I

Basic Detection Theory and One-Interval Designs

Part I introduces the *one-interval design*, in which a single stimulus is presented on each trial. The simplest and most important example is a correspondence experiment in which the stimulus is drawn from one of two stimulus classes and the observer tries to say from which class it is drawn. In auditory experiments, for example, the two stimuli might be a weak tone and no sound, tone sequences that may be slow or fast, or passages from the works of Mozart and Beethoven.

We begin by describing the use of one-interval designs to measure *discrimination*, the ability to tell two stimuli apart. Two types of such experiments may be distinguished. If one of the two stimulus classes contains only the null stimulus, as in the tone-versus-background experiment, the task is called *detection*. (This historically important application is responsible for the use of the term *detection theory* to refer to these methods.) If neither stimulus is null, the experiment is called *recognition*, as in the other examples. The methods for analyzing detection and recognition are the same, and we make no distinction between them (until chap. 10, where we consider experiments in which the two tasks are combined).

In chapters 1 and 2, we focus on designs with two possible responses as well as two stimulus classes. Because the possible responses in some applications (e.g., the tone detection experiment) are "yes" and "no," the paradigm with two stimuli, one interval, and two responses is sometimes termed *yes-no* even when the actual responses are, say, "slow" and "fast." Performance can be analyzed into two distinct elements: the degree to which the observer's responses mirror the stimuli (chap. 1) and the degree to which they display bias (chap. 2). Measuring these two elements requires a theory; we use the most common, normal-distribution variant of detection theory to

accomplish this end. Chapter 4 broadens the perspective on yes-no sensitivity and bias to include three classes of alternatives to this model: threshold theory, choice theory, and "nonparametric" techniques.

One-interval experiments may involve more than two responses or more than two possible stimuli. As an example of a larger response set, listeners could rate the likelihood that a passage was composed by Mozart rather than Beethoven on a 6-point scale. One-interval rating designs are discussed in chapter 3. As an example of a larger stimulus set, listeners could hear sequences presented at one of several different rates. If the requirement is to assign a different response to each stimulus, the task is called *identification*; if the stimuli are to be sorted into a smaller number of classes (perhaps slow, medium, and fast), it is *classification*. Chapter 5 applies detection-theory tools to identification and classification tasks, but only those in which elements of the stimulus sets differ in a single characteristic such as tempo. Identification and classification of more heterogeneous stimulus sets are considered in Part II.

1

The Yes-No Experiment: Sensitivity

In this book, we analyze experiments that measure the ability to distinguish between stimuli. An important characteristic of such experiments is that observers can be more or less *accurate*. For example, a radiologist's goal is to identify accurately those X-rays that display abnormalities, and participants in a recognition memory study are accurate to the degree that they can tell previously presented stimuli from novel ones. Measures of performance in these kinds of tasks are also called *sensitivity measures: High sensitivity* refers to good ability to discriminate, *low sensitivity* to poor ability. This is a natural term in detection studies—a sensitive listener hears things an insensitive one does not—but it applies as well to the radiology and memory examples.

Understanding Yes-No Data

Example 1: Face Recognition

We begin with a memory experiment. In a task relevant to understanding eyewitness testimony in the courtroom, participants are presented with a series of slides portraying people's faces, perhaps with the instruction to remember them. After a period of time (and perhaps some unrelated activity), recognition is tested by presenting the same participants with a second series that includes some of the same pictures, shuffled to a new random order, along with a number of "lures"—faces that were not in the original set. Memory is good if the person doing the remembering properly recognizes the Old faces, but not New ones. We wish to measure the ability to distinguish between these two classes of slides. Experiments of this sort have been performed to compare memory for faces of different races, orientations (upright vs. inverted), and many other variables (for a review, see Shapiro & Penrod, 1986).

Let us look at some (hypothetical) data from such a task. We are interested in just one characteristic of each picture: whether it is an Old face (one presented earlier) or a New face. Because the experiment concerns two kinds of faces and two possible responses, "yes" (I've seen this person before in this experiment) and "no" (I haven't), any of four types of events can occur on a single experimental trial. The number of trials of each type can be tabulated in a stimulus-response matrix like the following.

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old	20	5	25
New	10	15	25

The purpose of this yes-no task is to determine the participant's sensitivity to the Old/New difference. High sensitivity is indicated by a concentration of trials counted in the upper left and lower right of the matrix ("yes" responses to Old stimuli, "no" responses to New).

Summarizing the Data

Conventional, rather military language is used to describe the yes-no experiment. Correctly recognizing an Old item is termed a *hit*; failing to recognize it, a *miss*. Mistakenly recognizing a New item as old is a *false alarm*; correctly responding "no" to an Old item is, abandoning the metaphor, a *correct rejection*. In tabular terms:

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old (S_2)	Hits	Misses	
	(20)	(5)	(25)
New (S_1)	False alarms	Correct rejections	
	(10)	(15)	(25)

We use S_1 and S_2 as context-free names for the two stimulus classes.

Of the four numbers in the table (excluding the marginal totals), only two provide independent information about the participant's performance. Once we know, for example, the number of hits and false alarms, the other two entries are determined by how many Old and New items the experimenter decided to use (25 of each, in this case). Dividing each number by

the total in its row allows us to summarize the table by two numbers: The *hit rate* (*H*) is the proportion of Old trials to which the participant responded "yes," and the *false-alarm rate* (*F*) is the proportion of New trials similarly (but incorrectly) assessed. The hit and false-alarm rates can be written as conditional probabilities[1]

$$H = P(\text{"yes"} \mid S_2) \tag{1.1}$$

$$F = P(\text{"yes"} \mid S_1)\,, \tag{1.2}$$

where Equation 1.1 is read "The proportion of 'yes' responses when stimulus S_2 is presented."

In this example, $H = .8$ and $F = .4$. The entire matrix can be rewritten with response rates (or proportions) rather than frequencies:

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old (S_2)	.8	.2	1.0
New (S_1)	.4	.6	1.0

The two numbers needed to summarize an observer's performance, *F* and *H*, are denoted as an ordered *(false-alarm, hit)* pair. In our example, *(F, H)* = (.4, .8).

Measuring Sensitivity

We now seek a good way to characterize the observer's sensitivity. A function of *H* and *F* that attempts to capture this ability of the observer is called a sensitivity *measure*, *index*, or *statistic*. A perfectly sensitive participant would have a hit rate of 1 and a false-alarm rate of 0. A completely insensitive participant would be unable to distinguish the two stimuli at all and, indeed, could perform equally well without attending to them. For this observer, the probability of saying "yes" would not depend on the stimulus presented, so the hit and false-alarm rates would be the same. In interesting cases, sensitivity falls between these extremes: *H* is greater than *F*, but performance is not perfect.

[1] Technically, *H* and *F* are *estimates* of probabilities—a distinction that is important in statistical work (chap. 13). Probabilities characterize the observer's relation to the stimuli and are considered stable and unchanging; *H* and *F* may vary from one block of trials to the next.

The simplest possibility is to ignore one of our two response rates using, say, H to measure performance. For example, a lie detector might be touted as detecting 80% of liars or an X-ray reader as detecting 80% of tumors. (Alternatively, the hit rate might be ignored, and evaluation might depend totally on the false-alarm rate.) Such a measure is clearly inadequate. Compare the memory performance we have been examining with that of another group:

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old	8	17	25
New	1	24	25

Group 1 successfully recognized 80% of the Old words, Group 2 just 32%. But this comparison ignores the important fact that Group 2 participants just did not say "yes" very often. The hit rate, or any measure that depends on responses to only one of the two stimulus classes, cannot be a measure of sensitivity. To speak of sensitivity to a stimulus (as was done, for instance, in early psychophysics) is meaningless in the framework of detection theory.[2]

An important characteristic of sensitivity is that it can only be measured between two alternative stimuli and must therefore depend on both H and F. A moment's thought reveals that not all possible dependencies will do. Certainly a higher hit rate means greater, not less, sensitivity, whereas a higher false-alarm rate is an indicator of less sensitive performance. So a sensitivity measure should increase when either H increases or F decreases.

A final possible characteristic of sensitivity measures is that S_1 and S_2 trials should have equal importance: Missing an Old item is just as important an error as incorrectly recognizing a New one. In general, this is too strong a requirement, and we will encounter sensitivity measures that assign different weights to the two stimulus classes. Nevertheless, equal treatment is a good starting point, and (with one exception) the indexes described in this chapter satisfy it.

[2]The term *sensitivity is* used in this way, as a synonym for the hit rate, in medical diagnosis. *Specificity* is that field's term for the correct-rejection rate.

Two Simple Solutions

We are looking for a measure that goes up when H goes up, goes down when F goes up, and assigns equal importance to these statistics. How about simply subtracting F from H? The difference $H - F$ has all these characteristics. For the first group of memory participants, $H - F = .8 - .4 = .4$; for the second, $H - F = .32 - .04 = .28$, and Group 1 wins.

Another measure that combines H and F in this way is a familiar statistic, the proportion of correct responses, which we denote $p(c)$. To find proportion correct in conditions with equal numbers of S_1 and S_2 trials, we take the average of the proportion correct on S_2 trials (the hit rate, H) and the proportion correct on S_1 trials (the correct rejection rate, $1 - F$). Thus:

$$\begin{aligned} p(c) &= \tfrac{1}{2}[H + (1 - F)] \\ &= \tfrac{1}{2}(H - F) + \tfrac{1}{2} \ . \end{aligned} \tag{1.3}$$

If the numbers of S_1 and S_2 trials are not equal, then to find the literal proportion of trials on which a correct answer was given the actual numbers in the matrix would have to be used:

$$p(c)^* = (\text{hits} + \text{correct rejections})/\text{total trials} \ . \tag{1.4}$$

Usually it is more sensible to give H and F equal weight, as in Equation 1.3, because a sensitivity measure should not depend on the base presentation rate.

Let us look at $p(c)$ for equal presentations (Eq. 1.3). Is this a better or worse measure of sensitivity than $H - F$ itself? Neither. Because $p(c)$ depends directly on $H - F$ (and not on either H or F separately), one statistic goes up whenever the other does, and the two are monotonic functions of each other. Two measures that are monotonically related in this way are said to be *equivalent* measures of accuracy. In the running examples, $p(c)$ is .7 for Group 1 and .64 for Group 2, and $p(c)$ leads to the same conclusion as $H - F$. For both measures, Group 1 outscores Group 2.

A Detection Theory Solution

The most widely used sensitivity measure of detection theory (Green & Swets, 1966) is not quite as simple as $p(c)$, but bears an obvious family re-

semblance. The measure is called d' ("dee-prime") and is defined in terms of z, the inverse of the normal distribution function:

$$d' = z(H) - z(F) . \tag{1.5}$$

The z transformation converts a hit or false-alarm rate to a z score (i.e., to standard deviation units). A proportion of .5 is converted into a z score of 0, larger proportions into positive z scores, and smaller proportions into negative ones. To compute z, consult Table A5.1 in Appendix 5. The table makes use of a symmetry property of z scores: Two proportions equally far from .5 lead to the same absolute z score (positive if $p > .5$, negative if $p < .5$) so that:

$$z(1 - p) = -z(p) . \tag{1.6}$$

Thus, $z(.4) = -.253$, the negative of $z(.6)$. Use of the Gaussian z transformation is dominant in detection theory, and we often refer to normal-distribution models by the abbreviation *SDT*.

We can use Equation 1.5 to calculate d' for the data in the memory example. For Group 1, $H = .8$ and $F = .4$, so $z(H) = 0.842$, $z(F) = -0.253$, and $d' = 0.842 - (-0.253) = 1.095$. When the hit rate is greater than .5 and the false-alarm rate is less (as in this case), d' can be obtained by adding the absolute values of the corresponding z scores. For Group 2, $H = .32$ and $F = .04$, so $d' = -0.468 - (-1.751) = 1.283$. When the hit and false-alarm rates are on the same side of .5, d' is obtained by subtracting the absolute values of the z scores. Interestingly, by the d' measure, it is Group 2 (the one that was much more stingy with "yes" responses) rather than Group 1 that has the superior memory.

When observers cannot discriminate at all, $H = F$ and $d' = 0$. Inability to discriminate means having the same rate of saying "yes" when Old faces are presented as when New ones are offered. As long as $H \geq F$, d' must be greater than or equal to 0. The largest possible *finite* value of d' depends on the number of decimal places to which H and F are carried. When $H = .99$ and $F = .01$, $d' = 4.65$; many experimenters consider this an effective ceiling.

Perfect accuracy, on the other hand, implies an infinite d'. Two adjustments to avoid infinite values are in common use. One strategy is to convert proportions of 0 and 1 to $1/(2N)$ and $1 - 1/(2N)$, respectively, where N is the number of trials on which the proportion is based. Suppose a participant has 25 hits and 0 misses ($H = 1.0$) to go with 10 false alarms and 15 correct rejections ($F = .4$). The adjustment yields 24.5 hits and 0.5 misses, so $H = .98$ and $d' = 2.054 - (-0.253) = 2.307$. A second strategy (Hautus, 1995; Miller,

1996) is to add 0.5 to *all* data cells regardless of whether zeroes are present. This adjustment leads to $H = 25.5/26 = .981$ and $F = 10.5/26 = .404$. Rounding to two decimal places yields the same value as before, but d' is slightly smaller if computed exactly.

