L, H_0)}. \quad (58) \]

If \( l \) is a sufficient statistic, then \( \Lambda(R) \) must reduce to \( \Lambda(L) \). This implies that the second terms in the numerator and denominator must be equal. In other words,

\[ p_{Y|H_0}(Y|L, H_0) = p_{Y|H_1}(Y|L, H_1) \]

because the density of \( y \) cannot depend on which hypothesis is true. We see that choosing a sufficient statistic simply amounts to picking a coordinate system in which one coordinate contains all the information necessary to making a decision. The other coordinates contain no information and can be disregarded for the purpose of making a decision.

In Example 1 the new coordinate system could be obtained by a simple rotation. For example, when \( N = 2 \),

\[ L = \frac{1}{\sqrt{2}} (R_1 + R_2), \]

\[ Y = \frac{1}{\sqrt{2}} (R_1 - R_2). \quad (60) \]

In Example 2 the new coordinate system corresponded to changing to polar coordinates. For \( N = 2 \)

\[ L = R_1^2 + R_2^2, \]

\[ Y = \tan^{-1} \frac{R_2}{R_1}. \quad (61) \]

Notice that the vector \( y \) can be chosen in order to make the demonstration of the condition in (59) as simple as possible. The only requirement is that the pair \( (l, y) \) must describe any point in the observation space. We should also observe that the condition

\[ p_{R|H_1}(Y|H_1) = p_{R|H_0}(Y|H_0) \]

(62)
does not imply (59) unless \( l \) and \( y \) are independent under \( H_1 \) and \( H_0 \).
Frequently we will choose \( y \) to obtain this independence and then use (62)
to verify that \( l \) is a sufficient statistic.

### 2.2.2 Performance: Receiver Operating Characteristic

To complete our discussion of the simple binary problem we must evaluate the performance of the likelihood ratio test. For a Neyman-Pearson test the values of \( P_F \) and \( P_D \) completely specify the test performance. Looking at (42) we see that the Bayes risk \( \mathcal{R}_\alpha \) follows easily if \( P_F \) and \( P_D \) are known. Thus we can concentrate our efforts on calculating \( P_F \) and \( P_D \).

We begin by considering Example 1 in Section 2.2.1.

**Example 1.** From (25) we see that an equivalent test is

\[
l = \frac{1}{\sqrt{N} \sigma} \sum_{i=1}^{N} R_i \left( \frac{\ln \eta}{\sqrt{N} m} - \frac{\sigma}{\sqrt{N} m} \ln \eta + \frac{\sqrt{N} m}{2 \sigma} \right)
\]

(63)

\[
P_{l|H_0}(L|H_0) \quad P_{l|H_1}(L|H_1)
\]

\[
d \triangleq \frac{\sqrt{N} m}{\sigma} \ln \eta + \frac{\sqrt{N} m}{2 \sigma} = \frac{\ln \eta}{d} + \frac{d}{2}
\]

(a)

\[
\text{Threshold: } -\frac{\sigma}{\sqrt{N} m} \ln \eta + \frac{\sqrt{N} m}{2 \sigma} = \frac{\ln \eta}{d} + \frac{d}{2}
\]

(b)

**Fig. 2.8** Error probabilities: (a) \( P_F \) calculation; (b) \( P_D \) calculation.

We have multiplied (25) by \( \sqrt{N} m \) to normalize the next calculation. Under \( H_1 \),
\( l \) is obtained by adding \( N \) independent zero-mean Gaussian variables with variance \( \sigma^2 \) and then dividing by \( \sqrt{N} \). Therefore \( l \) is \( N(0,1) \).

Under \( H_1 \), \( l \) is \( N(\sqrt{N} m/\sigma, 1) \). The probability densities on the two hypotheses are sketched in Fig. 2.8a. The threshold is also shown. Now, \( P_F \) is simply the integral of \( p_{l|H_0}(L|H_0) \) to the right of the threshold.

