Chapter 4

Measures of detection performance

In the equal-variance Gaussian model, the distributions of X_n and X_s differ only by the parameter d'. This quantity easily summarizes detection performance. For the unequal-variance model, where the distributions differ in both mean and variance, finding a single measure of detectability is less obvious. The differences between the noise and signal configurations now involve two parameters, μ_s and σ_s^2. These cannot be summarized by one number. Any choice of a single detection index necessarily emphasizes some aspects of the difference and minimizes others. This chapter describes these alternatives. The first three sections present four measures of detection. No one of these measures clearly dominates the others in all situations. Each has its own advantages and liabilities, which are summarized, with recommendations, in Section 4.4. Although most studies are concerned primarily with measuring detectability, questions concerning bias also arise. Section 4.5 discusses the possible measures. Finally, although the prototypical signal-detection study is one in which extensive data are collected from a single observer, many studies are run in which smaller amounts of data are collected from many observers. The combination of results from different observers presents special problems that are treated in Section 4.6.

The issues discussed in this chapter apply to any study in which sufficient data are available to draw an operating characteristic. These data could come from manipulations of the bias in separate conditions or from the rating-scale procedure that will be discussed in Chapter 5. As an example for use in this chapter, Figure 4.1 gives the results of four conditions with responses from 200 trials in each condition. The figure also shows the operating characteristic in standard and Gaussian coordinates. A look at the transformed operating characteristic shows that the Gaussian model

Set	F.A.	Hits	f	h	$Z(f)$	$Z(h)$
1	33	81	0.165	0.405	−0.974	−0.240
2	55	121	0.275	0.605	−0.598	0.266
3	86	127	0.430	0.635	−0.176	0.345
4	115	150	0.575	0.750	0.189	0.674

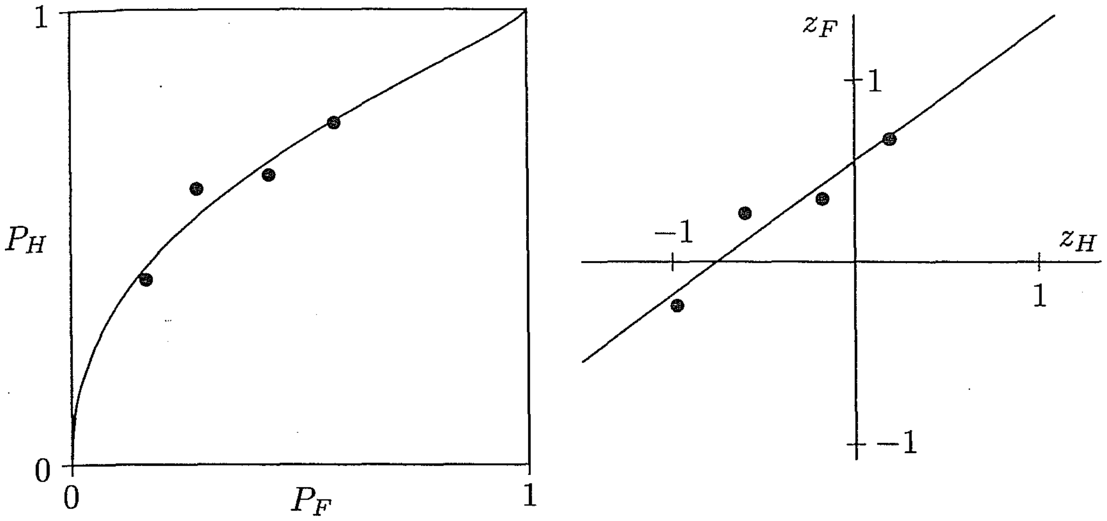

Figure 4.1: Detection performance in four bias conditions, each with 200 signal trials and 200 noise trials. The plots show the operating characteristic fitted from the unequal-variance Gaussian model. The fitted line, $z_H = 0.550 + 0.767 z_F$, crosses the axes at $x_0 = -0.747$ and $y_0 = 0.550$.

fits adequately (the points lie roughly on a straight line) and that the equal variance model is unsatisfactory (that line is not at 45°). The slope is less than 45°, which implies that $\sigma_s^2 < \sigma_n^2$. The issue now is how to find a number that measures the detectability for these data.

4.1 The distance between distributions

Any measure of detectability that integrates across conditions must be based, at least implicitly, on a theoretical representation that ties the observations together. For data such as those in Figure 4.1, the Gaussian model is a good choice. When this model has been fitted, a natural candidate for a measure of the difference between the two distributions of this model is the difference between their means, often denoted by[1]

$$\Delta m = \mu_s - \mu_n. \tag{4.1}$$

[1] The uppercase Greek delta here is a difference operator, and so Δm is not the product of Δ and m, but a difference between two values of the mean m (or μ).

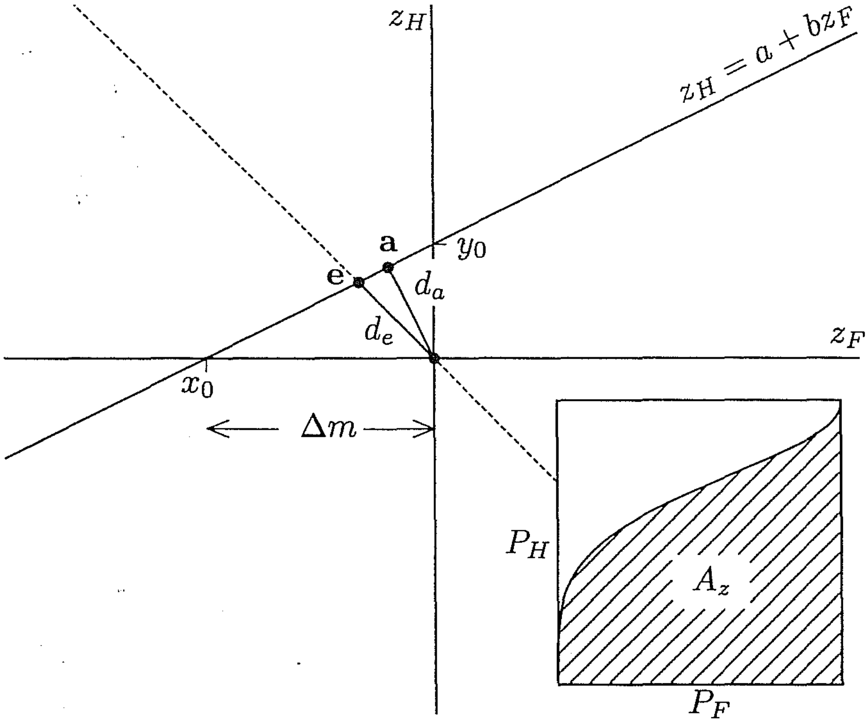

Figure 4.2: The detection measures shown as distances Δm, d_e, and d_a on the operating characteristic in Gaussian coordinates. The inset operating characteristic shows the area A_z under the curve in probability coordinates.