Most experiments avoid chance and perfect performance. Proportions correct between .6 and .9 correspond roughly to d' values between 0.5 and 2.5. Correct performance on 75% of both S_1 and S_2 trials yields a d' of 1.35; 69% for both stimuli gives $d' = 1.0$.

It is sometimes important to calculate d' when only $p(c)$ is known, not H and F. (Partial ignorance of this sort is common when reanalyzing published data.) Strictly speaking, the calculation cannot be done, but an approximation can be made by assuming that the hit rate equals the correct rejection rate so that $H = 1 - F$. For example, if $p(c) = .9$, we can guess at a measure for sensitivity: $d' = z(.9) - z(.1) = 1.282 - (-1.282) = 2.56$. To simplify the calculation, notice that one z score is the negative of the other (Eq. 1.6). Hence, in this special case:

$$d' = 2\, z[p(c)]\ . \tag{1.7}$$

This calculation is *not* correct in general. For example, suppose $H = .99$ and $F = .19$, so that H and the correct rejection rate are not equal. Then $p(c)$ still equals .9, but $d' = z(.99) - z(.19) = 2.326 - (-0.878) = 3.20$ instead of 2.56, a considerable discrepancy.

Implied ROCs

ROC Space and Isosensitivity Curves

What justifies the use of d' as a summary of discrimination? Why is this measure better, according to detection theory, than the more familiar $p(c)$? A good sensitivity measure should be invariant when factors other than sensitivity change. Participants are assumed by detection theory to have a fixed sensitivity when asked to discriminate a specific pair of stimulus classes. One aspect of responding that is up to them, however, is their willingness to respond "yes" rather than "no." If d' is an invariant measure of sensitivity, then a participant whose false-alarm and hit rates are (.4, .8) can also produce the performance pairs (.2, .6) and (.07, .35); all of these pairs indicate a d' of about 1.09, and differ only in response bias.

The locus of (false-alarm, hit) pairs yielding a constant d' is called an *isosensitivity* curve because all points on the curve have the same sensitivity.

This term was proposed by Luce (1963a) as more descriptive that the original engineering nomenclature *receiver operating characteristic* (ROC). Swets (1973) reinterpreted the acronym to mean *relative operating characteristic*. We use all these terms interchangeably.

Figure 1.1 shows ROCs implied by d'. The axes of the ROC are the false-alarm rate, on the horizontal axis, and the hit rate, plotted vertically. Because both H and F range from 0 to 1, the *ROC space*, the region in which ROCs must lie, is the unit square. For every value of the false-alarm rate, the plot shows the hit rate that would be obtained to yield a particular sensitivity level. Algebraically, these curves are calculated by solving Equation 1.5 for H; different curves represent different values of d'.

When performance is at chance ($d' = 0$), the ROC is the major diagonal, where the hit and false-alarm rates are equal. For this reason, the major diagonal is sometimes called the *chance line*. As sensitivity increases, the curves shift toward the upper left corner, where accuracy is perfect ($F = 0$ and $H = 1$). These ROC curves summarize the predictions of detection theory: If an observer in a discrimination experiment produces a (F, H) pair that lies on a particular implied ROC, that observer should be able to display any other (F, H) pair on the same curve.

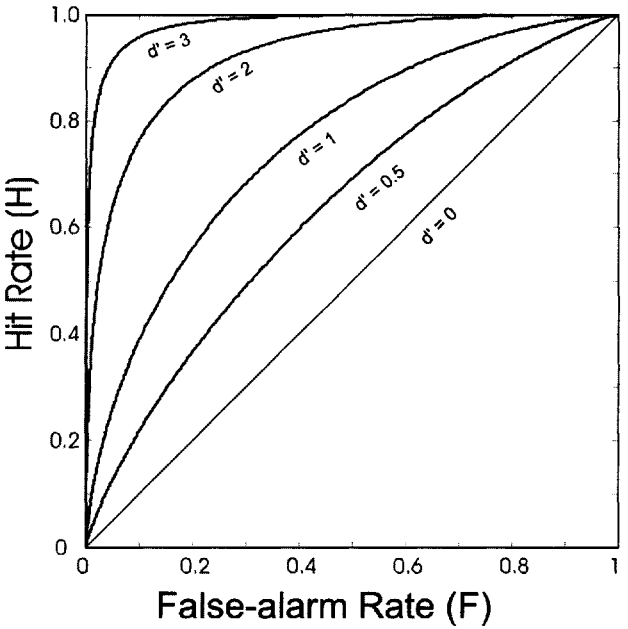

FIG. 1.1. ROCs for SDT on linear coordinates. Curves connect locations with constant d'.

The theoretical isosensitivity curves in Fig. 1.1 have two important characteristics. First, the price of complete success in recognizing one stimulus class is complete failure in recognizing the other. For example, to be perfectly correct with Old faces and have a hit rate of 1, it is also necessary to have a false-alarm rate of 1, indicating total failure to correctly reject New faces. Similarly, a false-alarm rate of 0 can be obtained only if the hit rate is 0. Isosensitivity curves that pass through (0, 0) and (1, 1) are called *regular* (Swets & Pickett, 1982).

Second, the slope of these curves decreases as the tendency to respond "yes" increases. The slope is the change in the hit rate, relative to the change in the false-alarm rate, that results from increasing response bias toward "yes." We shall see in a later section that this systematic slope change characterizes all ROCs.

ROCs in Transformed Coordinates

The features of regularity and decreasing slope are clear in Fig. 1.1, but other aspects of ROC shape are easier to see using a different representation of the ROC, one that takes advantage of our earlier description of a sensitivity measure as the difference between the transformed hit and false-alarm rates.

Look again at Equation 1.5, which describes the isosensitivity curve for d'. To find an algebraic expression for the ROC, we would need to solve this equation for H as a function of F. A simpler task is to solve for $z(H)$ as a function of $z(F)$:

$$z(H) = z(F) + d' \quad . \tag{1.8}$$

Equation 1.8 describes a *transformed ROC*, specifically a zROC, in which both axes are marked off in equal z scores rather than in equal proportion units. The range of values in these new units is from minus to plus infinity, although scores of more than 2.5 (i.e., 2.5 standard deviations from the mean) are rarely encountered. In these coordinates, the ROC has a particularly simple shape: It is a straight line with unit slope, as shown in Fig. 1.2.

The linearity of zROCs can be used to make a prediction about how much the false-alarm rate will go up if the hit rate increases (or vice versa). For example, suppose the false-alarm/hit pair (.2, .5) is on the ROC. Consulting Table A5.1, the z scores for F and H are –0.842 and 0. If we add the same number to each z score, the resulting scores correspond to another point on the ROC. Let us add 1.4, giving us the new z scores of 0.558 and 1.4. The table shows that the corresponding proportions are (.71, .92).

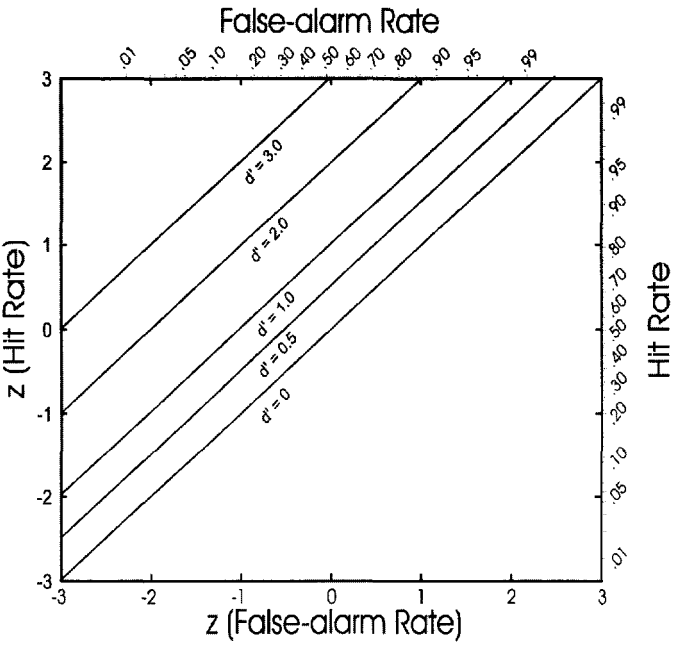

FIG. 1.2. ROCs for SDT on z coordinates.

The transformed ROC of Equation 1.8 provides a simple graphical interpretation of sensitivity: d' is the intercept of the straight-line ROC in Fig. 1.2, the vertical distance in z units from the ROC to the chance line at the point where $z(F) = 0$. In fact, because the ROC has slope 1, the distance between these two lines is the same no matter what the false-alarm rate is, and d' equals the vertical (or horizontal) distance between them at any point.

ROCs Implied by p(c)

Any sensitivity index has an implied ROC, that is, a curve in ROC space that connects points of equal sensitivity as measured by that index. To extend our comparison of d' with proportion correct, we now plot the ROC implied by $p(c)$. The trick is to take the definition of $p(c)$ in Equation 1.3 and solve it for H:

$$H = F + [2\,p(c) - 1] \ . \tag{1.9}$$

Equation 1.9 is a straight line of unit slope. Implied ROCs for $p(c)$ are shown in Fig. 1.3 for $p(c) = .5$, .65, and .8. The intercepts equal $2p(c) - 1$, that is, 0, .3, and .6.

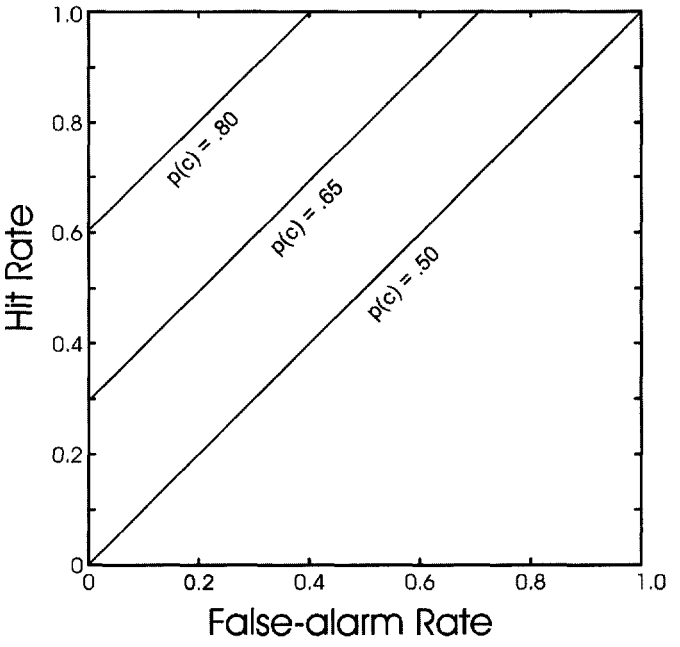

FIG. 1.3. ROCs implied by $p(c)$ on linear coordinates.

Consider again the false-alarm/hit pair (.2, .5). If we add the same number to each of these scores (without any transformation), the resulting scores correspond to another point on the ROC. Let us add .42, giving us the new hit and false-alarm proportions of (.62, .92). Simply using $p(c)$ as a measure of performance thus makes a prediction about how much the false-alarm rate will go up if the hit rate increases, and it is different from the prediction of detection theory.

Which Implied ROCs Are Correct?

The validity of detection theory clearly depends on whether the ROCs implied by d' describe the changes that occur in H and F when response bias is manipulated. Do empirical ROCs (the topic of chap. 3) look like those implied by d', those implied by $p(c)$, or something else entirely? It turns out that the detection theory curves do a much better job than those for $p(c)$. In early psychoacoustic research (Green & Swets, 1966) and subsequent work in many content areas (Swets, 1986a), ROCs were found to be regular, to have decreasing slope on linear coordinates, and to follow straight lines on z coordinates.

One property of the *z*ROCs described by Equation 1.8 that is *not* always observed experimentally is that of unit slope. When response bias changes, the value of d' calculated from Equation 1.5 may systematically increase or decrease instead of remaining constant. The unit-slope property reflects the equal importance of S_1 and S_2 trials to the corresponding sensitivity measure. In chapter 3, we discuss modified indexes that allow for unequal treatment.

When ROCs do have unit slope, they are symmetrical around the minor diagonal. Making explicit the dependence of sensitivity on a hit and false-alarm rate, we can express this property as

$$d'(1 - H, 1 - F) = d'(F, H) \ . \tag{1.10}$$

That is, if an observer changes response bias so that the new false-alarm rate is the old miss rate $(1 - H)$, then the new hit rate will be the old correct-rejection rate $(1 - F)$. For example, $d'(.6, .9) = d'(.1, .4)$. Mathematically, this occurs because $z(1 - p) = -z(p)$ (Eq. 1.6). Figure 1.4 provides a graphical interpretation of this relation, showing that (F, H) and $(1 - H, 1 - F)$ are on the same unit-slope ROC.