Thus

\[
P_F = \int_{\ln \eta/d + d/2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{x^2}{2} \right) \, dx,
\]

(64)

where \( d \triangleq \sqrt{N} m/\sigma \) is the distance between the means of the two densities. The integral in (64) is tabulated in many references (e.g., [3] or [4]).

We generally denote

\[
\text{erf}_\ast(X) \triangleq \int_{-\infty}^{X} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{x^2}{2} \right) \, dx,
\]

(65)

where \( \text{erf}_\ast \) is an abbreviation for the error function and

\[
\text{erfc}_\ast(X) \triangleq \int_{X}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{x^2}{2} \right) \, dx
\]

(66)

is its complement. In this notation

\[
P_F = \text{erfc}_\ast \left( \frac{\ln \eta}{d} + \frac{d}{2} \right).
\]

(67)

Similarly, \( P_D \) is the integral of \( p_{l|H_1}(L|H_1) \) to the right of the threshold, as shown in Fig. 2.8b:

\[
P_D = \int_{\ln \eta/d + d/2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{(x - d)^2}{2} \right] \, dx
\]

\[
= \int_{\ln \eta/d + d/2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{y^2}{2} \right) \, dy \triangleq \text{erfc}_\ast \left( \frac{\ln \eta}{d} - \frac{d}{2} \right).
\]

(68)

In Fig. 2.9a we have plotted \( P_D \) versus \( P_F \) for various values of \( d \) with \( \eta \) as the varying parameter. For \( \eta = 0 \), \( \ln \eta = -\infty \), and the processor always guesses \( H_1 \). Thus \( P_F = 1 \) and \( P_D = 0 \). As \( \eta \) increases, \( P_F \) and \( P_D \) decrease. When \( \eta = \infty \), the processor always guesses \( H_0 \) and \( P_F = P_D = 0 \).

As we would expect from Fig. 2.8, the performance increases monotonically with \( d \). In Fig. 2.9b we have replotted the results to give \( P_D \) versus \( d \) with \( P_F \) as a parameter on the curves. For a particular \( d \) we can obtain any point on the curve by choosing \( \eta \) appropriately (\( 0 \leq \eta \leq \infty \)).

The result in Fig. 2.9a is referred to as the receiver operating characteristic (ROC). It completely describes the performance of the test as a function of the parameter of interest.

A special case that will be important when we look at communication systems is the case in which we want to minimize the total probability of error

\[
\Pr(\ast) \triangleq P_F P_D + P_F P_M.
\]

(69a)

† The function that is usually tabulated is \( \text{erf}_\ast(X) \triangleq \sqrt{2/\pi} \int_{0}^{X} \exp (-y^2) \, dy \), which is related to (65) in an obvious way.
us to discuss its approximate behavior analytically. For \( X > 0 \)
\[
\frac{1}{\sqrt{2\pi}X} \left( 1 - \frac{1}{X^2} \right) \exp \left( -\frac{X^2}{2} \right) < \text{erfc}_* (X) < \frac{1}{\sqrt{2\pi}X} \exp \left( -\frac{X^2}{2} \right). \tag{71}
\]
This can be derived by integrating by parts. (See Problem 2.2.15 or Feller [30].) A second bound is
\[
\text{erfc}_* (X) < \frac{1}{4} \exp \left( -\frac{X^2}{2} \right), \quad x > 0, \tag{72}
\]

Fig. 2.9 (a) Receiver operating characteristic: Gaussian variables with unequal means.

The threshold for this criterion was given in (40). For the special case in which \( P_0 = P_1 \), the threshold \( \eta \) equals one and
\[
\Pr (\epsilon) = \frac{1}{2} \Pr (x). \tag{69b}
\]
Using (67) and (68) in (69), we have
\[
\Pr (\epsilon) = \int_{-\frac{d}{2}}^{\frac{d}{2}} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{x^2}{2} \right) dx = \text{erfc}_* \left( \frac{d}{2} \right). \tag{70}
\]
It is obvious from (70) that we could also obtain the \( \Pr (\epsilon) \) from the ROC. However, if this is the only threshold setting of interest, it is generally easier to calculate the \( \Pr (\epsilon) \) directly.