With the noise mean μ_n set to zero, Δm is the same thing as the signal mean μ_s, which, as the notation suggests, is interpreted as a difference between means. This measure is simple and is easily determined from a straight line fitted to a series of detection conditions in Gaussian coordinates. Figure 4.2 shows this measure (and other measures to be discussed later) for an unequal-variance operating characteristic, both in Gaussian and probability coordinates. The difference Δm equals that between the origin and the point x_0 on the horizontal axis where the operating characteristic crosses it (recall Equation 3.7 on page 54).

Equation 4.1 is a theoretical statement about the parameters of the model. A comparable equation expresses the relationship among the estimates of these quantities:

$$\widehat{\Delta m} = \widehat{\mu}_s - \widehat{\mu}_n.$$

Because μ_n is by definition zero, the relationship becomes simply $\widehat{\Delta m} = \widehat{\mu}_s$. The value of $\widehat{\Delta m}$ is taken directly from the calculations in Section 3.5.

Although it is important not to confuse the theoretical quantities with their estimates, the equations that relate the theoretical quantities in this chapter are identical to their counterparts for the estimates. Little is gained by writing them twice. Hence, to avoid duplication and notational clutter, the distinction between a quantity and its estimate is not preserved notationally here. The context in which an equation is used makes it clear whether the theoretical parameters of the signal-detection model or their estimates from data are under consideration. In applications—say, when discussing data—the estimates are often not specifically marked, but in theoretical discussions it is valuable to preserve the distinction.

Example 4.1: Find Δm from the four conditions in Figure 4.1.

Solution: Figure 4.1 shows the frequencies converted to the proportions f and h and transformed into $Z(f)$ and $Z(h)$. The straight line $z_H = a + bz_F$ fitted to the transformed points (using a computer program) is

$$z_H = 0.550 + 0.736z_F,$$

with axis crossing points at $x_0 = -0.747$ and $y_0 = 0.550$. The operating characteristic line crosses the z_F axis at $-\Delta m$. This value can be read directly from the graph or it can be calculated from the numerical slope and intercept:

$$\Delta m = -x_0 = \frac{a}{b} = \frac{0.550}{0.736} = 0.747.$$

A convenient feature of Δm as a detection measure is its closeness to the signal-detection model. It is directly equivalent to a parameter of the Gaussian model, so that it is easy to interpret once that model is understood. When the variances of the signal distribution and the noise distribution are the same, Δm has the same value as d', so it is interpreted in the same way, particularly when the difference between the variances is small.

As a measure of detection, Δm suffers from two disadvantages, one related to its interpretation, the other statistical. The interpretational difficulty occurs because Δm does not completely capture the extent to which the distributions overlap. The magnitude of the overlap depends on both the difference between means and the variance of the signal distribution. Figure 4.3 illustrates the problem. In both panels, $\Delta m = 2$. In the top panel σ_s is half the size of σ_n. There is not a great deal of overlap in the distributions and a good observer can be correct as much as 92% of the time. In the bottom panel, the relative sizes of the standard deviations are reversed, so σ_s is twice σ_n. There is more overlap and the observer now can at best be correct only 77% of the time. In spite of the equal value of Δm, it seems unreasonable to treat the two signals as equally detectable.

The statistical difficulty with Δm is of greatest concern when there is substantial variability in the data. The extent to which a quantity varies accidentally from one replication to another under identical conditions is

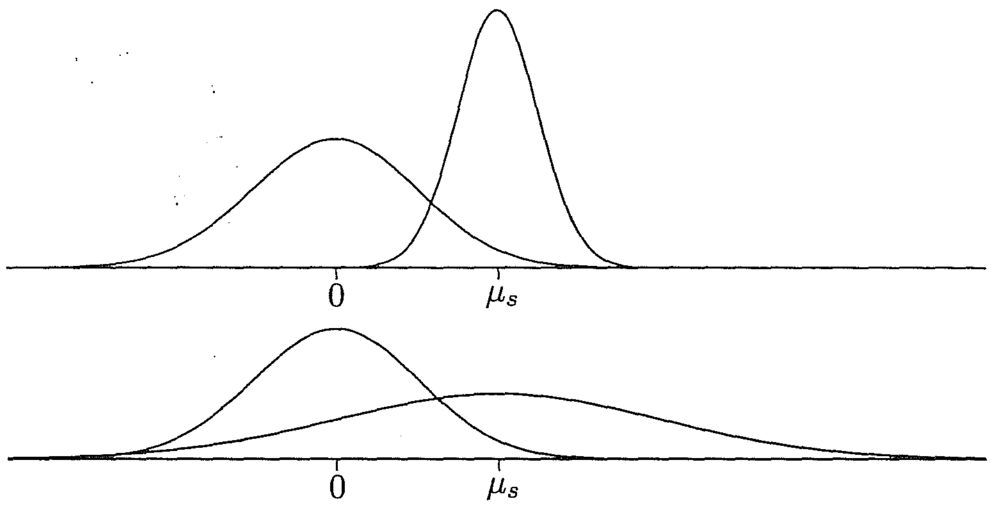

Figure 4.3: Two pairs of noise and signal distributions with $\sigma_n^2 \neq \sigma_s^2$. *Although* $\Delta m = 2$ *for both pairs, the overlap of the distributions is greater and optimal performance is lower in the pair for which* σ_s^2 *is large.*

measured by its standard error. The standard error of Δm is larger than it is for several comparable measures. The techniques in the chapter on statistical treatment can be used to make exact calculations (see Section 11.3), but the problem can be understood geometrically. A line fitted to a swarm of points is most accurately determined in the middle of the swarm. Its accuracy is greatest at the mean of the observations and falls off as one moves away from this center, an effect discussed for regression lines in many statistics books. A measure that depends on the position of a single point on the line can be determined with the least error when that point lies in the middle of the points used to estimate the line. However, most of the data points in a typical experiment fall to the right of the intercept that determines Δm (e.g., as in Figure 3.7 on page 48). The place where the isosensitivity line crosses the horizontal axis is away from the center of the data, making Δm less accurate as a measure of the line's location than a point in the middle of the data.