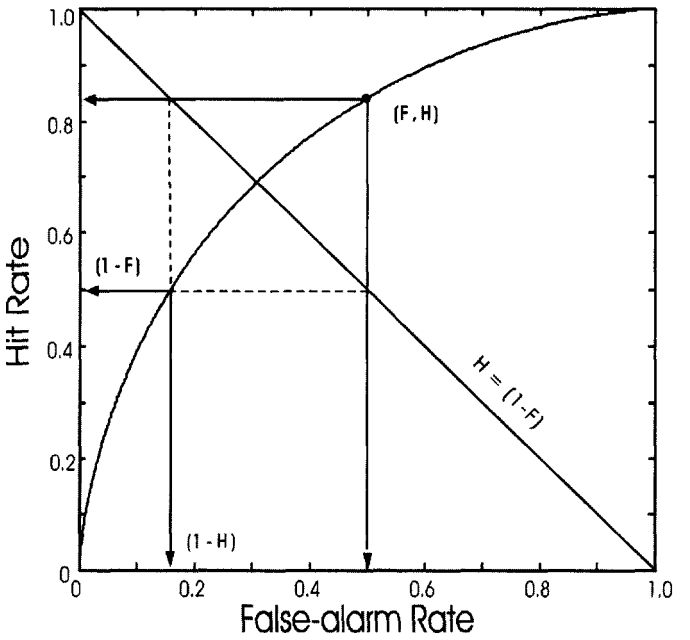

FIG. 1.4. The points (F, H) and $(1 - H, 1 - F)$ lie on the same symmetric ROC curve.

Sensitivity as Perceptual Distance

Stimuli that are easy to discriminate can be thought of as perceptually far apart; in this metaphor, a discrimination statistic should measure perceptual distance, and d' has the mathematical properties of distance measures (Luce, 1963a): The distance between an object and itself is 0, all distances are positive (*positivity*), the distance between objects x and y is the same as between y and x (*symmetry*), and

$$d'(x, w) \leq d'(x, y) + d'(y, w) . \tag{1.11}$$

Equation 1.11 is known as the *triangle inequality*.

Because they have true zeroes and are unique except for the choice of unit, distance measures have *ratio scaling* properties. That is, when discriminability is measured by d', it makes sense to say that stimuli a and b are twice as discriminable as stimuli c and d. Suppose, for example, that two participants in our face-recognition experiment produce d' values of 1.0 and 2.0. In a second test, a day later, their sensitivities fall to 0.5 and 1.0. Although the change in d' is twice as great for Participant 2, we can say that Old and New items are half as perceptually distant, for both participants, as on the first day. No corresponding statement can be made in the language of $p(c)$.

The positivity property means that d' should not be negative in the long run. Negative values can arise by chance when calculated over a small number of trials and are not a cause for concern. The temptation to whitewash such negative values into zeroes should be resisted: When a number of measurements are averaged, this strategy inflates a true d' of 0 into a positive one.

The triangle inequality (Eq. 1.11) is sometimes replaced by a stronger assumed relation—namely,

$$d'(x, w)^n = d'(x, y)^n + d'(y, w)^n . \tag{1.12}$$

When $n = 2$, this is the Euclidean distance formula. When $n = 1$, Equation 1.12 describes the "city-block" metric; an important special case (discussed in chap. 5) arises when stimuli differ perceptually along only one dimension.

Another distance property of d' is *unboundedness*: There is no maximum value of d', and perfect performance corresponds to infinity. In practice, occasional hit rates or false-alarm rates of 1 or 0 may occur, and a correction such as one of those discussed earlier must be made to subject the data to detection theory analysis. Any such correction is predicated on the belief that

the perfect performance arises from statistical ("sampling") error. If, on the contrary, stimulus differences are so great that confusions are effectively impossible then the experiment suffers from a ceiling effect, and should be redesigned.

The Signal Detection Model

The question under discussion to this point has been how best to measure accuracy. We have defended d' on pragmatic grounds. It represents the difference between the transformed hit and false-alarm rates, and it provides a good description of the relation between H and F when response bias varies. Now we ask what our measures imply about the process by which discrimination (in our example, face recognition) takes place. How are items represented internally, and how does the participant make a decision about whether a particular item is Old or New?

Underlying Distributions and the Decision Space

Detection theory assumes that a participant in our memory experiment is judging a single attribute, which we call *familiarity*. Each stimulus presentation yields a value of this decision variable. Repeated presentations do not always lead to the same result, but generate a distribution of values. The first panel of Fig. 1.5 presents the probability distribution (or likelihood distribution, or probability density) of familiarity values for New faces (stimulus class S_1). Each value on the horizontal axis has some likelihood of arising from New stimuli, indicated on the ordinate. The probability that a value above the point k will occur is the proportion of area under the curve above k (see Appendix 1 for a review of probability concepts).

On the average, Old items are more familiar than New ones—otherwise, the participant would not be able to discriminate. Thus, the whole of the distribution of familiarity due to Old (S_2) stimuli, shown in the second panel, is displaced to the right of the New distribution. There must be at least some values of the decision variable that the participant finds ambiguous, that could have arisen either from an Old or a New face; otherwise performance would be perfect. The two distributions together comprise the *decision space*—the internal or *underlying* problem facing the observer. The participant can assess the familiarity value of the stimulus, but of course does not know which distribution led to that value. What is the best strategy for deciding on a response?

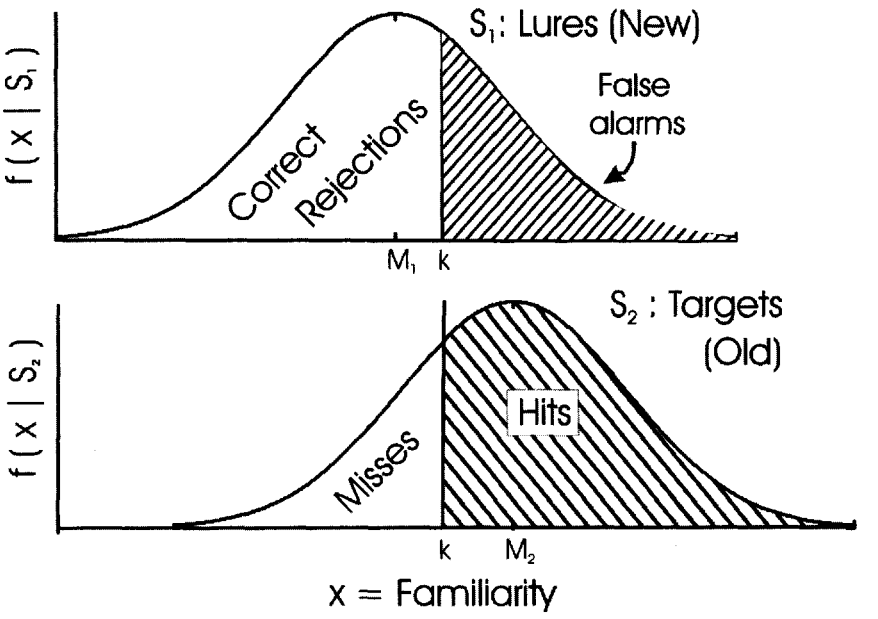

FIG. 1.5. Underlying distributions of familiarity for Old and New items. Top curve shows distribution due to New (S_1) items; values above the criterion k lead to false alarms, those below to correct rejections. Lower curve shows distribution due to Old (S_2) items; values above the criterion k lead to hits, those below to misses. The means of the distributions are M_1 and M_2. (In this and subsequent figures, the height of the probability density curve is denoted by f.)

Response Selection in the Decision Space

The optimal rule (see Green & Swets, 1966, ch. 1) is to establish a *criterion* that divides the familiarity dimension into two parts. Above the criterion, labeled k in Fig. 1.5, the participant responds "yes" (the face is familiar enough to be Old); below the criterion, a "no" is called for. The four possible stimulus–response events are represented in the figure. If a value above the criterion arises from the Old stimulus class, the participant responds "yes" and scores a hit. The hit rate H is the proportion of area under the Old curve that is above the criterion; the area to the left of the criterion is the proportion of misses. When New stimuli are presented (upper curve), a familiarity value above the criterion leads to a false alarm. The false-alarm rate is the proportion of area under the New curve to the right of the criterion, and the area to the left of the criterion equals the correct-rejection rate.

The decision space provides an interpretation of how ROCs are produced. The participant can change the proportion of "yes" responses, and generate different points on an ROC, by moving the criterion: If the criterion is raised, both H and F will decrease, whereas lowering the criterion will increase H and F.

We saw earlier that an important feature of ROCs is regularity: If $F = 0$, then $H = 0$; if $H = 1$, then $F = 1$. Examining Fig. 1.5, this implies that if the criterion is moved so far to the right as to be beyond the entire S_1 density (so that $F = 0$), it will be beyond the entire S_2 density as well (so that $H = 0$). The other half of the regularity condition is interpreted similarly. The distributions most often used satisfy this requirement by assuming that *any* value on the decision axis can arise from either distribution.

Sensitivity in the Decision Space

We have seen that k, the criterion value of familiarity, provides a natural interpretation of response bias. What aspect of the decision space reflects sensitivity? When sensitivity is high, Old and New items differ greatly in average familiarity, so the two distributions in the decision space have very different means. When sensitivity is low, the means of the two distributions are close together. Thus, the mean difference between the S_1 and S_2 distributions—the distance $M_2 - M_1$ in Fig. 1.5—is a measure of sensitivity. We shall soon see that this distance is in fact identical to d'.

Distance along a line, as in Fig. 1.5, can be measured from any zero point; so we measure mean distances relative to the criterion k. Thus expressed, the mean difference equals $(M_2 - k) - (M_1 - k)$: Sensitivity is the difference between these two distances, the distance from the S_1 mean to the criterion and the (negative, in this case) distance from the S_2 mean to the criterion. We now show that these two mean-to-criterion distances can be estimated using the z transformation discussed earlier in the chapter.

Underlying Distributions and Transformations

Figure 1.6 shows how the distances between the means of underlying distributions and the criterion are related to the response rates in our experiment. For each value of $M - k$, the figure shows the proportion of the area of an underlying distribution that is above the criterion. When $M - k = 0$, for example, the "yes" rate is 50%; large positive differences correspond to high "yes" rates and large negative differences to low ones. The curve in Fig. 1.6 is called a *(cumulative) distribution function*; in the language of calculus, it is the integral of the probability distributions shown in Fig. 1.5.

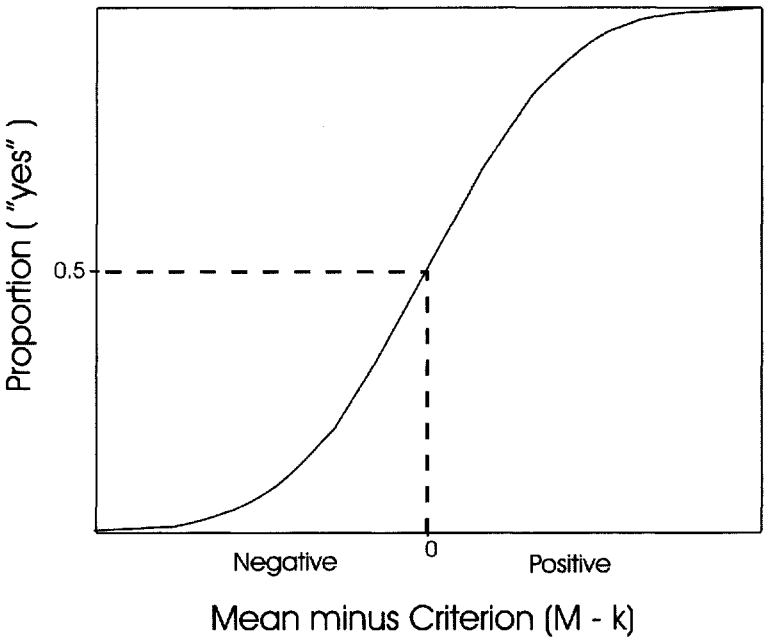

FIG. 1.6. A cumulative distribution function (the integral of one of the densities in Fig. 1.5) giving the proportion of "yes" responses as a function of the difference between the distribution mean and the criterion.

We can use the distribution function to translate any "yes" proportion into a value of $M - k$. This is the tie between the decision space and our sensitivity measures: For any hit rate and false-alarm rate (both "yes" proportions), we can use the distribution function to find two values of $M - k$ and subtract them to find the distance between the means. The distribution function transforms a distance into a proportion; we are interested in the inverse function, from proportions to distances, denoted z. In Fig. 1.7, the hit and false-alarm proportions from our face-recognition example are ordinate values, and the corresponding values $z(H)$ and $z(F)$ are abscissa values. The distance between these abscissa points, $z(H) - z(F)$, is the distance between the S_1 and S_2 means in Fig. 1.5. It is also, by Equation 1.5, equal to d'. Because z measures distance in standard deviation units, so does d'. Thus, the sensitivity measure d' is the distance between the means of the two underlying distributions in units of their common standard deviation.