Before calculating the performance of the other two examples, it is worthwhile to point out two simple bounds on \( \text{erfc}_* (X) \). They will enable

Fig. 2.9 (b) detection probability versus \( d \).
which can also be derived easily (see Problem 2.2.16). The four curves are plotted in Fig. 2.10. We note that \( \text{erfc}_* (X) \) decreases exponentially.

The receiver operating characteristics for the other two examples are also of interest.

**Example 2.** In this case the test is

\[
I(R) = \sum_{i=1}^{N} R_{ir} = \frac{X \sum_{i=1}^{N} 2 \sigma_i^2 \gamma}{\sigma_1^2 - \sigma_2^2} \left( \ln \gamma - N \ln \frac{\sigma_2}{\sigma_1} \right) = \gamma, \quad (\sigma_1 > \sigma_2).
\]  

(73)

The performance calculation for arbitrary \( N \) is somewhat tedious, so we defer it to Section 2.6. A particularly simple case appearing frequently in practice is \( N = 2 \). Under \( H_0 \) the \( r_i \) are independent zero-mean Gaussian variables with variances equal to \( \sigma_0^2 \):

\[
P_{F} = \Pr \left( l \geq \gamma | H_0 \right) = \Pr \left( r_1^2 + r_2^2 \geq \gamma | H_0 \right).
\]  

(74)

To evaluate the expression on the right, we change to polar coordinates:

\[
r_1 = z \cos \theta, \quad z = \sqrt{r_1^2 + r_2^2}
\]

\[
r_2 = z \sin \theta, \quad \theta = \tan^{-1} \frac{r_2}{r_1}
\]  

(75)

Then

\[
\Pr \left( z^2 \geq \gamma | H_0 \right) = \int_0^{2\pi} d\theta \int_{\sqrt{\gamma}}^{\infty} Z \frac{1}{2\sigma_0^2} \exp \left( - \frac{Z^2}{2\sigma_0^2} \right) dZ.
\]  

(76)

Integrating with respect to \( \theta \), we have

\[
P_{F} = \int_{\sqrt{\gamma}}^{\infty} \frac{1}{2\sigma_0^2} \exp \left( - \frac{Z^2}{2\sigma_0^2} \right) dZ.
\]  

(77)

We observe that \( l \), the sufficient statistic, equals \( z^2 \). Changing variables, we have

\[
P_{F} = \int_{\frac{\gamma}{2\sigma_0^2}}^{\infty} \frac{1}{2\sigma_0^2} \exp \left( - \frac{L}{2\sigma_0^2} \right) dL = \exp \left( - \frac{\gamma}{2\sigma_0^2} \right).
\]  

(78)

(Note that the probability density of the sufficient statistic is exponential.)

Similarly,

\[
P_{D} = \exp \left( - \frac{\gamma}{2\sigma_0^2} \right).
\]  

(79)

To construct the ROC we can combine (78) and (79) to eliminate the threshold \( \gamma \).

This gives

\[
P_{D} = \left( P_{F} \right)^{\sigma_1^2 / \sigma_2^2}.
\]  

(80)

In terms of logarithms

\[
\ln P_{D} = \frac{\sigma_2^2}{\sigma_1^2} \ln P_{F}.
\]  

(81)

As expected, the performance improves monotonically as the ratio \( \sigma_1^2 / \sigma_2^2 \) increases.

We shall study this case and its generalizations in more detail in Section 2.6.

The two Poisson distributions are the third example.