4.2 Distances to the isosensitivity line

As Figure 4.2 shows, the detection measure Δm is the horizontal distance between the origin in Gaussian coordinates and the isosensitivity line. The detectability of the signal, in this sense, is measured by how far this line is from the origin. The point on the line to which this distance is measured corresponds to a condition with a hit rate of $P_H = \frac{1}{2}$, which is a peripheral condition in most studies. One way to get a better measure is to derive it from a more central point, which will be more stable. Two measures have been proposed that are based on this idea.

One measure, known as d_e, derives from the point corresponding to a condition with symmetrical error rates. For this balanced conditions, the false-alarm rate equals the miss rate. It is likely to fall in the middle of the range of points produced by bias manipulations in the typical experiment. At this point, the equality of the error rates means that $z_H = -Z[P(\text{miss})] = -z_F$. The point, labeled **e** in Figure 4.2, lies at the intersection of the operating characteristic $z_H = a + bz_F$ and the minor diagonal $z_H = -z_F$. Solving these simultaneous equations shows that they intersect at $\mathbf{e} = (-z_e, z_e)$, with $z_e = a/(1+b)$. One could express the relationship of this point to the origin $\mathbf{0}$ either by the distance from $\mathbf{0}$ to **e** or by z_e directly. However, it is useful to make the measure coincide with d' when the equal-variance model holds. For that model, $a = d'$ and $b = 1$, so that $z_e = d'/2$. Accordingly, define the measure d_e to be twice z_e. This measure can be written in terms of the slope and intercept, the parameters of the signal-detection model, or the axis crossing points of the operating characteristic (using Equations 3.4 and 3.6):

$$d_e = \frac{2a}{1+b} = \frac{2\mu_s}{1+\sigma_s} = \frac{2x_0y_0}{x_0 - y_0}. \tag{4.2}$$

When $\sigma_s = 1$, then $d_e = d'$.

The second measure, known as d_a, derives from the point **a** on the operating characteristic that is closest to the origin in Gaussian coordinates. The point of closest approach of a line to a point lies in a direction perpendicular to the line (see Figure 4.2). Lines perpendicular to the isosensitivity line $z_H = a + bz_F$ have a slope of $-1/b$, so the closest approach to the origin is along the line $z_H = -z_F/b$, which has this slope and passes through the origin. Solving this pair of simultaneous equations, as in finding **e**, shows that the lines intersect at the point $\mathbf{a} = [-ab/(1+b^2), a/(1+b^2)]$. The distance from the origin to **a** is $a/\sqrt{1+b^2}$. Again, it is helpful to bring the measure in line with d', which is done by multiplying the distance by $\sqrt{2}$. The closest-approach measure is

$$d_a = \frac{\sqrt{2}\,a}{\sqrt{1+b^2}} = \frac{\sqrt{2}\,\mu_s}{\sqrt{1+\sigma_s^2}} = \frac{\sqrt{2}\,x_0y_0}{\sqrt{x_0^2+y_0^2}}. \tag{4.3}$$

Example 4.2: Calculate the distance detection statistics for the data in Figure 4.1.

Solution: Working from the slope and intercept found in Example 4.1, and using Equation 4.2, the distance measure d_e is

$$d_e = \frac{2a}{1+b} = \frac{2 \times 0.550}{1+0.736} = 0.634.$$

Likewise, from Equation 4.3, the distance measure d_a is

$$d_a = \frac{\sqrt{2}\,a}{\sqrt{1+b^2}} = \frac{1.414 \times 0.550}{\sqrt{1+(0.736)^2}} = 0.626.$$

Although d_a is perforce a little smaller than d_e (it is based on the closest approach to the origin), there is little difference between them. They both differ from the estimate $\Delta m = 0.747$ found for these data in Example 4.1.

Starting with a sketched line, one can get the same answers from the crossing points. For example,

$$d_a = \frac{\sqrt{2}\,x_0 y_0}{\sqrt{x_0^2+y_0^2}} = \frac{1.414 \times (-0.747) \times 0.550}{\sqrt{(0.550)^2+(-0.747)^2}} = 0.626.$$

Although measures d_a and d_e are not as tightly linked to the parameters of the Gaussian signal-detection model as is Δm, they make better indices of detection. There is not a great deal of difference between them. The symmetrical error structure associated with **e** makes d_e a plausible choice. The point **a** does not have such a simple interpretation, and when thought of as describing the outcome of a study, d_a seems a little arbitrary. However, it turns out that d_a is closely related to the measure A_z that is discussed in the next section, which gives it an advantage.

In most studies, the points **e** and **a** that determine either d_e and d_a fall near the middle of the data points that are used to determine the operating characteristic. These points are more stable and less influenced by accidents than points farther away from the center. Thus, except with very unusual sets of data, either d_e or d_a is statistically superior to Δm.

4.3 The area under the operating characteristic

The three detection measures discussed so far all derive from the Gaussian model: Δm is a parameter of the model and d_s and d_a depend on the fact that the isosensitivity line in Gaussian coordinates is straight. The final detection measure considered here is less directly tied to that model, at least superficially. It is simply the area under the isosensitivity curve when it is plotted in ordinary probability coordinates. This area, denoted generically by A, is shaded in the inset to Figure 4.2.

A look at the sequence of isosensitivity contours for the equal-variance model plotted in Figure 3.4 on page 43 shows how as d' increases, the operating characteristics shift up and to the left. The area below the chance function ($d' = 0$) is $1/2$, and as d' gets large it increases to 1. These characteristics apply to models other than the equal-variance Gaussian: the area is $1/2$ when the responses are unrelated to the stimulus, it increases with the detectability of the signal, and it asymptotes to unity for very strong signals.