The distance between the means of distributions is a congenial interpretation of d' because it is unchanged by response bias. No matter where the

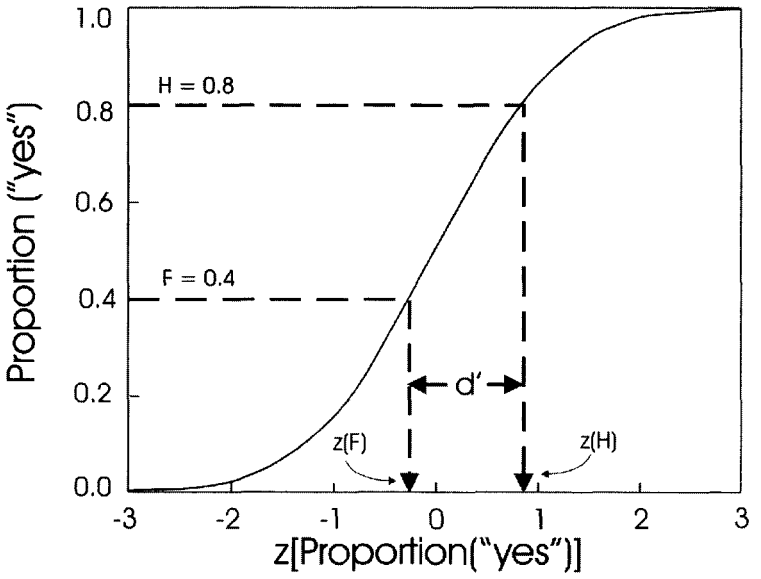

FIG. 1.7. A cumulative normal distribution function. The inverse function can be used to transform the proportions H and F into z scores, and sensitivity is the difference between $z(H)$ and $z(F)$.

participant locates the criterion, d' equals the same number. This relation is not specific to normal-distribution SDT: Any sensitivity measure obtained by subtracting transformed hit and false-alarm rates can be represented as the distance between the means of two distributions whose shape is given by the inverse of the transformation.

We can now venture a "definition" that will at least delimit the contents of this book. By *detection theory* we mean a theory relating choice behavior to a psychological decision space. An observer's choices are determined by the distances between distributions in the space due to different stimuli (sensitivities) and by the manner in which the space is partitioned to generate the possible responses (response biases).

Calculational Methods

Calculation of d' (and other statistics yet to be introduced) can be accomplished at several levels of technical sophistication. As we have seen, a table of the normal distribution is sufficient in principle. Computer programs have been developed specifically for this job and are much more convenient when the amount of data to be analyzed is large. Appendix 6 contains one

such program; it uses the most accurate algebraic approximation to z, according to Brophy (1985). A more complex program, which can also be used for the discrimination paradigms to be introduced later in the book, is *d' plus* (Macmillan & Creelman, 1997), which is available on the Internet.[3]

It is also easy to find d' using the "inverse normal" functions of spreadsheet programs; this is especially appealing for the many laboratories in which the data are collected or stored into spreadsheets. Basic calculations are illustrated in Table 1.1 for Excel, but are very similar in QuattroPro and other programs. The function z is written = NORMSINV. The indexes to be entered or computed are listed in Column A, and formulas are given that can be inserted in Rows 5 to 11 of Column B, then copied to subsequent columns. Sorkin (1999) explored the use of spreadsheets for SDT calculations in greater detail.

Detection theory procedures are also available as part of standard statistical packages such as Systat and SPSS. Because many users of detection theory make routine use of such packages, this is an attractive option. Data can be entered either as frequencies (number of hits, number of misses, etc.) or trial by trial, as they would be collected in an experiment. These packages can also be used when there are more than two response alternatives; we discuss them further in the context of rating designs (chap. 3).

TABLE 1.1 *Formulas for Spreadsheet (Excel) Calculation of SDT Statistics With Examples*

	A (Labels Only)	*Formula (for Column B; Then Copy to C and Other Columns)*	*B (Set 1)*	*C (Set 2)*
1	# hits		10	9
2	# misses		0	1
3	# false alarms		2	0
4	# correct rejections		8	10
5	*H* (hit rate)	= IF(B2>0, B1/(B1 + B2), (B1 – 0.5)/(B1 + B2))	.950	.900
6	*F* (false-alarm rate)	= IF(B3>0, B3/(B3+B4), 0.5/(B3+B4))	.200	.050
7	$z(H)$	= NORMSINV(B5)	1.645	1.282
8	$z(F)$	= NORMSINV(B6)	–0.842	–1.645
9	d'	= B7 – B8	2.486	2.926
10	c	= –0.5*(B7 + B8)	–0.402	0.182
11	β	=EXP(B9*B10)	0.368	1.702

[3]The site is http://psych.utoronto.ca/~creelman/.

Essay: The Provenance of Detection Theory

Psychophysics, the oldest psychology, has continually adapted itself to the substantive concerns of experimentalists. In particular, detection theory is well suited to cognitive psychology and might indeed be considered one of its sources. No grounding in history is needed to use this book, but some appreciation of the intellectual strains that meet here will help place these tools in context.

The term *psychophysics* was invented by Gustav Fechner (1860), the 19th-century physicist, philosopher, and mystic. He was the first to take a mathematical approach to relating the internal and external worlds on the basis of experimental data. Some present-day psychophysicists directly pursue Fechner's interest in relating mental experience to the physical world, usually in simple perceptual experiments. Measuring the way in which the reported experience of loudness grows with physical intensity is a psychophysical problem of this sort; we consider a detection theory approach to this problem in chapter 5.

This book is part of a second Fechnerian legacy, also methodological, but more general than the first. Fechner developed, tested, and described experimental methods for estimating the *difference threshold*, or *just noticeable difference (jnd)*, the minimal difference between two stimuli that leads to a change in experience. Fechner's assumption that the jnd could be the unit of measurement, the fundamental building block or atom of experience, was central to Wundt's and Titchener's structuralism, the first experimentally based theory of perception. The analogy to 19th-century chemistry was close: Theory and experiment should focus on uncovering the basic units and the laws of combination of those units.

Fechner's methods were adopted and became topics of investigation in their own right; they still form the backbone of experimental psychology. Attempts to measure jnds led to two complications: (a) The threshold appeared not to be a fixed quantity because, as the difference between two stimuli increases, correct discrimination becomes only gradually more likely (Urban, 1908); and (b) different methods produced different values for the jnd.

The concept of the jnd survived the first problem by redefinition: The jnd is now considered to be the stimulus difference that can be discriminated on some fixed percentage of trials (see chaps. 5 and 11). Two early reactions to the problem of continuity in psychophysical data are recognizable in modern research (see Jones, 1974).

One line of thought retained the literal notion of a sensory threshold, building mechanical and mathematical models to explain the gradual nature of observed functions (see chap. 4 for the current status of such models). The threshold idea was congenial with early 20th-century behaviorist and operationist attitudes: Sensory function could be studied and measured without invoking unpopular notions of mental content (Garner, Hake, & Eriksen, 1956). The threshold, in this view, was a construct derived from data and did not have to relate to any internal and unobservable mental process. The solution to method dependence was merely to subscript thresholds to indicate the method by which they were obtained (Graham, 1950; Osgood, 1958).

The second response to the variability problem, instigated, according to Jones (1974), by Delboef (1883), substituted a continuum of experience for the discrete processes of the threshold; it is this view that informs most contemporary psychophysics. One approach to measuring such continuous experience was Stevens' (1975) magnitude estimation, which used direct verbal estimates. Detection measurement, in contrast, relies on underlying random variation or noise. Psychologists' realization of the importance of random variation dates at least to Fullerton and Cattell (1892), who invoked it in a rigorous quantitative way to account for inconsistency in response with repetitions of identical stimuli. Variability later served as the key building block for the pioneering work of Thurstone (1927a, 1927b) in measuring distances along sensory continua indirectly.

The idea of variability or noise as an explanatory concept also arose in engineering, with the development and evaluation of radar detection apparatus. Radar and sonar are limited in performance by intrinsic noise in the input signal. Any input from an antenna or sensor can be due to noise alone or to a signal of interest embedded in the background noise. Groups at the University of Michigan (Peterson, Birdsall, & Fox, 1954), MIT (van Meter & Middleton, 1954), and in the Soviet Union (Kotel'nikov, 1960) recognized that the physical noise that was mixed with all signals, and that could mimic signal presence, was a major limitation to detection performance.

Knowing that stimulus environments are noisy does not, in itself, tell an observer how best to cope with them. An approach to this problem was contributed by another applied science: statistical decision theory. Decision theorists pointed out that information derived from noisy signals could lead to action only when evaluated against well-defined goals. Decisions (and thus action) should depend not only on the stimulus, but on the expected outcomes of actions. The viewer of a radar display that might or might not

contain a blip, for example, should consider the relative effects of failing to detect a real bomber and of detecting a phantom before deciding on a response to that display.

W. P. Tanner, Jr., working with J. A. Swets at the University of Michigan, realized that these engineering notions could be applied to psychology and appropriated them directly into the psychophysical experiment (Tanner & Swets, 1954). By separating the world of stimuli and their perturbations from that of the decision process, detection theory was able to offer measures of performance that were not specific to procedure and that were independent of motivation. Procedure and motivation could influence data, but affected only the decision process, leaving measurable aspects of the internal stimulus world unchanged and capable of being evaluated separately.

According to detection theory, the observer's access to the stimuli being discriminated is indirect: An intelligent, not entirely reliable process makes inferences about them and acts according to the demands of the experimental situation. One might say that detection theory "deals with the processes by which [a decision about] a perceived, remembered, and thought-about world is brought into being from [an] unpromising beginning" (Neisser, 1967, p. 4). Neisser's landmark book linked perception and cognition into a unified framework after a hiatus of many decades. The constructionist (although not complicated) decision processes of detection theory mark it as an early example of cognitive psychology. The ideas behind detection theory are the everyday assumptions of behavioral experimenters in the cognitive era, and the theory itself is central to a wide range of research areas in cognitive science. Perhaps Estes' (2002) assessment is not an overstatement: "... [SDT is] the most towering achievement of basic psychological research of the last half century" (p. 15).

Summary

The results of a one-interval discrimination experiment can be described by a hit and a false-alarm rate, which in turn can be reduced to a single measure of sensitivity. Good indexes can be written as the difference between the hit and false-alarm rates when both are appropriately transformed. The sensitivity measure proposed by detection theory, d', uses the normal-distribution z transformation. The primary rationale for d' as a measure accuracy is that it is roughly invariant when response bias is manipulated; simpler indexes such as proportion correct do not have this property. The use of d' implies a model in which the two possible stimulus classes lead to normal

distributions differing in mean, and the observer decides which class occurred by comparing an observation with an adjustable criterion.

Conditions under which the methods described in this chapter are appropriate are spelled out in Chart 2 of Appendix 3.

Problems

1.1. Suppose you are measuring the sensitivity of a polygraph ("lie detector"). What are "hits," "misses," "false alarms," and "correct rejections"?

1.2. The following tables give the number of trials in three conditions of a detection experiment on which participants responded "yes" or "no" to S_1 or S_2. (a) Calculate H and F. (b) Find $H - F$, $p(c)$, and $p(c)$*. For these data sets, can $H - F$ be greater than $p(c)$ in one case and the reverse ordering occur in another, or is one index *always* greater than the other?

(a)	"yes"	"no"
S_2	9	6
S_1	7	8

(b)	"yes"	"no"
S_2	55	45
S_1	5	25

(c)	"yes"	"no"
S_2	45	55
S_1	25	5

1.3 (a). In Problem 1.2(a), the numbers of S_1 and S_2 trials are equal, but in (b) and (c) they are not. Does this matter computationally? experimentally?

(b). Is it possible to calculate $p(c)$ for S_2 trials only? What would this statistic measure?

1.4. Compute d' for the following (F, H) pairs:

(a) (.16, .84), (.01, .99), (.75, .75).

(b) (.6, .9), (.5, .9), (.05, .9).

1.5 (a). If $p(c) = .8$ and H and F are unknown, estimate d'.
(b). If $p(c) = .8$, the numbers of S_1 and S_2 trials are equal, and $F = .05$, find H and d'.

1.6 (a). Suppose $d' = 1$. What is H if $F = .01, .1, .5$?
(b) Plot the ROC from these points on linear and z coordinates, and use the zROC to confirm the value of d'.

1.7. For the data matrixes of Problem 1.2, find d' from H and F and also from $p(c)$. Is there a pattern to the results?

1.8. Are the points (.3, .9) and (.1, .7) on the same ROC according to detection theory (i.e., do they imply the same value of d')? Do they imply the same value of $p(c)$?

1.9. Suppose $(F, H) = (.2, .6)$. If F is unchanged, what would H have to be to double the participant's sensitivity, according to detection theory? If H is unchanged, what would F have to be?