**Example 3.** From (38), the likelihood ratio test is

\[
n \geq \gamma \frac{m_1 - m_0}{m_0 \ln \frac{m_1}{m_0}} = \gamma, \quad (m_1 > m_0).
\]  

(82)

Because \( n \) takes on only integer values, it is more convenient to rewrite (82) as

\[
n \geq \gamma \frac{\gamma_1}{\gamma_0}, \quad \gamma_1 = 0, 1, 2, \ldots
\]  

(83)
where $\gamma_l$ takes on only integer values. Using (35),

$$P_D = 1 - e^{-m_1} \sum_{n=0}^{\gamma_l-1} \frac{(m_1)^n}{n!}, \quad \gamma_l = 0, 1, 2, \ldots \tag{84}$$

and from (36)

$$P_F = 1 - e^{-m_0} \sum_{n=0}^{\gamma_l-1} \frac{(m_0)^n}{n!}, \quad \gamma_l = 0, 1, 2, \ldots \tag{85}$$

The resulting ROC is plotted in Fig. 2.11a for some representative values of $m_0$ and $m_1$.

We see that it consists of a series of points and that $P_F$ goes from 1 to $1 - e^{-m_0}$ when the threshold is changed from 0 to 1. Now suppose we wanted $P_F$ to have an intermediate value, say $1 - \frac{1}{2} e^{-m_0}$. To achieve this performance we proceed in the following manner. Denoting the LRT with $\gamma_l = 0$ as LRT No. 0 and the LRT with $\gamma_l = 1$ as LRT No. 1, we have the following table:

<table>
<thead>
<tr>
<th>LRT</th>
<th>$\gamma_l$</th>
<th>$P_F$</th>
<th>$P_D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>$1 - e^{-m_0}$</td>
<td>$1 - e^{-m_1}$</td>
</tr>
</tbody>
</table>

Fig. 2.11 (a) Receiver operating characteristic, Poisson problem.

![Fig. 2.11 (b) Receiver operating characteristic with randomized decision rule.](image)

To get the desired value of $P_F$ we use LRT No. 0 with probability $\frac{1}{2}$ and LRT No. 1 with probability $\frac{1}{2}$. The test is

If $n = 0$, say $H_1$ with probability $\frac{1}{2}$,

say $H_0$ with probability $\frac{1}{2}$,

$n \geq 1$ say $H_1$.

This procedure, in which we mix two likelihood ratio tests in some probabilistic manner, is called a randomized decision rule. The resulting $P_D$ is simply a weighted combination of detection probabilities for the two tests.

$$P_D = 0.5(1) + 0.5(1 - e^{-m_1}) = (1 - 0.5 e^{-m_1}). \tag{86}$$

We see that the ROC for randomized tests consists of straight lines which connect the points in Fig. 2.11a, as shown in Fig. 2.11b. The reason that we encounter a randomized test is that the observed random variables are discrete. Therefore $A(R)$ is a discrete random variable and, using an ordinary likelihood ratio test, only certain values of $P_F$ are possible.
Looking at the expression for $P_F$ in (56) and denoting the threshold by $\eta$, we have

$$P_F(\eta) = \int_0^\infty \rho_{X|H_0}(x|H_0) \, dx.$$  \hspace{1cm} (87)

If $P_F(\eta)$ is a continuous function of $\eta$, we can achieve a desired value from 0 to 1 by a suitable choice of $\eta$ and a randomized test will never be needed. This is the only case of interest to us in the sequel (see Prob. 2.2.12).

With these examples as a background, we now derive a few general properties of receiver operating characteristics. We confine our discussion to continuous likelihood ratio tests.

Two properties of all ROC's follow immediately from this example.

**Property 1.** All continuous likelihood ratio tests have ROC's that are concave downward. If they were not, a randomized test would be better. This would contradict our proof that a LRT is optimum (see Prob. 2.2.12).

**Property 2.** All continuous likelihood ratio tests have ROC's that are above the $P_D = P_F$ line. This is just a special case of Property 1 because the points $(P_F = 0, P_D = 0)$ and $(P_F = 1, P_D = 1)$ are contained on all ROC's.

**Property 3.** The slope of a curve in a ROC at a particular point is equal to the value of the threshold $\eta$ required to achieve the $P_D$ and $P_F$ of that point.