The value of A has an interpretation in procedural terms that is one of its most attractive features. It is equal to the probability of a correct response in the *two-alternative forced-choice experiment* that will be discussed in Chapter 6. In that procedure, the observer is presented on each trial with two stimuli, one that contains the signal and one that does not. The observer's task is to decide which is which. As will be shown in Section 6.4, a result known as the *area theorem* shows that the probability of a correct response by an observer who is performing optimally in the forced-choice task is equal to the area under that observer's isosensitivity curve in a detection experiment using the same stimuli. This interpretation of A as a correct-response probability is easy to understand.

As just defined, the area measure A does not depend on any particular assumptions about the form of X_s and X_n, or even on the use of a distribution-based model. If an operating characteristic can be determined, then the area beneath it can be calculated without reference to a particular detection model. In this sense, A is sometimes described as a "nonparametric" or "distribution-free" measure of detection. As will be seen, this designation must be qualified.

From a mathematical point of view, the area under the theoretical isosensitivity curve is calculated by an integral. Suppose that for a particular detection model, the operating characteristic is given by the function $P_H = R(P_F)$. By definition, the area A under this line is the integral of $R(p)$ between $p = 0$ and $p = 1$:

$$A = \int_0^1 R(p)\, dp. \tag{4.4}$$

For any model that is specific enough to determine the function $R(p)$, this area can be calculated, either analytically or numerically. As in the formal analysis of Section 3.2, write the area in terms of the functions that relate the criterion to the false-alarm rate and hit rate, $P_F = F(\lambda)$ and $P_H = H(\lambda)$ (Equation 3.2 on page 45). The equation of the operating characteristic is $R(p) = H[F^{-1}(p)]$, and the area is

$$A = \int_0^1 H[F^{-1}(p)]\, dp. \tag{4.5}$$

This integral gives the area.

Now turn to the Gaussian model, for which the hit and false-alarm functions are

$$F(\lambda) = \Phi(-\lambda) \qquad \text{and} \qquad H(\lambda) = \Phi\left(\frac{\mu_s - \lambda}{\sigma_s}\right).$$

A calculation that will be presented in Section 6.4 shows that, with these functions, the area under the operating characteristic is

$$A_z = \Phi\left(\frac{\mu_s}{\sqrt{1+\sigma_s^2}}\right). \tag{4.6}$$

The subscript z here is a reminder that this area is derived from the Gaussian model. In terms of the isosensitivity function and the crossing points,

$$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) = \Phi\left(\frac{x_0 y_0}{\sqrt{x_0^2+y_0^2}}\right). \tag{4.7}$$

Example 4.3: Find the area under the Gaussian isosensitivity contour for the data in Figure 4.1.

Solution: Using the estimated slope and intercept obtained in Example 4.1, Equation 4.7 gives

$$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) = \Phi\left(\frac{0.550}{\sqrt{1+(0.736)^2}}\right) = \Phi(0.443) = 0.671.$$

The term within the distribution function Φ in Equation 4.6 is the same as the distance from the origin to the point **a** where the operating characteristic passes closest to the origin (Equation 4.3). As a result, the area measure is related to the distance measure d_a by

$$A_z = \Phi(d_a/\sqrt{2}). \tag{4.8}$$

The area measure lies between $1/2$ and 1 and the distance measure between 0 and ∞. The relationship between them looks like half of a cumulative Gaussian distribution function (Figure 4.4). When the equal-variance model holds, the distance-based measures are identical, and d' can be used instead of d_a in Equation 4.8 or on the abscissa of Figure 4.4.

The fact that area A is defined very generally for any operating characteristic suggests that it might be possible to calculate this area directly from a set of observed data without going through a theoretical model. It is possible to do something of the sort, although the resulting area has some serious defects. The idea is to connect the observed points by straight lines, then break up the area below them by trapezoids whose areas are easily

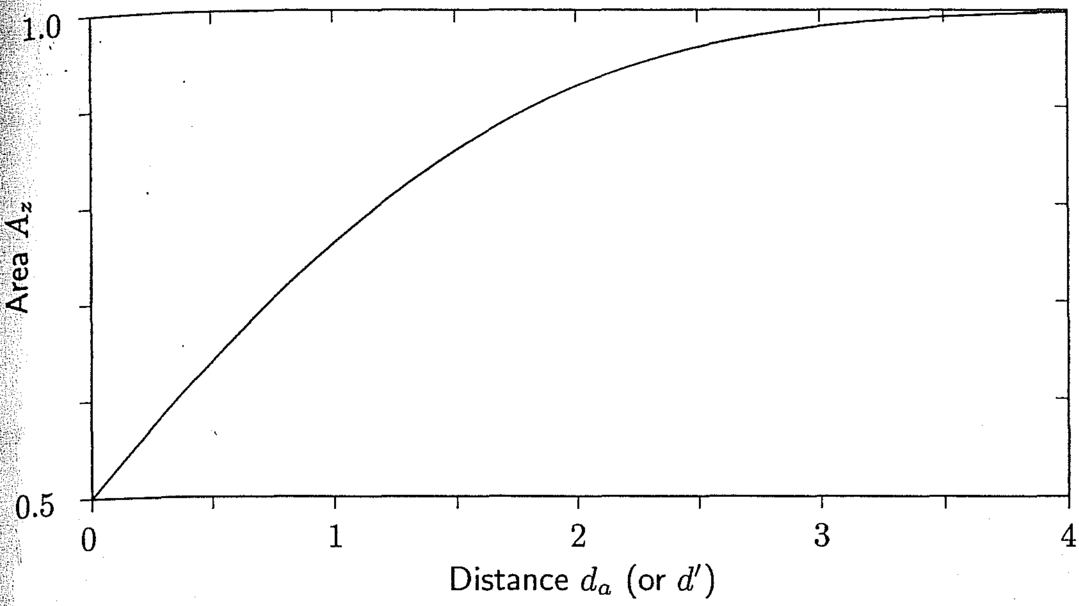

Figure 4.4: The relationship between the area A_z under the Gaussian operating characteristic and the distance d_a (or d' for the equal-variance model).