1.10. Plot the ROCs implied by the following measures, on both linear and z coordinates: $H^2 - F^2$, $H^{1/2} - F^{1/2}$, H/F^2, H^2/F. Which measures are best? worst?

1.11. Suppose a face-recognition experiment yields 20 hits and 10 false alarms in 45 trials. Can you compute d'? If not, is it possible to narrow down the possibilities? *Hint*: The stimulus–response matrix looks like this:

20		
10		
		45

What happens if there are 0 misses, or 0 correct rejections?

2

The Yes-No Experiment: Response Bias

In dealing with other people, "bias" is the tendency to respond on some basis other than merit, showing a degree of favoritism. In a correspondence experiment, *response bias* measures the participant's tilt toward one response or the other.

The sensitivity measure d' depends on stimulus parameters, but is untainted by response bias: To a good approximation, it remains constant in the face of changes in response popularity. We now adopt the complementary perspective, seeking an index of response bias that is uncolored by sensitivity. Conceptually, d' corresponds to a fixed aspect of the observer's decision space, the difference between the means of underlying distributions; a measure of bias should also reflect an appropriate characteristic of the perceptual representation. How can we assign a value to the participant's preference for one of the two responses?

Two Examples

Example 2a: Face Recognition, Continued

Consider again the face-recognition experiment of chapter 1, in which viewers discriminated Old from New faces. Suppose the investigator now repeats the experiment, this time hypnotizing the participants in an effort to improve their memory, and obtains the following results from a representative observer:

	Normal		*Hypnotized*	
	"Yes"	"No"	"Yes"	"No"
Old	69	31	89	11
New	31	69	59	41

Applying the analyses of chapter 1 reveals that hypnosis has not affected sensitivity: d' is approximately 1.0 in both the normal and hypnotized conditions.

Hypnosis *does* appear to affect willingness to say "yes"; there are many more positive responses in the hypnotized condition than in the control data. (For a discussion of whether hypnotism actually has this effect, see Klatzky & Erdelyi, 1985.) In this example, therefore, an experimental manipulation affects bias, but not sensitivity. In the next example, a single variable affects both.

Example 2b: X-ray Reader Training

Apprentice radiologists must be trained to distinguish normal from abnormal X-rays (see Getty, Pickett, D'Orsi, & Swets, 1988, for a description of one training program). In this field, a hit is conventionally defined to be the correct diagnosis of a tumor from an X-ray, and a false alarm is the incorrect labeling of normal tissue as tumorous. Consider three readers who before training are equally able to distinguish X-rays displaying real tumors from X-rays of normal tissue, attaining exactly the same performance, but emerge from training with different scores on a posttest:

	Before Training	*After Training*
Trainee 1	$H = .89$	$H = .96$
	$F = .59$	$F = .39$
Trainee 2	$H = .89$	$H = .993$
	$F = .59$	$F = .68$
Trainee 3	$H = .89$	$H = .915$
	$F = .59$	$F = .265$

The trained readers are more sensitive—two of them show both a higher proportion of hits and a lower proportion of false alarms than before training. But has there also been a change in willingness to say "yes"? In the hypnotic recognition experiment, a response bias change merely masked the constancy of sensitivity; in this second example, there is clear evidence for a sensitivity change, but an interesting response-bias question remains.

Measuring Response Bias

Characteristics of a Good Response-Bias Measure

Because a response-bias index is intended to measure the participant's willingness to say "yes," we expect it to depend systematically on both the hit

and false-alarm rates and in the same direction—either increasing or decreasing in both. Sensitivity measures, remember, increase with H and decrease with F, an analogous property. A response-bias index should depend on the *sum* of terms involving H and F, whereas the sensitivity statistic d' depends on the *difference* of H and F terms.

Response-bias statistics can reflect either the degree to which "yes" responses dominate or the degree to which "no" responses are preferred. All the measures in this book index a leaning in the same direction: A positive bias is a tendency to say "no," whereas a negative bias is a tendency to say "yes." The rationale for these apparently illogical pairings will become clear when we discuss the representation.

Criterion Location (c)

The basic bias measure for detection theory, called c (for *criterion*), is defined as:

$$c = -\tfrac{1}{2}[z(H) + z(F)] \ . \tag{2.1}$$

When the false-alarm and miss rates are equal, $z(F) = z(1 - H) = -z(H)$ and c equals 0. Negative values arise when the false-alarm rate exceeds the miss rate, and positive values arise when it is lower. Extreme values of c occur when H and F are both large or both small: If both equal .99, for example, $c = -2.33$, whereas if both equal .01, $c = +2.33$. The range of c is therefore the same as that of d', although 0 is at the center rather than an endpoint. Figure 2.1 shows the locus of positive, negative, and 0 values of response bias in the part of ROC space where sensitivity is above chance.

Table A5.1, which was introduced in chapter 1 as a tool for calculating d', can also be used to compute the bias measure c. Spreadsheets accomplish the table-lookup task automatically (see Table 1.1, which includes some bias measures). Analyzing the face-recognition results, we find that c shifts from 0 to –0.73 under hypnosis, reflecting an increase in "yes" responses.

To interpret these numbers according to our model, consider the decision space in Fig. 2.2. The familiarity decision axis is labeled in standard deviation units, 0 being the point midway between the two distributions. Because $d' = 1.0$, the mean of the Old distribution is at 0.5, the mean of the New at –0.5. The participant's decision rule is to divide the familiarity axis into "yes" and "no" regions at a *criterion*.

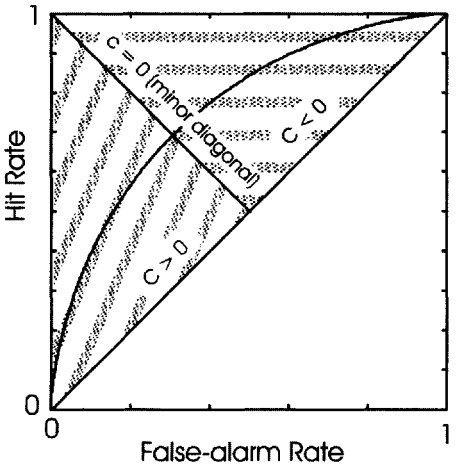

FIG. 2.1. The representation of criterion location in ROC space. Points in the shaded regions arise from criteria that are positive (below the minor diagonal) and negative (above the minor diagonal). Points in the unshaded region below the major diagonal result from negative sensitivity.

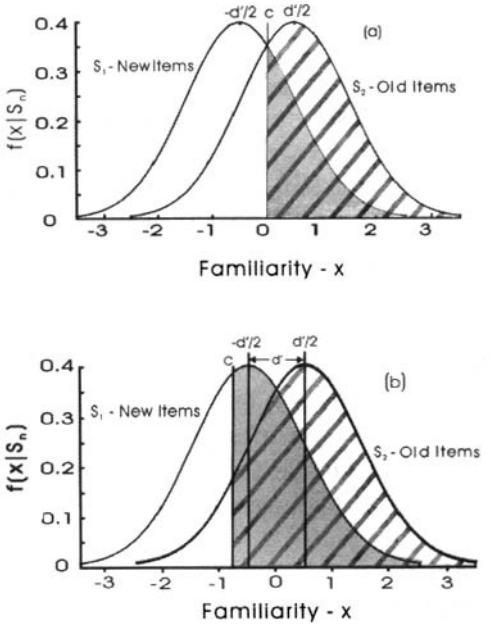

FIG. 2.2. Decision spaces for the Normal and Hypnotized conditions of Example 2a, according to SDT. Shaded area corresponds to *F*, diagonally striped area to *H*. (a) Normal controls have a symmetric criterion, $d' = 1.0$. (b) Hypnotized participants display identical sensitivity but a lower criterion, and thus have higher hit and false-alarm rates.

A simple calculation shows that the value of this criterion, in standard deviation units from the midpoint, is the bias parameter c. In chapter 1, we saw that the z score of the "yes" rate corresponds to the mean-minus-criterion distance. For the S_1 distribution, this implies

$$-d'/2 - c = z(F)\,, \tag{2.2a}$$

and for the S_2 distribution

$$d'/2 - c = z(H)\,. \tag{2.2b}$$

Adding these two equations produces Equation 2.1.

The different values of response bias in the normal and hypnotized conditions of our face-recognition experiment, therefore, correspond to different criterion locations. In the control condition (Fig. 2.2a), the criterion is located at 0, exactly halfway between the two distributions, and the participant is said to be "unbiased." Under hypnosis (Fig. 2.2b), the participant's criterion is much lower, below the mean of the New distribution. Because it is 0.73 standard deviations below the zero-bias point, $c = -0.73$.

Analysis of the radiology training data from Example 2b is equally straightforward. All trainees improve in sensitivity: d' about doubles. Values of c can be calculated from Equation 2.1. Trainee 1 maintains the same criterion location after training as before ($c = -0.74$). Trainee 2 has a more extreme bias (–1.46), and Trainee 3 has a less extreme one (–0.37). The degree to which the criteria differ among trainees is easily seen in Fig. 2.3, which shows the decision space and criterion settings for each reader: The first row represents the pretraining decision space of all trainees, and the other rows represent the posttraining spaces of each one individually.

Alternative Measures of Bias

Detection theory offers one measure of sensitivity (for two-response experiments), but is more generous with bias parameters. Besides criterion location, just described, bias can be specified by *relative criterion* location and *likelihood ratio*.

Relative Criterion Location (c′)

In this measure of bias, we scale the criterion location relative to performance. A rationale for such scaling is that with easier discrimination tasks a

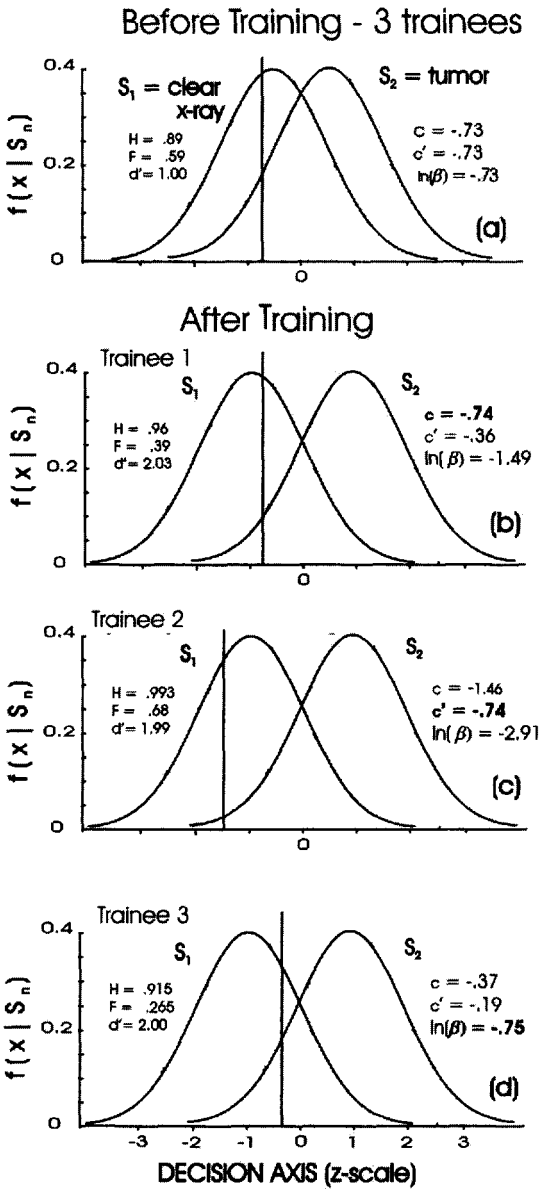

FIG. 2.3. Decision spaces for the three radiology trainees of Example 2b. In each case the hit rate, false-alarm rate, sensitivity, and three alternative criterion measures are shown. (a) Before training, $d' = 1.0$. The criterion c, the relative criterion c', and log likelihood ratio equal –0.73 for all trainees. (b) Trainee 1, after training; increased sensitivity and approximately the same criterion location c as before training. (c) Trainee 2, after training; increased sensitivity and approximately the same relative criterion location c' as before training. (d) Trainee 3, after training; increased sensitivity and approximately the same value of log likelihood ratio [$\ln(\beta)$] as before training.

more extreme criterion (as measured by c) would be needed to yield the same amount of bias.

Look again at the radiography training data of Example 2b. The first radiologist's criterion location is indeed the same distance from 0 (the equal-bias point) before and after training, but whether this is to be called "no change" can be argued. The criterion was initially below the mean of the S_1 distribution, but is above it afterward. If distance from the criterion to a distribution mean is the key to bias, this observer's bias has become less extreme. Would it not be sensible to calculate the criterion distance as a proportion of sensitivity distance? The alternative bias measure suggested by this reasoning is:

$$c' = \frac{c}{d'} = -\frac{1}{2}\frac{[z(H)+z(F)]}{[z(H)-z(F)]} . \tag{2.3}$$

Calculated values for c' are given in Fig. 2.3. It happens in this example that before training, $c = c'$, but only because $d' = 1.0$. After training, c' is half the magnitude of c because $d' = 2$. When d' varies, one must decide whether in discussing "bias" one wishes to take account of sensitivity. Of the three radiologists, it is Trainee 2 who maintains the same bias in the sense of c' and Trainee 1 whose bias is unchanged in the sense of c.