**Proof.**

$$P_D = \int_0^\infty \rho_{X|H_1}(x|H_1) \, dx,$$

$$P_F = \int_0^\infty \rho_{X|H_0}(x|H_0) \, dx.$$  \hspace{1cm} (88)

Differentiating both expressions with respect to $\eta$ and writing the results as a quotient, we have

$$\frac{dP_D}{d\eta} = -\frac{\rho_{X|H_1}(\eta|H_1)}{\rho_{X|H_0}(\eta|H_0)} = \frac{dP_D}{d\eta},$$  \hspace{1cm} (89)

We now show that

$$\frac{\rho_{X|H_1}(\eta|H_1)}{\rho_{X|H_0}(\eta|H_0)} = \eta.$$  \hspace{1cm} (90)

Let

$$\Omega(\eta) \triangleq \{ R|\Lambda(R) \geq \eta \} = \{ R|\rho_{X|H_1}(R|H_1) \geq \eta \} = \{ R|\rho_{X|H_0}(R|H_0) \geq \eta \}.$$  \hspace{1cm} (91)

Then

$$P_F(\eta) \triangleq \Pr \{ \Lambda(R) \geq \eta | H_1 \} = \int_{\Omega(\eta)} \rho_{X|H_1}(R|H_1) \, dR$$

$$= \int_{\Omega(\eta)} \Lambda(R)p_{X|H_0}(R|H_0) \, dR,$$  \hspace{1cm} (92)

where the last equality follows from the definition of the likelihood ratio. Using the definition of $\Omega(\eta)$, we can rewrite the last integral

$$P_D(\eta) = \int_{\Omega(\eta)} \Lambda(R)p_{X|H_0}(R|H_0) \, dR = \int_{\eta}^\infty \Lambda p_{X|H_0}(x|H_0) \, dx.$$  \hspace{1cm} (93)

Differentiating (93) with respect to $\eta$, we obtain

$$\frac{dP_D(\eta)}{d\eta} = -\eta \rho_{X|H_0}(\eta|H_0).$$  \hspace{1cm} (94)

Equating the expression for $dP_D(\eta)/d\eta$ in the numerator of (89) to the right side of (94) gives the desired result.

We see that this result is consistent with Example 1. In Fig. 2.9a, the curves for nonzero $d$ have zero slope at $P_F = P_D = 1$ ($\eta = 0$) and infinite slope at $P_F = P_D = 0$ ($\eta = \infty$).

**Property 4.** Whenever the maximum value of the Bayes risk is interior to the interval $(0, 1)$ on the $P_1$ axis, the minimax operating point is the intersection of the line

$$(C_{11} - C_{00}) + (C_{01} - C_{11})(1 - P_D) - (C_{10} - C_{00})P_F = 0,$$  \hspace{1cm} (95)

and the appropriate curve of the ROC (see 46). In Fig. 2.12 we show the special case defined by (50),

$$C_F P_F = C_M P_F = C_M (1 - P_D),$$  \hspace{1cm} (96)

**Fig. 2.12 Determination of minimax operating point.**
and the a priori probabilities are \( P_i \). The model is shown in Fig. 2.13. The expression for the risk is

\[
\mathcal{R} = \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} P_i C_{ij} \int_{Z_i} \mathbb{P}_{i|j}(R|H_i) dR. \tag{98}
\]

To find the optimum Bayes test we simply vary the \( Z_i \) to minimize \( \mathcal{R} \). This is a straightforward extension of the technique used in the binary case. For simplicity of notation we shall only consider the case in which \( M = 3 \) in the text.

Noting that \( Z_0 = Z - Z_1 - Z_2 \), because the regions are disjoint, we obtain

\[
\mathcal{R} = P_0 C_{00} \int_{Z_0} p_{r|H_0}(R|H_0) dR + P_0 C_{10} \int_{Z_1} p_{r|H_0}(R|H_0) dR + P_0 C_{20} \int_{Z_2} p_{r|H_0}(R|H_0) dR
\]

\[
+ P_1 C_{01} \int_{Z_0} p_{r|H_1}(R|H_1) dR + P_1 C_{11} \int_{Z_1} p_{r|H_1}(R|H_1) dR + P_1 C_{21} \int_{Z_2} p_{r|H_1}(R|H_1) dR
\]

\[
+ P_2 C_{02} \int_{Z_0} p_{r|H_2}(R|H_2) dR + P_2 C_{12} \int_{Z_1} p_{r|H_2}(R|H_2) dR + P_2 C_{22} \int_{Z_2} p_{r|H_2}(R|H_2) dR.
\]