calculated. The procedure is illustrated by a two-condition study that is analyzed in Figure 4.5. The two observations lie at $f_1 = 0.133$ and $h_1 = 0.333$ and at $f_2 = 0.333$ and $h_2 = 0.660$. The empirical operating characteristic goes from the origin to the first point, then to the second point, and finally to the point (1, 1). The area under this curve is the sum of the area of the triangle A_1 and that of the two trapezoids A_2 and A_3. The area of a triangle is half the product of its base and its height, while the area of a trapezoid is the base times the average of the two heights. The three sections sum to give the total shaded area in Figure 4.5:

$$A_{\text{trap}} = A_1 + A_2 + A_3 = \frac{f_1 h_1}{2} + \frac{h_1 + h_2}{2}(f_2 - f_1) + \frac{h_2 + 1}{2}(1 - f_2).$$

Substituting the numerical values gives the area $A_{\text{trap}} = 0.649$. For a study with n points, A_{trap} is given by a generalization and rearrangement of the two-point formula:

$$A_{\text{trap}} = \tfrac{1}{2}\left[1 + \sum_{i=2}^{n}(h_{i-1}f_i - h_i f_{i-1}) + h_n - f_n\right]. \tag{4.9}$$

Although A_{trap} is easier to calculate than the other measures, its usefulness is limited by two problems. First, in many sets of data the points do not form a monotonically increasing sequence. Without such a sequence, the trapezoids cannot be cleanly drawn. For example, the four points in

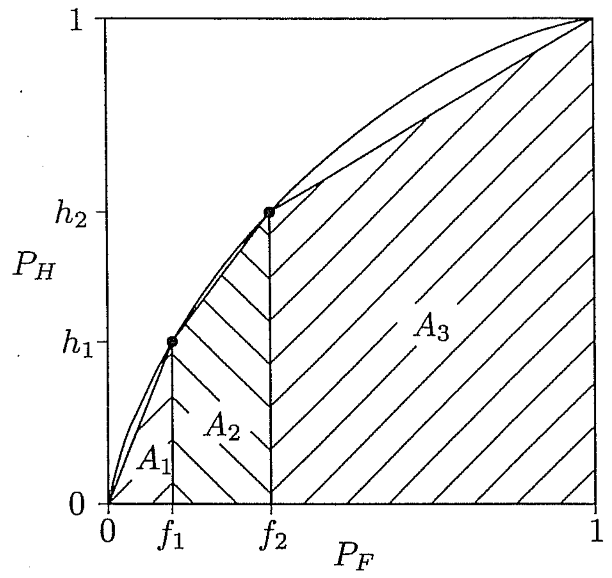

Figure 4.5: The trapezoidal area A_{trap} under an empirical isosensitivity curve. The smooth curve gives the equal-variance Gaussian function fitted to the same data.

Figure 4.1 do not form a proper operating characteristic, but one whose slope goes from steep to flatter to steep again. Still worse are the data in Example 3.1 on page 47. A line connecting the points in Figure 3.7 is not monotonic and does not correspond to a reasonable operating characteristic. In both cases, some sort of smoothing of the data is necessary. This smoothing cannot be done without an idea of what the operating characteristic should look like. The best way to regularize the results is to fit them with a theoretical model. In Figure 4.5, this smoothing has been done by fitting the equal-variance Gaussian model to the two points, and the result is shown by the smooth line. The area under this line is $A_z = 0.685$, which is greater than $A_{\text{trap}} = 0.649$. Other methods of smoothing could be used, but each implies something about what the true operating characteristic should look like. There is no good reason to prefer any of these over the Gaussian fit.

The second problem with A_{trap} is its bias. The empirical operating characteristic was formed by connecting the points by straight lines. However, regular isosensitivity contours are almost always bowed and concave downward, like those of the Gaussian model. Even when straight line operating characteristics are found (as they are for the finite state models of Chapter 8), the transitions from one segment to another are unlikely to be at the observed data points. As a result, the function formed by connecting the data points lies below the true operating characteristic—it is actually the lower bound of all possible regular operating characteristics that pass through the points. The area A_{trap} is an underestimate of the true area.

The bias in A_{trap} would be less of a matter for concern were its magnitude not dependent on characteristics of the experiment itself. The amount of bias is small when many points on the empirical function have been found and their false-alarm rates are spread out evenly between zero and one. The bias is larger when only a few points are measured or they are bunched together. For example, the bias of A_{trap} in Figure 4.5 would be reduced by adding a point that lies on the true function and has a false-alarm rate between f_2 and 1. This dependence of the area on the conditions used in the experiment undercuts the distribution-free claims for A_{trap}. One way to eliminate the bias is to use the same approach that was necessary to deal with the sampling irregularities. The calculation must be based on a model such as the Gaussian model. Even if this model is not quite correct, the errors that it introduces are surely less than the inevitable negative bias of A_{trap}.

When only a single pair of hit and false-alarm probabilities is observed, any estimate of the area under the operating characteristic requires a considerable extrapolation, as the entire curve must be reconstructed from a single point. Of necessity, one must go beyond the data here. A good estimate of the area, derived from the Gaussian model in Equation 4.8, is $A_z = \Phi(d'/\sqrt{2})$. Some other measures of area have been proposed that are sometimes described as better approximations to the true area than A_z. Perhaps the most compelling of these are the measures based on the triangles or trapezoids formed by passing lines from the origin $\mathbf{0}$ or the point $\mathbf{1} = (1, 1)$ through the observed point $\mathbf{p} = (f, h)$ (Figure 4.6). One such measure, often denoted A', averages the area under the line extended from $\mathbf{0}$ through $\mathbf{p}$ to the point marked $\mathbf{b}$ in the figure with the area under the line extended from $\mathbf{1}$ through $\mathbf{p}$ to $\mathbf{a}$:

$$A' = \tfrac{1}{2}[(\text{area below } \mathbf{0} \text{ to } \mathbf{b} \text{ line}) + (\text{area below } \mathbf{a} \text{ to } \mathbf{1} \text{ line})].$$

An analysis of the geometry involved shows that

$$A' = 1 - \tfrac{1}{4}\left[\frac{1-h}{1-f} + \frac{f}{h}\right]. \tag{4.10}$$

Although the average A' does not equal the area under the operating characteristic associated with any particularly natural detection process, the two areas that are averaged do have interpretations in terms of the threshold models discussed in Chapter 8 (see Figure 8.3 on page 137).