Likelihood Ratio (β)

The third measure of bias is found by an apparently different strategy. In the decision space, each value x on the decision axis has two associated "likelihoods," one for each distribution. Each likelihood is the height of one of the distributions; we denote this height at the location x by $f(x)$, and to distinguish the two distributions we refer to the heights of S_1 and S_2 as $f(x|S_1)$ and $f(x|S_2)$. The relative likelihood of S_2 versus S_1, obtained by dividing these, is called the *likelihood ratio*:

$$\mathrm{LR}(x) = f(x|S_2)/f(x|S_1) . \tag{2.4}$$

Each point x has an associated value of likelihood ratio: It is 1.0 at the center (where the two distributions cross), greater than 1.0 to the right, and between 0 and 1.0 to the left. One measure of response bias, therefore, is the value of likelihood ratio at the criterion.

Equation 2.4 suggests an interesting interpretation of likelihood ratio in terms of the ROC. Consider two points very close together on the decision

axis—imagine they are a small value ε units apart, as shown in Fig. 2.4a. The change in the hit rate between the two points is approximately $f(x|S_2)\varepsilon$, the height of the S_2 distribution multiplied by the width of a tiny rectangle. The change in the false-alarm rate, by the same token, equals $f(x|S_1)\varepsilon$. The ratio of these changes, which is the slope of the ROC, is $f(x|S_2)/f(x|S_1)$. Notice that this slope exactly equals the likelihood ratio. The assertion in chapter 1 that the slope of the ROC continuously decreases follows from the equivalence of likelihood ratio and ROC slope. As the criterion goes from large to small values of c, the likelihood ratio must decrease, and so therefore must the slope.

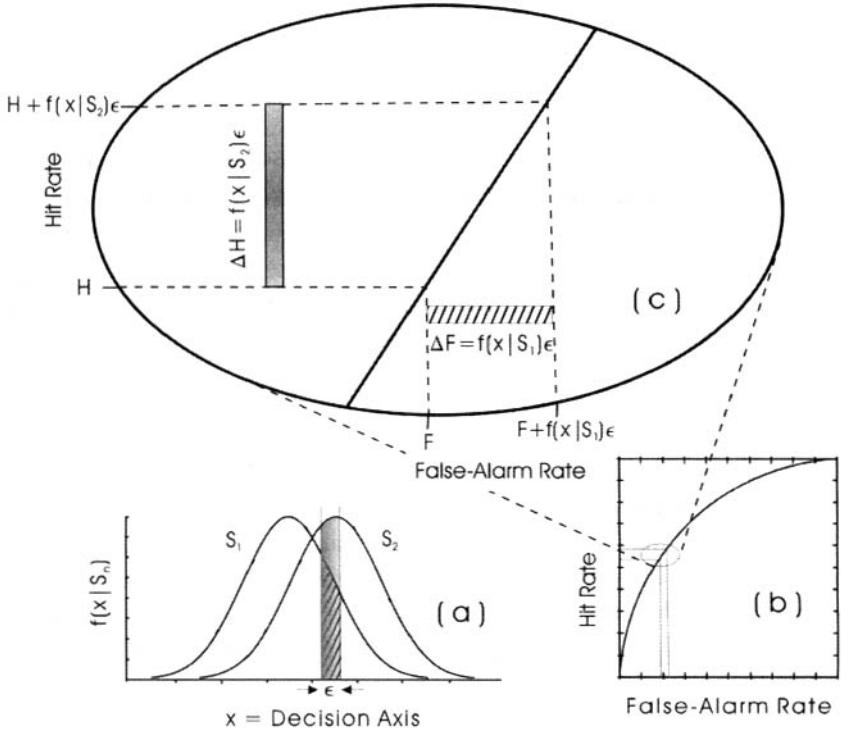

FIG. 2.4. Geometric demonstration that the slope of the ROC at any point is the likelihood ratio at the criterion value that yields that point. (a) In the decision space, two criteria are shown that differ by a small amount ε. For the lower criterion, the hit rate is greater by an amount equal to the area of the filled rectangle, and the false-alarm rate is greater by an amount equal to the area of the diagonally shaded rectangle. (b) The two criteria correspond to two points on an ROC curve. (c) An expanded view of the relevant section of the ROC. The lower point (higher criterion) is (F, H). At the higher point, the hit and false-alarm rates increase by the areas of the rectangles in (a). The slope of the ROC, the ratio of these two increments, is $f(x|S_2)/f(x|S_1)$, which is the likelihood ratio.

This conclusion does not depend on any assumptions about the shape of the underlying distributions, but actual calculation of likelihood ratio does require such a commitment. In the normal-distribution model we have been exploring, the height of the likelihood function, denoted ϕ, depends on x and on the distribution's mean μ and standard deviation σ according to the equation

$$\phi(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} . \tag{2.5}$$

Values of ϕ are given in Table A5.1.

The general strategy for finding the likelihood ratio can now be applied to the normal model. The likelihood function f in Equation 2.4 equals ϕ, and the likelihood ratio is the ratio of two values of $\phi(x)$—one for the S_2 distribution and one for S_1. A little calculation (to be found in the Computational Appendix) shows that the likelihood ratio, usually called β in the normal model, depends on sensitivity and the criterion location in a simple way:

$$\beta = e^{cd'}.$$

An equivalent form can be found by taking logarithms:

$$\ln(\beta) = cd' = -\tfrac{1}{2}[z(H)^2 - z(F)^2] . \tag{2.6}$$

Likelihood ratio can be calculated either directly from likelihoods (given by Eq. 2.5 or Table A5.1) or from its relation to c and d' (Eq. 2.6). For Trainee 3, the likelihood ratio equals $\phi(.915)/\phi(.265) = .1556/.3276 = 0.475$, and $\ln(\beta) = -0.75$. Alternatively, because $d' = 2.00$ and $c = -0.373$ for this observer, $\ln(\beta) = cd' = -0.75$. By this measure, Trainee 3 maintained the same response bias before and after training, whereas the other trainees adopted more extreme criteria (–1.49 and –2.91) after training. Summarizing bias using β [or $\ln(\beta)$] leads us to a different conclusion about our radiologists than did c or c'.

Isobias Curves

The isosensitivity curve, which describes the relation between the hit and false-alarm rates when bias (but not sensitivity) changes, is useful in evalu-

ating measures of accuracy. A function relating H and F for changing sensitivity (but not bias) is equally important in understanding bias statistics. The locus of points in ROC space that reflect equal bias is called an *isobias curve*.

For a particular bias parameter, the isobias curve is a prediction about how performance changes when bias is held fixed while sensitivity varies. Consider Fig. 2.5, which locates the performance of all the radiologists of Example 2b in ROC space. Trainee 1, remember, displayed the same value of c before and after training, so the points B and T_1 lie on the isobias contour for c defined by $c = -0.73$. Other points for which c takes on this value are connected by a continuous curve. Similarly, Trainees 2 and 3 generate two points on the isobias curves for c' and β, respectively. Clearly the three measures predict very different patterns of performance when sensitivity changes and bias remains the same.

To derive the form of an index's isobias curve, it is necessary to solve the equation defining the measure for H as a function of F. For example, the isobias function for c is found from Equation 2.1 to be

$$z(H) = -2c - z(F) . \tag{2.7}$$

As can be seen in Fig. 2.6 (upper right panel), this relation is a straight line in z coordinates. Families of curves for all three measures are shown in Fig. 2.6, in both linear and z coordinates.

Comparing the Bias Measures

How can a choice be made among the bias statistics available? The three bias measures, all quite plausible, are simply related. The criterion location relative to the zero-bias point, c, is *divided* by d' to obtain the relative bias c', and *multiplied* by d' to obtain the likelihood ratio measure $\ln(\beta)$. Because the logarithm is a monotonic function, $\ln(\beta)$ is equivalent to likelihood ratio itself.

Likelihood ratio is the most general of these three concepts: Unlike absolute or relative criterion location, it is meaningful for representations of any complexity. Early writers on detection theory (Licklider, 1959; Peterson, Birdsall, & Fox, 1954), therefore, placed great stress on the likelihood ratio as the basis for decision. When sensitivity is constant, d' serves as an arbitrary scale factor on the interval-scaled decision axis, and one may fairly say that log likelihood ratio *is* the decision variable.

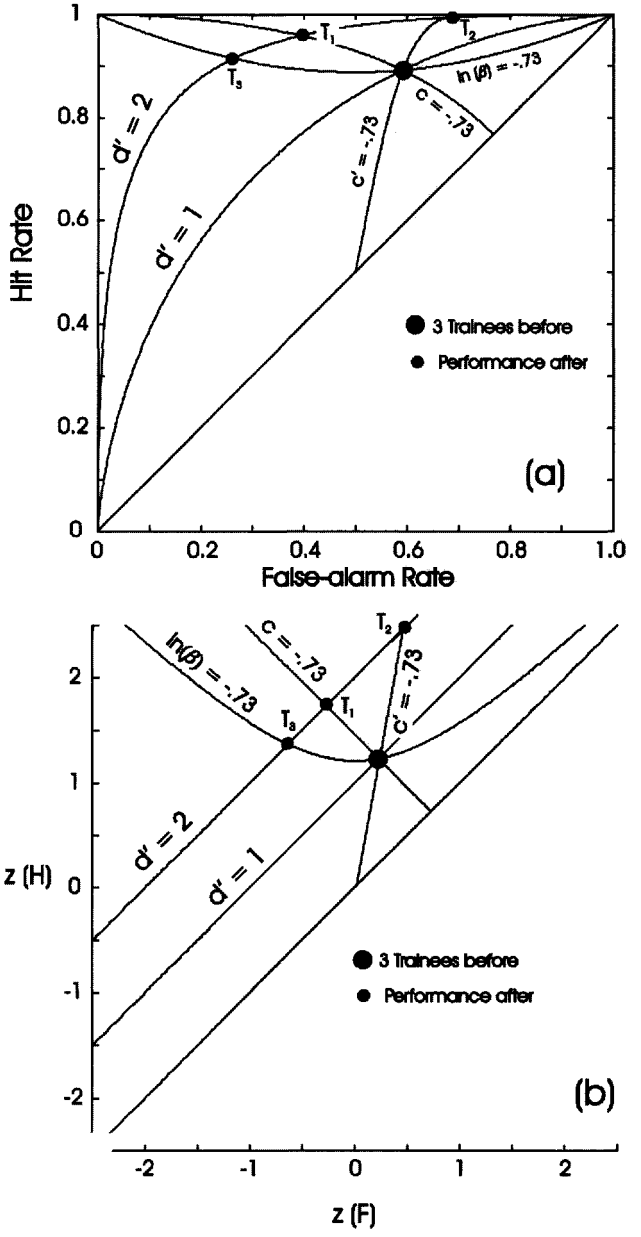

FIG. 2.5. Two ROCs and three isobias curves for the data of Example 2b. One ROC describes the sensitivity for all three trainees before training ($d' = 1$), the other sensitivity after ($d' = 2$). The isobias curves are for constant c (Trainee 1), constant c' (Trainee 2), and constant β (Trainee 3). Linear coordinates are used in (a), z coordinates in (b).