This reduces to

\[
\mathcal{R} = P_0 C_{00} + P_1 C_{11} + P_2 C_{22}
\]

\[
+ \int_{Z_0} [P_0 (C_{02} - C_{22}) p_{r|H_0}(R|H_2) + P_1 (C_{01} - C_{11}) p_{r|H_1}(R|H_2)] dR
\]

\[
+ \int_{Z_1} [P_0 (C_{10} - C_{00}) p_{r|H_0}(R|H_0) + P_2 (C_{12} - C_{22}) p_{r|H_2}(R|H_2)] dR
\]

\[
+ \int_{Z_2} [P_0 (C_{20} - C_{00}) p_{r|H_0}(R|H_0) + P_1 (C_{21} - C_{11}) p_{r|H_1}(R|H_2)] dR.
\]

As before, the first three terms represent the fixed cost and the integrals represent the variable cost that depends on our choice of \( Z_0, Z_1, \) and \( Z_2 \). Clearly, we assign each \( R \) to the region in which the value of the integrand is the smallest. Labeling these integrands \( I_0(R) \), \( I_1(R) \), and \( I_2(R) \), we have the following rule:

\[
\text{if } I_0(R) < I_1(R) \text{ and } I_2(R), \text{ choose } H_0,
\]

\[
\text{if } I_1(R) < I_0(R) \text{ and } I_2(R), \text{ choose } H_1,
\]

\[
\text{if } I_2(R) < I_0(R) \text{ and } I_1(R), \text{ choose } H_2. \tag{101}
\]
We can write these terms in terms of likelihood ratios by defining
\[
\Lambda_1(R) = \frac{P_{R|H_1|R|H_0}}{P_{R|H_0}},
\]
\[
\Lambda_2(R) = \frac{P_{R|H_2|R|H_0}}{P_{R|H_0}}.
\]

Using (102) in (100) and (101), we have
\[
P_1(C_01 - C_1, ...) \Lambda_1(R) \overset{H_1 \text{ or } H_2}{\geq} P_0(C_{10} - C_{00}) + P_2(C_{12} - C_{02}) \Lambda_2(R),
\]
\[
P_2(C_{02} - C_{21}) \Lambda_2(R) \overset{H_2 \text{ or } H_1}{\geq} P_0(C_{20} - C_{00}) + P_1(C_{21} - C_{01}) \Lambda_1(R),
\]
\[
P_2(C_{12} - C_{22}) \Lambda_2(R) \overset{H_2 \text{ or } H_0}{\geq} P_0(C_{20} - C_{10}) + P_1(C_{21} - C_{11}) \Lambda_1(R).
\]

We see that the decision rules correspond to three lines in the \( \Lambda_1, \Lambda_2 \) plane. It is easy to verify that these lines intersect at a common point and therefore uniquely define three decision regions, as shown in Fig. 2.14. The decision space is two-dimensional for the three-hypothesis problem. It is easy to verify that \( M \) hypotheses always lead to a decision space which has, at most, \( M - 1 \) dimensions.

Several special cases will be useful in our later work. The first is defined by the assumptions
\[
C_{00} = C_{11} = C_{22} = 0, \quad C_{ij} = 1, \quad i \neq j.
\]

These equations indicate that any error is of equal importance. Looking at (98), we see that this corresponds to minimizing the total probability of error.