It is sometimes asserted that because Equation 4.10 has been developed from the areas directly, it does not involve assumptions about the distributions of X_n and X_s. By this argument, A' is "nonparametric" and preferable to measures such as A_z, which are based on the Gaussian distribution. However, this argument is not correct. One cannot extrapolate

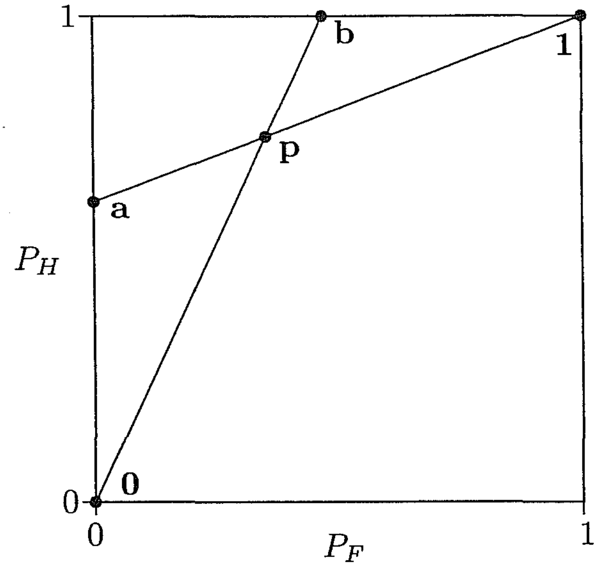

Figure 4.6: Trapezoidal areas constructed from a single observation with $h = 0.75$ and $f = 0.35$. The letters **0**, **a**, **b**, **p**, *and* **1** *identify points.*

from one point to an area without assumptions, and those that underlie A' are at least as stringent as those that underlie A_z (see reference note). Moreover, when data from several points are available, the operating characteristic usually looks more like the curve used to calculate A_z than like that implied by A'. Unless one has specific reasons to question it, A_z is preferable.

4.4 Recommendations

At this point a summary is in order, with some recommendation (which inevitably reflect my own preferences and biases). The goal of signal-detection theory, insofar as measuring detection is concerned, is to separate the aspects of the situation that depend on the intrinsic strength of the signal's effect on the observer from those that depend on the observer's decision to make a particular response. This separation cannot be made without some description of the decision process. The basic representation emphasized in this book is that of the Gaussian model. This description was chosen for several reasons. Among these are its historical importance and wide use, its simplicity and naturalness, its links to the statistical literature, and, most essential, the fact that it fits a great many sets of data well.

The Gaussian model is not the only possible representation. Different assumptions about the nature of the signal and the detection process lead to other models. Prominent here are the choice models mentioned briefly at the end of Section 2.1, the finite-state models to be described in Chapter 8, and the general likelihood approach of Chapter 9. Each of these representations suggests its own ways to measure the detectability of a signal. When a theoretical analysis based on one of these descriptions (or one not covered here) is adopted, the measures that flow from it should be used. A non-Gaussian model might also be chosen when it provides a better fit to a set of data, although one should be wary of sacrificing clarity of explanation for small improvements in empirical fit. Although the examination of one's data could indicate that a non-Gaussian model is required, these tests take a considerable amount of data and are not often run. On the whole, adhering to the Gaussian assumptions is acceptable.

That being said, consider the Gaussian case. When a single pair of hit and false-alarm rates is available, the natural statistic to use is d'. Either it or a transformation of it to A_z (Equation 4.8) measures detection. This much is straightforward. Either measure separates detection performance from the decision process as well as possible with such limited data.

When two or more points on the isosensitivity contour are available, a more complex representation is possible, and one must choose among several competing measures of detectability. The simple equal-variance model was extended to its more general form by adding complexity to the representation of the signal by changing X_s from $\mathcal{N}(d', 1)$ to $\mathcal{N}(\mu_s, \sigma_s^2)$, not by changing the decision component (the criterion λ or the bias $\log\beta$). Extracting a single number from this two-dimensional description of the signal invariably emphasizes one aspect of the stimulus over another. Each of the detection measures discussed in this chapter (and several others, not discussed) places this emphasis in a different way. It would be nice if psychological or perceptual theory could clearly identify which of these measures is most meaningful, but these theories are not sufficiently advanced. Indeed, it seems improbable that one representation can be the best description of the diverse situations to which signal-detection theory applies.

Lacking a strong psychological theory, one must turn to a measure that has satisfactory general properties. The strongest contenders are the measures related to the area under the operating characteristic. They integrate (both figuratively and literally) the performance over the range of potential biases. They are also somewhat less dependent on the details of the Gaussian model than are the distributional parameters, for most other representations have much the same area under their operating characteristic. Either the direct area A_z or its monotonic transformation to the distance d_a might be used—they are equivalent in that they imply the same ordering for any set of stimuli. Which of the two to prefer is less clear. For signals of moderate intensity they behave similarly. The relationship between them, shown in Figure 4.4, can be treated as a straight line out to about $d_a = 2$ without undue violence. Beyond that, the two measures differ mainly in the extent to which they separate easily detectable signals. Values of d_a map strong signals out to infinity, while values of A_z are bounded by one. It is not clear which behavior is preferable. On the one hand, d_a captures the fact that small improvements in performance at the high end may be very hard to come by and require a big difference in the physical stimulus. On the other hand, A_z is more stable and less influenced by relatively uninformative differences among conditions in which $h \approx 1$ or $f \approx 0$.

The other two measures Δm and d_e are less plausible choices. The instabilities noted at the end of Section 4.1 make the signal mean μ_s (or its alias Δm) clearly inferior to the other measures. The distance-based d_e is unobjectionable, but there is no reason to prefer it to d_a.