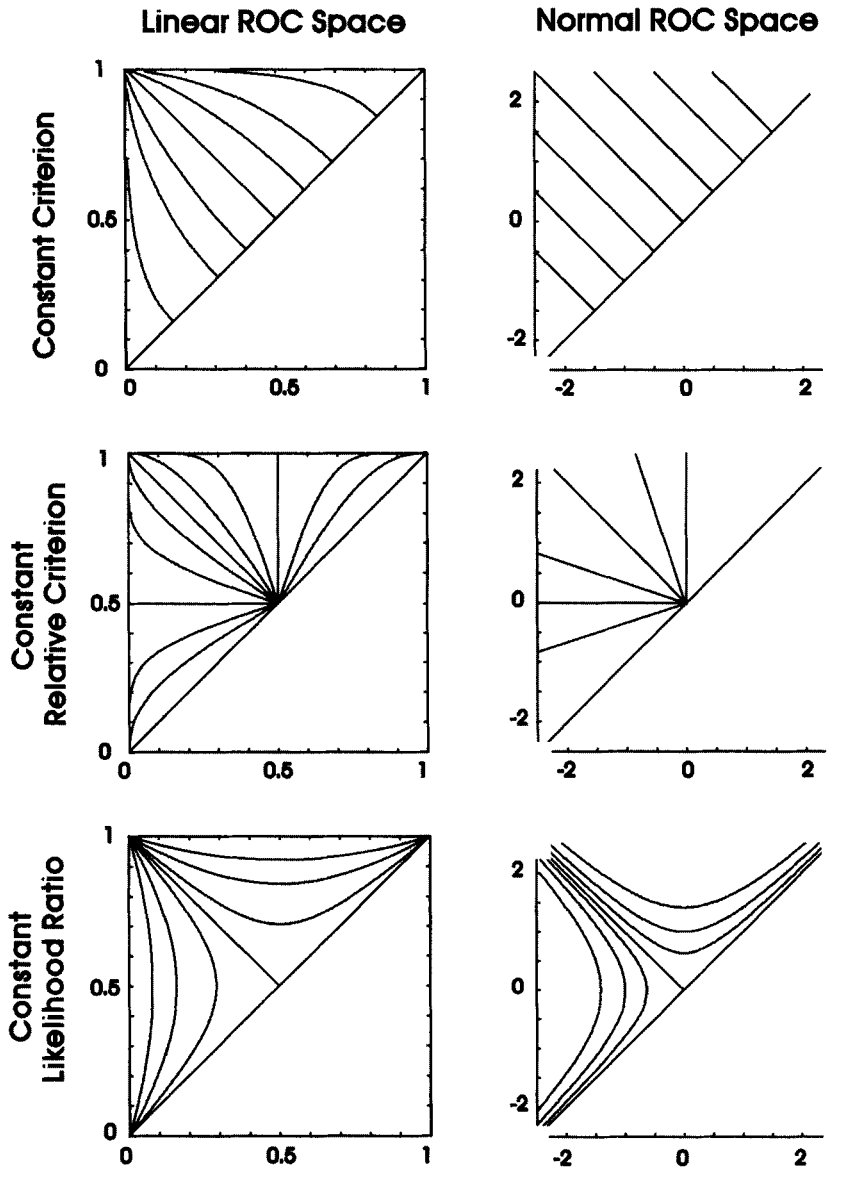

FIG. 2.6. Families of isobias curves (hit rate vs. false-alarm rate, d' varying) for constant criterion c, relative criterion c', and likelihood ratio β, on linear and z axes.

But this does not mean that $\ln(\beta)$ is necessarily to be preferred as a response bias measure for comparing experimental conditions—when detectability is constant, any measure will do. In Example 2a, d' equals 1.0 in both face-recognition conditions; thus, the bias parameters c, c', and $\ln(\beta)$ give the same values. If d' were constant but not equal to 1.0, and the three statistics therefore differed numerically, they would still lead to the same conclusion in any comparison between conditions. But when d' varies, as in Example 2b, the three measures of bias support different interpretations.

We consider three standards to which a candidate bias measure might be held: (a) Its isobias curve should be supported empirically, (b) it should depend monotonically on H and F in the same direction, and (c) it should be independent of the sensitivity index. The second and third standards favor the criterion location c; the first has not provided clear-cut support for any one measure.

Form of Empirical Isobias Curves

The curves in Fig. 2.6 are similar, but they differ substantially in certain parts of the space (at the corners and near the major diagonal), and the reader may eagerly expect that, as with sensitivity, we shall be able to decide among the implied measures on the basis of data. Dusoir (1975, 1983) compared isobias curves for c, c', β, and several other potential bias statistics with data from an auditory detection experiment. Sensitivity was varied by changing tone intensity and bias via instructions, allowing isobias curves to be constructed.

Dusoir found great individual differences in shapes of isobias curves and could not support any one measure as superior to any other. This lack of unanimity reflects an important asymmetry between sensitivity and bias: To derive measures that yield constant sensitivity requires, at most, a theory of the stimuli; to do the same for bias indexes requires, at least, a theory of the instructions. Theories of tone detection, and even theories of memory, are a good deal more advanced than the kind of theory of language understanding needed to predict isobias performance. Dusoir (1983) concluded that participants may vary in their understanding of bias-inducing instructions so that different observers may all be holding some—but not the same—bias parameter fixed.

Two more recent applications produced more internally consistent results, although they are not in agreement with each other. See, Warm, Dember, and Howe (1997) examined a "vigilance" situation in which signals occur infrequently and trials are not defined by the experimenter. Such

tasks are of interest as models of, for example, radar monitoring, and observers typically show a decrement in performance after less than 1 hour (Davies & Parasuraman, 1982). See et al. asked observers to detect small increases in the height of lines on a computer screen, and they manipulated response bias by varying the probability of signal occurrence and monetary payoffs. In a critical test, they chose two levels of "salience" (detectability), and were thus able to construct two-point isobias curves; these curves had the form predicted for c.

Recognition memory data relevant to the isobias question have been reported by Stretch and Wixted (1998), who reanalyzed the data of Ratcliff, McKoon, and Tindall (1994) and also conducted new experiments. Sensitivity was manipulated by varying the time or rate of exposure of words, or the number of times items were presented. Response bias was evaluated through a rating design; we discuss the details of this procedure in the next chapter. Although they did not plot isobias curves per se, their data come closest to the form predicted for β.

It is tempting to conclude that the bias manipulations used by See et al. (1997) and Stretch and Wixted (1998) are superior to the instructional method used by Dusoir (1983), at least in the sense that individual differences may be smaller. But changes in presentation probability have their own problems, as we shall see in chapter 3. In any case, these two studies reach different conclusions. Perhaps this is not something to worry about: Although one would like theories of response bias to be oblivious to the stimulus domain being studied, such a goal may be too optimistic.

Monotonicity of Theoretical Isobias Curves

Our general condition on bias measures, according to which an increase in either the hit rate or false-alarm rate should mean a decrease in bias, imposes a restriction on isobias curves: As F increases, H must decrease. All measures satisfy this condition in the upper left quadrant of ROC space, where $H \geq .5$ and $F \leq .5$, but both relative criterion and likelihood ratio violate it elsewhere.

Two other regions of ROC space, in which the curves of Fig. 2.6 show different behaviors, are the chance line $H = F$ and the area below it. When sensitivity is 0, it is still meaningful to talk about bias: An observer for whom $H = F = .1$ clearly has a different bias from one operating at $H = F = .9$. The criterion location c does take on different values along the diagonal,

but c' and β do not. In fact, when underlying distributions of likelihood are identical, both of these statistics are undefined.[1]

Below-chance behavior may seem uninteresting or even illogical, but statistical fluctuations can easily lead to such performance. Two points in ROC space that are symmetrically located across the chance line—(F, H) and (H, F)—should intuitively show the same or similar biases. By this test, criterion location is again the best measure, giving the same value for the two points. Both c' and $\ln(\beta)$ show discontinuities, changing sign as they cross the diagonal (see Macmillan & Creelman, 1990, for more detail). These measures can be salvaged by multiplying their values by –1 below the chance line [as has been suggested for $\ln(\beta)$ by Waldmann & Göttert, 1989, and for other, "nonparametric" bias measures by Aaronson & Watts, 1987, and Snodgrass & Corwin, 1988].

Independence of Bias From Sensitivity

That response bias be independent of accuracy is clearly a desirable outcome, but we must be careful what we wish for, because *independence* has multiple meanings. First, consider statistical independence, the condition that in repeated tests neither of two measures affects the other. Only c is independent of d' in this sense (see the Computational Appendix for proof that it is).

Second, we can examine the dependence of bias measures on stimulus strength—there should be none. An analogous strategy, finding noneffects of bias on sensitivity, provided some of the earliest support for the use of d'. Dusoir (1983) applied this test, but the results were inconclusive: Of 21 comparisons (each representing a single observer in a single experimental condition), c, c', and β were each significantly correlated with sensitivity 12 times. Other statistics considered by Dusoir (some of which we encounter in chap. 4) were at least equally unsuccessful. In the See et al. (1997) experiment, c showed a slight dependence on d' in one of three experiments, but the correlations between β and d' were both more widespread and stronger.

A final type of independence can be seen intuitively in the ranges of these variables: The range of c does not depend on d', whereas the range of the other measures does (Banks, 1970; Ingham, 1970). When d' is large, c' has a small range and β a large one; when d' is small, the reverse is true. The criterion location c is the only index whose magnitude can be interpreted with-

[1]In one interpretation, likelihood ratio equals 1 for the zero-sensitivity case. As noted earlier, however, the decision axis itself may be considered to be likelihood ratio, so the decision space collapses to a single point.

out knowledge of d' (Macmillan & Creelman, 1990; Snodgrass & Corwin, 1988). A caveat remains, however: Perhaps the range of biases is truly *not* the same at different levels. Thus, Stretch and Wixted (1998) concluded from the nature of the relation between criterion and stimulus strength in their memory experiment that the range of biases was narrower at high levels of accuracy.

How Does the Participant Choose a Decision Rule?

Whatever the best response bias measure is, the decision process leading to it is not in dispute.[2] The participant establishes a criterion at some point on a relevant internal dimension and uses it to partition the dimension into regions of "yes" and "no" responses. Two questions remain: (a) Is this always the best thing to do? If so, (b) where should the criterion be located?

As long as the stimuli (and thus sensitivities) are fixed, using a criterion to determine responses is, indeed, always the right strategy, and for an interesting reason. As we have seen, the decision axis is a monotonic function of likelihood ratio in the fixed-sensitivity case, so the question becomes whether it is optimal to use likelihood ratio to make decisions.[3] To answer this question requires consensus about what the "best" decision rule should accomplish, something about which reasonable people can agree. Green and Swets (1966: ch. 1) nominated four decision goals; for each a likelihood ratio decision rule is indicated and the optimal likelihood ratio at the criterion can be calculated:

1. *Maximize proportion correct.* When presentation probabilities are equal, a participant who wishes to maximize proportion correct must treat the two stimulus classes symmetrically, preferring to make neither false alarms nor misses more often. This is accomplished by setting c to equal 0, the zero-bias point. Likelihood ratio at this point is 1. If S_2 is presented more often than S_1, however, it will pay the participant to be more willing to respond "yes," and a lower criterion should be set. If $p(S_1)$ and $p(S_2)$ are the a priori probabilities of presenting the two stimuli, then the optimal value of likelihood ratio is $p(S_1)/p(S_2)$.

[2]Well, not in *much* dispute. One alternative interpretation of detection data rejects the whole idea of criterion shifts (Balakrishnan, 1999) in favor of changes in the distributions themselves.

[3]In chapter 3, we encounter a case in which the likelihood ratio is not monotonic with the decision axis; even then the likelihood ratio rule is best.

2. *Maximize a weighted combination of hits and correct rejections.* An observer may be more interested in hits than in false alarms, or vice versa, for reasons other than presentation probability. For example, the X-ray readers of Example 2b should be much more willing to make a false alarm (detecting a tumor when none is there) than a miss (failing to detect). To maximize a weighted average—say, three times the hit rate, plus the correct rejection rate—the observer should set a criterion at the "importance ratio," in this example, three. That is, only if the X-ray under examination is at least three times as likely to be normal as pathological should the observer say, "no, there is no tumor." If the importance ratio equals the ratio of presentation probabilities, this objective is the same as maximizing proportion correct.

3. *Maximize expected value.* The decision rule suggested for the X-ray reader, just above, was based on the relative value of the two kinds of correct decisions. This can be made explicit, at least in experimental situations: Participants can be rewarded for hits and correct rejections, or they can be penalized for false alarms and misses. In the laboratory, the rewards are sometimes small financial ones, sometimes merely "points" (see chap. 3).

The ideal value of likelihood ratio in such a situation depends on the reward function R that specifies the payoff for each experimental outcome:

$$LR(x) = \beta = \frac{[R(correct\ rejection) - R(false\ alarm)]}{[R(hit) - R(miss)]} \frac{p(S_1)}{p(S_2)} \quad (2.8)$$

Normally, the "rewards" for false alarms and misses are negative. If an observer is paid 10 cents for each correct response and is penalized 1 cent for misses and a dollar for false alarms, the optimal value for the criterion (assuming equal presentation probabilities) is $[0.10 - (-1.00)]/[(0.10 - (-0.01)] = 10$; that is, the observer should insist that the odds favoring S_2 given the data be 10 to 1 or larger before hazarding a "yes" response.

People rarely adopt such extreme criteria; when payoffs are changed to favor "no," criteria generally shift, but not to the degree prescribed by Equation 2.8. The theoretical analysis is sometimes salvaged by reference to "subjective" rewards, which are presumed to lag behind real ones. We are aware of no attempt to verify that the subjective criteria actually used by participants in payoff-driven experi-

ments are, in any sense, optimal. Many practical matters, such as participants' (usually negative) attitudes toward piece work, competition among participants, and a tendency to see performance (in what is frequently, after all, a professor's laboratory) as a measure of intelligence, all suggest that Equation 2.8 captures only some of the real basis for human decision making.

4. *Test a statistical hypothesis.* A decision maker is often instructed, explicitly or implicitly, to obtain as high a hit rate as possible while holding the false-alarm rate to some predetermined level, a goal called the *Neyman–Pearson objective*. Thus, our X-ray reader might be advised to keep the false-alarm rate below .5; for an air traffic controller, an acceptable value of F might also be quite high, because it is the misses—failures to notice impending collisions—that have to be minimized. Clients undergoing audiological testing often adopt a much more severe criterion, being unwilling to make more than a few false alarms, which they view as lies.