Substituting into (103)–(105), we have
\[
P_1 \Lambda_1(R) \overset{H_1 \text{ or } H_2}{\geq} P_0,
\]
\[
P_2 \Lambda_2(R) \overset{H_2 \text{ or } H_1}{\geq} P_0,
\]
\[
P_2 \Lambda_2(R) \overset{H_2 \text{ or } H_0}{\geq} P_1 \Lambda_1(R).
\]
2.3 \textit{Hypotheses}

The decision regions in the \((\Lambda_1, \Lambda_2)\) plane are shown in Fig. 2.15a. In this particular case, the transition to the \((\ln \Lambda_1, \ln \Lambda_2)\) plane is straightforward (Fig. 2.15b). The equations are

\[
\ln \Lambda_1(R) \begin{cases} H_1 \text{ or } H_2 & \geq \ln \frac{P_0}{P_1} \\ H_0 \text{ or } H_2 & \geq \ln \frac{P_1}{P_2} \end{cases}
\]

\[
\ln \Lambda_2(R) \begin{cases} H_1 \text{ or } H_2 & \geq \ln \frac{P_0}{P_1} \\ H_0 \text{ or } H_1 & \geq \ln \frac{P_1}{P_2} \end{cases} \quad (108)
\]

The expression in (107) and (108) are adequate, but they obscure an important interpretation of the processor. The desired interpretation is obtained by a little manipulation.

Substituting (102) into (103–105) and multiplying both sides by \(P_{r|H_0}(R|H_0)\), we have

\[
P_1P_{r|H_1}(R|H_1) \geq \frac{H_1 \text{ or } H_2}{H_0 \text{ or } H_2} \left\{ P_0P_{r|H_0}(R|H_0), \right. \]

\[
P_2P_{r|H_2}(R|H_2) \geq \frac{H_1 \text{ or } H_2}{H_0 \text{ or } H_1} \left\{ P_0P_{r|H_0}(R|H_0), \right. \]

\[
P_2P_{r|H_2}(R|H_2) \geq \frac{H_1 \text{ or } H_2}{H_0 \text{ or } H_0} \left\{ P_1P_{r|H_1}(R|H_1), \right. \]

Looking at (109), we see that an equivalent test is to compute the a posteriori probabilities \(Pr[H_0|R]\), \(Pr[H_1|R]\), and \(Pr[H_2|R]\) and choose the largest. (Simply divide both sides of each equation by \(P_r(R)\) and examine the resulting test.) For this reason the processor for the minimum probability of error criterion is frequently referred to as a maximum a posterior probability computer. The generalization to \(M\) hypotheses is straightforward.

The next two topics deal with degenerate tests. Both results will be useful in later applications. A case of interest is a degenerate one in which we combine \(H_1\) and \(H_2\). Then

\[
C_{12} = C_{21} = 0, \quad (110)
\]

and, for simplicity, we can let

\[
C_{01} = C_{10} = C_{20} = C_{02} \quad (111)
\]

and

\[
C_{00} = C_{11} = C_{22} = 0. \quad (112)
\]

Then (103) and (104) both reduce to

\[
P_1\Lambda_1(R) + P_2\Lambda_2(R) \geq \frac{H_1 \text{ or } H_2}{H_0} \quad P_0 \quad (113)
\]

and (105) becomes an identity.

The decision regions are shown in Fig. 2.16. Because we have eliminated all of the cost effect of a decision between \(H_1\) and \(H_2\), we have reduced it to a binary problem.

We next consider the dummy hypothesis technique. A simple example illustrates the idea. The actual problem has two hypotheses, \(H_1\) and \(H_2\), but occasionally we can simplify the calculations by introducing a dummy hypothesis \(H_0\) which occurs with zero probability. We let

\[
P_0 = 0, \quad P_1 + P_2 = 1, \quad (114)
\]

and

\[
C_{12} = C_{02}, \quad C_{21} = C_{01}. \quad (114)
\]

Substituting these values into (103–105), we find that (103) and (104) imply that we always choose \(H_1\) or \(H_2\) and the test reduces to

\[
P_2(C_{12} - C_{22}) \Lambda_2(R) \begin{cases} H_2 & \geq \frac{H_1}{H_1} \left\{ P_1(C_{21} - C_{11}) \Lambda_1(R), \right. \end{cases} \quad (115)
\]