There remains a collection of measures that are not directly derived from an underlying detection model. These include the trapezoidal area A_{trap}, the average A' mentioned at the end of Section 4.3, and such measures as the difference between the hit rate and the false-alarm rate. Seemingly, these measures are not tied to any particular description of the detection process and give a way to get away from theoretical models. However, that is a false hope. Each such measure can be derived from a description of the detection process. To use a measure implies a willingness to either adopt that description or one equivalent to it. Typically, the structures implied by these measures are less plausible than one based on the type of graded evidence implied by the Gaussian models. It is simply not possible to develop measures of detectability without reference to what evidence was used for the decision and how that decision was made, either explicitly or implicitly. One has to have some sort of model or description process, and the Gaussian model is a good choice.

4.5 Measures of bias

Many detection studies are concerned only with measuring the detectability of the signal. The role of signal-detection theory here is to remove the contaminating effects of response bias and obtain cleaner results. However, sometimes comparisons of the bias in different conditions must be made. For this comparison, a measure of the bias is needed.

Measurement of bias is in some ways more difficult than measurement of detectability. Even with the equal-variance model, different measures with different properties were available. As discussed in Section 2.4, both λ_{center} and $\log \beta$ are plausible measures, but, as Figure 3.5 on page 44 showed, their isobias contours are quite unlike each other. It is difficult to argue on principle that one measure is better than the other. An understanding of the psychological factors that are involved in bias shifts could help, but such a theory is, if anything, farther away than an understanding of stimulus effects.

Somewhat surprisingly, the situation with the unequal-variance model is simpler than it is for the equal-variance model. A centered criterion such as λ_{center} is hard to justify when the distributions have unequal variance (where should the center be placed?). In contrast, the heights of the density functions at the criterion are clearly defined, as is their ratio. The logarithm of the likelihood ratio at the criterion (Equation 2.8) makes a clear and defensible measure of the bias:

$$\log \beta = \log\left[\frac{f_s(\lambda)}{f_n(\lambda)}\right] = \log f_s(\lambda) - \log f_n(\lambda).$$

Moreover, this measure is closely linked to the more general decision models that are discussed in Chapters 9 and 10, which gives it both a wider applicability and a plausible intuitive interpretation.

The likelihood ratio at the criterion, $\beta = f_s(\lambda)/f_n(\lambda)$, is determined by the shape of the operating characteristic. The numerator of this ratio is the rate at which probability is added to P_H as λ changes, and the denominator is the rate at which probability is added to P_F (recall that the probabilities are integrals of these functions; see Equations 3.1 on page 45). The ratio β of these rates gives the rate at which P_H changes with P_F, which is the slope of the operating characteristic.[2] As a result, one can actually use the slope of the operating characteristic as a measure of bias. It is, of course, equivalent to $\log \beta$.

The density functions for the unequal-variance Gaussian model (Equation A.42 on page 237) are

$$f_s(x) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left[\frac{-(x-\mu_s)^2}{2\sigma_s^2}\right] \qquad \text{and} \qquad f_n(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}x^2\right].$$

Substituting $x = \lambda$, taking the logarithms, and subtracting gives the bias:

$$\begin{aligned} \log \beta &= \log f_s(\lambda) - \log f_n(\lambda) \\ &= \left[-\log\sqrt{2\pi} - \log \sigma_s - \frac{(\lambda-\mu_s)^2}{2\sigma_s^2}\right] - \left[-\log\sqrt{2\pi} - \tfrac{1}{2}\lambda^2\right] \\ &= \tfrac{1}{2}\left[\lambda^2 - \frac{(\lambda-\mu_s)^2}{\sigma_s^2}\right] - \log \sigma_s. \end{aligned} \tag{4.11}$$

[2] This result can be proved formally by differentiating the function $P_H = H[F^{-1}(P_F)]$ (Equation 3.2) using the chain rule.

Written using the slope and intercept of the isosensitivity line (i.e., substituting $\mu_s = a/b$ and $\sigma_s = 1/b$), this equation is

$$\log\beta = \tfrac{1}{2}[\lambda^2 - (a - b\lambda)^2] + \log b. \tag{4.12}$$

The bias of a point $(z_F,\, z_H)$ on the fitted line (i.e., one with $z_F = -\lambda$ and $z_H = a + bz_F$) is found from this equation to be

$$\log\beta = \tfrac{1}{2}(z_F^2 - z_H^2) + \log b. \tag{4.13}$$

Each of these equations reduces to the equal-variance form in Section 2.4 when $\sigma_s^2 = b = 1$.

To apply Equation 4.13, the point $(z_F,\, z_H)$ must be on the fitted isosensitivity line, and in Equations 4.11 and 4.12 the value of λ is that of this point. When the data are scattered around the best-fitting line, as in Figure 3.7 on page 48, observed values of $Z(f)$ and $Z(h)$ cannot be substituted for z_F and z_H, nor can $-Z(f)$ be taken for λ. The observed point must be adjusted to bring it into line with the model. Computer programs that fit the signal-detection model usually provide these corrected estimates of λ (or even give estimates of the bias) as part of their output. However, a calculational procedure is useful when the line has been sketched by eye or when published data are used. A good way to make this adjustment is to transfer the observed points to the closest point on the line. This adjustment is not completely optimal, as it does not take account of the sampling variability of f and h, but is close to the best solution. Figure 4.7 illustrates the process. The observed point $[Z(f),\, Z(h)]$ is translated to the point $(\tilde{z}_F,\, \tilde{z}_H)$ along a line perpendicular to the fitted line. The geometry here is identical to that used to determine the closest approach to the origin when finding d_a (Equation 4.3). In this case, it gives

$$\tilde{z}_F = \frac{b[Z(h) - a] + Z(f)}{1 + b^2} \qquad \text{and} \qquad \tilde{z}_H = a + b\tilde{z}_F. \tag{4.14}$$

These values are then substituted into Equation 4.13.

Example 4.4: Determine the bias parameters for the conditions in Figure 4.1.