Satisfaction of the Neyman–Pearson objective also requires a likelihood ratio criterion decision rule, with the value of likelihood ratio set to produce the desired false-alarm rate. Jerzy Neyman and Karl Pearson were among the founding fathers of modern statistics, and their objective is exactly that met by conventional statistical hypothesis testing. False alarms in that context are called Type I errors, and the false-alarm rate is arbitrarily set to a small value, typically .05 or .01. Observations (sample means, sample mean differences, etc.) above the criterion lead to rejection of the null hypothesis, either correctly (hits) or, with fixed low probability, incorrectly (false alarms).

Coda: Calculating Hit and False-Alarm Rates From Parameters

The outcome of a yes-no discrimination experiment, we have seen, can be characterized by either of two pairs of parameters: the hit and false-alarm rates, or sensitivity and bias. Detection theory asserts that the latter pair is more illuminating. These first two chapters have therefore focused on expressions for sensitivity and bias in terms of H and F. When solved for H, these expressions describe isosensitivity and isobias curves.

It is sometimes useful, however, to reverse this process and calculate H and F from detection theory parameters. We do this here according to the following plan. In the decision space, the hit and false-alarm rates—both proportions of "yes" responses—correspond to the area under a probability

function above the criterion. The (cumulative) distribution function at the criterion gives the complementary probability of an observation falling below criterion. This distribution function can be easily calculated, because it is the inverse of the z transformation that converts proportions to distances. The calculation is illustrated in Fig. 2.7.

The z transformation converts a proportion to a standardized distance from the mean. The inverse of z, which gives the "no" rate when the criterion is at z, is the normal distribution function, denoted $\Phi(z)$. The value of $\Phi(z)$ can be found from a normal table, but Table A5.1 is not ideally arranged for this purpose. In that table, p values are given in units of .01, which is helpful when p is known, as in data analysis. Table A5.2 gives the same information, but for z scores in units of .01, which is more convenient when z is known. The probability p corresponding to a z score is $\Phi(z)$.

The "yes" rate is $1 - \Phi(z)$; because the normal distribution is symmetric, this equals $\Phi(-z)$. Expressed as a z score, the criterion equals $c - d'/2$ for the S_2 distribution and $c - (-d'/2)$ for S_1; so

$$H = \Phi(d'/2 - c) \tag{2.9}$$

$$F = \Phi(-d'/2 - c) .$$

For an unbiased observer, $c = 0$, $H = \Phi(d'/2)$, and $F = \Phi(-d'/2)$. In this case, the hit and correct rejection rates both equal proportion correct, so

$$p(c) = \Phi(d'/2) . \tag{2.10}$$

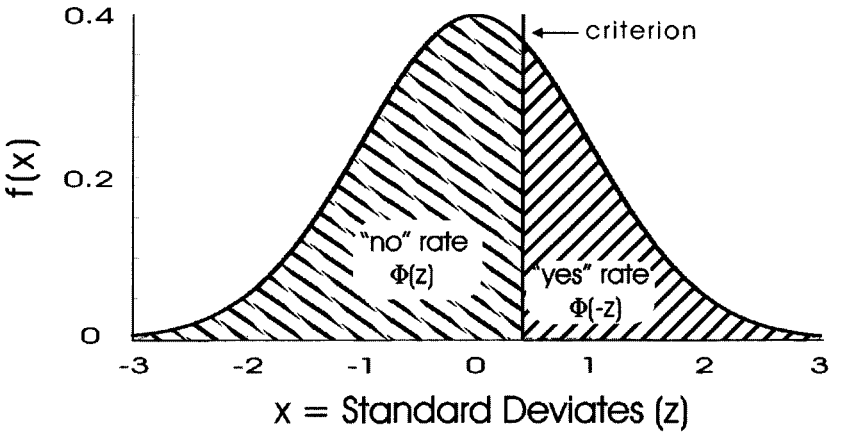

FIG. 2.7. Relation between underlying distributions and "yes" rates (hit and false-alarm rates). When the criterion is at z, the yes rate is $\Phi(-z)$.

Essay: On Human Decision Making

Much of the large literature on decision making by human beings (see, e.g., Kahnemann, Slovic, & Tversky, 1982) asks how closely our behavior corresponds to what we "should" do. The decision problem described in this chapter is in many ways rather simple: Only one dimension is relevant, the stimuli are presented at predictable times (in most applications), and repeated trials allow the observer to focus on relevant aspects of the stimulus display. Does the observer indeed deal with this problem in the "right" way—by establishing a criterion and using it?

At least two nonoptimal strategies have occurred to most psychophysicists who have studied (and, frequently, served as participants in) correspondence experiments: *inattention* and *inconsistency.* An inattentive observer dozes off, or at least drifts into reverie, on some proportion of trials; because failing to respond is usually discouraged, this leads to an unknown number of $d' = 0$ trials, ones on which the observer responds despite not having paid attention, mixed in with the others. An inconsistent participant uses a criterion, but changes the location of the cutoff from trial to trial; because the criterion must be compared to a sensory event, the movement adds an unknown amount of variance to the underlying distributions (Wickelgren, 1968). Both strategies, if they may be called that, serve to reduce observed sensitivity.

Do these effects occur? Almost certainly, but little is known about how badly they contaminate experiments. Training provided before experimental data are collected may serve to reduce these errors; observers who fail to improve during practice may be suspected of persisting in a nonoptimal strategy. In most applications, small amounts of inattention or inconsistency matter little. Stimulus pairs that yield high performance levels are an exception: The experimenter who wishes to make a precise estimate of a d' of 4 or so will be frustrated by even an occasional lapse. If lapses are part of the human condition, such estimates are doomed to unreliability.

We have been speaking of optimal strategies; what about optimal *use* of strategies? Given that an observer is using a criterion in the manner we suppose, are there ways we can encourage "unbiased" decision making, that is, symmetric criterion placement? Arguments are sometimes put forward that one or another experimental technique will accomplish this goal, which is sometimes a valuable one (see especially chap. 11). Often, however, there is no reason to aim for a symmetric criterion. After all, the sensitivity measure

with which detection theory provides us is unaffected by bias, so why worry? Perhaps only because in common parlance (but not in psychophysics) *bias* is a pejorative term, something worth avoiding.

Another appeal of unbiased responding is that it makes almost any measure of sensitivity satisfactory, eliminating the need for complex psychophysics. The search for unbiased responding may thus be a vestige of the belief that, really, simple, untransformed measures are to be trusted more than theoretical ones. We shall critically evaluate this possibility in chapter 4.

Finally, the concept of bias in detection theory has sometimes been misunderstood in a way that makes neutral bias qualitatively different from other values. The location of the criterion can, we have seen, be manipulated by instructions: Apparently, then, observers can consciously choose to change it. If no instructions are given, however, observers are not aware of the possibility of varying a criterion. Thus, the argument goes, instructions to change bias provide conscious interference with a normally unconscious process. In our view, the distinction between consciousness and its lack has nothing to do with either the existence or location of a criterion. Detection theory takes no stand on the conscious status of a criterion, and in any case observers do *not* naturally choose a neutral value. We shall encounter this issue again in chapter 10 when we briefly discuss the alleged phenomenon of subliminal perception. An observer who responds "no" when a stimulus is presented because of a high criterion is not necessarily aware of the possibility that a "yes" response would have been possible had the criterion been set lower.

Summary

Whereas a good sensitivity statistic is the difference between the transformed hit and false-alarm rates (chap. 1), a good measure of response bias is the sum of the same two quantities. In the decision space, this index describes the location of a criterion that divides the decision axis between values that lead to "yes" and "no" responses. Other measures—relative criterion and likelihood ratio—are equivalent when sensitivity is unvarying, but not when accuracy changes across conditions. Criterion location has advantages, both logically and, in some cases, empirically. Using a criterion to partition the decision axis is an optimal response strategy. The optimal location of the criterion can be calculated if the performance goal is specified.

Conditions under which the methods described in this chapter are appropriate are spelled out in Chart 3 of Appendix 3.

Computational Appendix

Derivation of Equation 2.6

The likelihood ratio is the ratio of the values of the S_2 and S_1 normal likelihood functions at the location $x = c$. The function $\phi(x)$ is defined by Equation 2.5. For both distributions, the standard deviation $\sigma = 1$; the mean μ equals d' in the numerator and 0 in the denominator. The resulting likelihood ratio, called β in SDT, is

$$\beta = e^{cd'}, \text{ or}$$

$$\ln(\beta) = cd' = -\tfrac{1}{2}[z(H) + z(F)][z(H) - z(F)] \qquad (2.11)$$

$$= \tfrac{1}{2}[z(H)^2 - z(F)^2] .$$

Statistical Independence of d' and c

That c is statistically independent of d' can be seen as follows: Hit and false-alarm rates are independently computed (from data on S_2 and S_1 trials, respectively); so $z(H)$ and $z(F)$ are independent, and thus uncorrelated across repeated estimates. The sum (c) and difference (d') of uncorrelated variables are also uncorrelated and, because $z(H)$ and $z(F)$ are normally distributed, independent. Neither c' nor β is independent of d' in this sense.

Problems

2.1. For the data of Example 2b, calculate $p(c)$ for each trainee. Do all readers improve, according to this measure?

2.2. For the data of Example 2b, suppose all trainees adopted symmetric criteria, both before and after training. (a) What values of $p(c)$ would they obtain? (b) How do c, c', and $\ln(\beta)$ compare?

2.3. For the matrixes of Problem 1.2, find c, c', and β.

2.4. (a) Suppose an investigator decides to use F itself to measure response bias. What is the isobias curve for this statistic? (b) Another simple statistic is the yes rate, $(H + F)/2$. Find the isobias curve.

2.5. Suppose $(F, H) = (.2, .6)$. If d' doubles and the observer maintains the "same bias," what will the new (F, H) point be? (Interpret "same

bias" to mean same criterion, same relative criterion, and same likelihood ratio; you will have three answers.)

2.6. Ekman, O'Sullivan, and Frank (1999) videotaped men either lying or telling the truth about social issues on which they held strong beliefs, and played the tapes to four groups of observers: trained interrogators, Los Angeles County sheriffs, clinical psychologists interested in deception, and academic psychologists. The proportions of correct responses to lies and truths were:

Experimental group	*P("lie"|lie)*	*P("truth"|truth)*
Interrogators	.80	.66
Sheriffs	.77	.56
Clinical psychologists	.71	.64
Academic psychologists	.57	.58

The authors concluded that "... the most accurate groups did especially well in judging the lies compared with the truths ...," and that this could not be attributed to response bias. What would a detection theory analysis have to say about bias in these four groups? About sensitivity?

2.7. Suppose an observer is paid 10 cents for all correct responses. (a) What does the payoff matrix look like? (b) What is the optimal value of likelihood ratio if the proportion of S_2 trials is .5? .25? .1? (c) If $d' = 1$, find the criterion location c for each case. (d) Again assume $d' = 1$. In each case, what would the hit and false-alarm rates be? (*Hint*: Use the "coda" section.)

2.8. In a grating detection experiment, observers try to distinguish the presence of a pattern of alternating light and dark stripes from a uniform grey patch. There are two experimental conditions, with the stimuli differing in *contrast*, such that the stripes are better defined in the high-contrast than the low-contrast condition. Two groups of participants each view both conditions, but differ in that one group is given feedback (told after their response whether the grating was present) and the other is not. For each set of observers, how would you expect criterion placement to differ with contrast? Would it make a difference if the high- and low-contrast gratings were presented in different blocks of trials, or mixed together in a single block?

2.9. In this chapter, an analogy between detection theory and conventional statistical hypothesis testing is presented. According to this

analogy, statistical results usually are summarized by a "false-alarm rate." Why are they not instead summarized by a sensitivity measure such as d'?

2.10. Suppose $d' = 2$. (a) What are H and F if $c = 0.5$? -0.5? What are H and F if $c' = 0.5$? -0.5?

3

The Rating Experiment and Empirical ROCs

In the last two chapters, we described correspondence experiments in which people report which of two events (such as seeing a New or Old face) had occurred. According to detection theory, they do this by comparing the strength of evidence, which we called *familiarity*, with a criterion. Observations of more than criterial familiarity are called "old," and those below criterion are called "new." The criterion is placed at a location of the observer's choice: Strict criteria serve to minimize false alarms, lax criteria to minimize misses.

If observers can set *different* criteria in different experimental conditions, they must know more about events in their experience than is needed to make a simple yes-no judgment. In this chapter, we see how observers can make graded reports about the degree of their experience by setting multiple criteria simultaneously. Our two primary examples are both tests of recognition memory, but for rather different materials: odors and words.

Design of Rating Experiments

Example 3a: Recognition Memory for Odors

How is memory for odors affected by the passage of time? Rabin and Cain (1984) presented participants with 20 odors to remember, then tested them at a delay of 10 minutes, 1 day, and 7 days. At each test, a different set of 20 New odors was intermixed with the Old stimuli.

Observers labeled each smell as "old" or "new" and also rated their confidence in these answers on a 5-point scale, which we have reduced to a 3-point scale for illustrative purposes. Thus, there are two kinds of stimuli