Looking at (12) and recalling the definition of \(\Lambda_1(R)\) and \(\Lambda_2(R)\), we see that this result is exactly what we would expect. [Just divide both sides of (12) by \(P_{r|H_0}(R|H_0)\).] On the surface this technique seems absurd, but it will turn out to be useful when the ratio

\[
\frac{P_{r|H_2}(R|H_2)}{P_{r|H_1}(R|H_1)}
\]

is difficult to work with and the ratios \(\Lambda_1(R)\) and \(\Lambda_2(R)\) can be made simple by a proper choice of \(P_{r|H_0}(R|H_0)\).

In this section we have developed the basic results needed for the \(M\)-hypothesis problem. We have not considered any specific examples.
because the details involved in constructing the likelihood ratios are the same as those in the binary case. Typical examples are given in the problems. Several important points should be emphasized.

1. The minimum dimension of the decision space is no more than \( M - 1 \). The boundaries of the decision regions are hyperplanes in the \((A_1, \ldots, A_{M-1})\) plane.

2. The optimum test is straightforward to find. We shall find however, when we consider specific examples that the error probabilities are frequently difficult to compute.

3. A particular test of importance is the minimum total probability of error test. Here we compute the a posteriori probability of each hypothesis \( \Pr (H_i | \mathbf{R}) \) and choose the largest.

These points will be appreciated more fully as we proceed through various applications.

These two sections complete our discussion of simple hypothesis tests. A case of importance that we have not yet discussed is the one in which several source outputs are combined to give a single hypothesis. To study this detection problem, we shall need some ideas from estimation theory. Therefore we defer the composite hypothesis testing problem until Section 2.5 and study the estimation problem next.

### 2.4 ESTIMATION THEORY

In the last two sections we have considered a problem in which one of several hypotheses occurred. As the result of a particular hypothesis, a vector random variable \( \mathbf{r} \) was observed. Based on our observation, we shall try to choose the true hypothesis.

In this section we discuss the problem of parameter estimation. Before formulating the general problem, let us consider a simple example.

**Example 1.** We want to measure a voltage \( a \) at a single time instant. From physical considerations, we know that the voltage is between \(-V\) and \(+V\) volts. The measurement is corrupted by noise which may be modeled as an independent additive zero-mean Gaussian random variable \( n \). The observed variable is \( r \). Thus

\[
r = a + n.
\]

The probability density governing the observation process is \( p_{r|a}(R|A) \). In this case

\[
p_{r|a}(R|A) = p_a(R - A) = \frac{1}{\sqrt{2\pi} \sigma_s} \exp \left( - \frac{(R - A)^2}{2\sigma_s^2} \right).
\]

The problem is to observe \( r \) and estimate \( a \).

This example illustrates the basic features of the estimation problem.

A model of the general estimation problem is shown in Fig. 2.17. The model has the following four components:

- **Parameter Space.** The output of the source is a parameter (or variable). We view this output as a point in a parameter space. For the single-parameter case, which we shall study first, this will correspond to segments of the line \(-\infty < A < \infty\). In the example considered above the segment is \((-V, V)\).

- **Probabilistic Mapping from Parameter Space to Observation Space.** This is the probability law that governs the effect of \( a \) on the observation.

- **Observation Space.** In the classical problem this is a finite-dimensional space. We denote a point in it by the vector \( \mathbf{R} \).

- **Estimation Rule.** After observing \( \mathbf{R} \), we shall want to estimate the value of \( a \). We denote this estimate as \( \hat{a}(\mathbf{R}) \). This mapping of the observation space into an estimate is called the estimation rule. The purpose of this section is to investigate various estimation rules and their implementations.

The second and third components are familiar from the detection problem. The new features are the parameter space and the estimation rule. When we try to describe the parameter space, we find that two cases arise.

In the first, the parameter is a random variable whose behavior is governed by a probability density. In the second, the parameter is an unknown quantity but not a random variable. These two cases are analogous to the