Solution: The program that fitted the model in Example 4.1 gave estimates of the criteria. For the first condition, λ_1 was 1.017. Then, using Equation 4.12 with $a = 0.550$ and $b = 0.736$,

$$\begin{aligned}\log\beta_1 &= \tfrac{1}{2}[\lambda_1^2 - (a - b\lambda_1)^2] + \log b\\ &= \tfrac{1}{2}[(1.017)^2 - (0.736\times1.017 - 0.550)^2] + \log(0.736) = -0.191.\end{aligned}$$

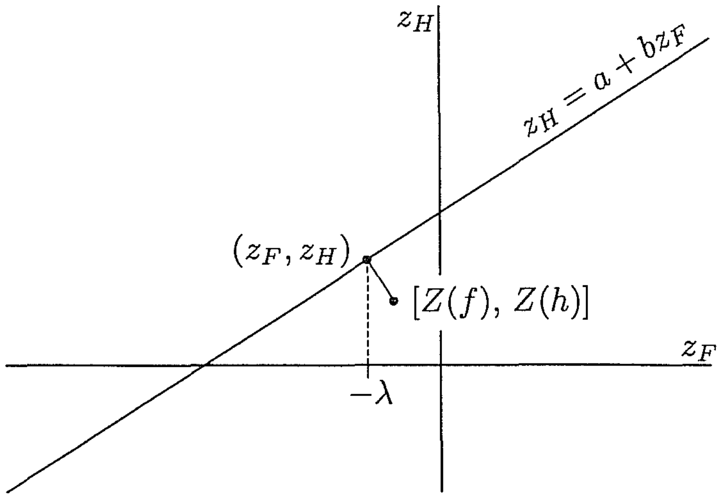

Figure 4.7: Shifting an observed point $[Z(f),\ Z(h)]$ to the perpendicularly nearest point $(\widetilde{z}_F,\ \widetilde{z}_H)$ on the fitted line to estimate the criterion or bias.

The other criteria are 0.518, 0.211, and −0.183, which give bias estimates of −0.186, −0.362, and −0.524, respectively.

If estimates of the criteria were not provided by the fitting program, then they need to be derived from the original point and the fitted line. For the first point, $f_1 = 0.165$ and $h_1 = 0.405$, so $Z(f_1) = -0.974$ and $Z(h_1) = -0.240$. Using Equation 4.14 to translate the observed point to the line gives

$$\begin{aligned} \widetilde{z}_{F_1} &= \frac{b[Z(h_1) - a] + Z(f_1)}{1 + b^2} \\ &= \frac{0.736(-0.240 - 0.550) + (-0.974)}{1 + (0.736)^2} = -1.009. \end{aligned}$$

The criterion $\lambda_1 = 1.009$ obtained from this coordinate is essentially the same as the value 1.017 produced by the computer algorithm, the difference being due to the slightly different estimation criteria and to rounding error. The transformed hit rate corresponding to the new estimate is

$$\widetilde{z}_{H_1} = a + bz_{F_1} = 0.550 + (0.736)(-1.009) = -0.193.$$

The point (1.009, −0.193) falls on the line in Figure 4.1 (showing that a computational error has not been made). The bias estimate from Equation 4.13 is

$$\begin{aligned} \log\beta &= \tfrac{1}{2}(\widetilde{z}_{F_1}^2 - \widetilde{z}_{H_1}^2) + \log b \\ &= \tfrac{1}{2}[(-1.009)^2 - (0.193)^2] + \log(0.736) = 0.184. \end{aligned}$$

Again, the small difference between this result and that of the program is unimportant.

4.6 Aggregation of detection statistics

The signal-detection measures in this chapter apply best to data obtained by collecting many observations from a single observer. In this way, good estimates of P_H and P_F can be made, and one can be reasonably confident that the underlying parameters are approximately stable. This type of data is frequently found in studies that are conducted in the psychophysical tradition from which signal-detection theory was developed. These studies typically use a small number of observers (sometimes just one), each of whom is analyzed separately and discussed individually.

In other domains, studies are common that use larger groups of observers, in this context usually called *subjects* or *participants*. In these studies, the researcher wishes to get some sort of average or mean result and draw inferences that apply to the group as a whole. For example, a memory researcher using a recognition paradigm may run 20 or more subjects in each of several conditions, then compare the conditions. A signal-detection analysis is necessary, because the researcher wishes to remove the response bias when making comparisons of recognition accuracy, particularly as it undoubtedly varies over subjects. The researcher takes each subject's rate of true and false recognition responses (i.e., hit rate and false-alarm rate) and calculates either d' or A_z. The data now must be aggregated over the subjects so that a single conclusion can be drawn.

Similar issues arise in a single-observer analysis when the data from a particular detection task are collected on a series of days. Separate sets of hit and false-alarm rates are obtained on each day. The researcher does not plan to discuss each day's data separately (even though they differ slightly), but wants a single measure to express the average performance.

There are two ways to pool signal-detection results over a set of entities, be they subjects or sessions. One possibility is to pool the observations over the individual entities to obtain the total number of trials, hits, false alarms, and so forth, then to run the analysis on these aggregate data. The other possibility is to conduct separate signal-detection analyses for each individual entity, then combine the results. The first procedure is the simpler to apply, but, when it is feasible, the latter approach is much to be preferred.

If the two approaches gave the same answer, then there would be no problem—either approach could be used. However, generally they give different answers. The more the entities vary from one to another, the more the procedures diverge. From a formal point of view, the problem is that the signal-detection statistics are nonlinear functions of the original data, and the average of a collection of nonlinear functions is not equal to that function applied to the average. Thus, a statistic such as d' that is calculated from the pooled hits and false alarms does not equal the average of the d'_i calculated on the individual entities. The difference is not large when the entities are very homogeneous in their detection properties, but can be substantial when the entities are heterogeneous.

Either approach to pooling the entities gives a valid answer for its particular form of combination. The proper measure to use is the one that best matches the researcher's intent. Almost always, the intent is to find the average value of the individual detection statistics. For example, the memory researcher knows that subjects differ in how well they remember material and in their biases. Pooling the data just blurs these differences. Furthermore, although there is no average subject whose performance is estimated by the pooled data, there is a mean over the population of subjects who might have been sampled for the study. Inferences about the situation are expressed in terms of this mean, and it is estimated by the average of the sample statistics, not by those derived from the pooled data. The same arguments apply to the psychophysical researcher who collects data from the same observer on 10 similar sessions. It is not surprising that the observer differs a bit from one session to the other, but these differences are unimportant, so the researcher is concerned with the average performance. Each researcher should separately calculate whatever index is important for each subject or session, then average them. Calculating individual measures has an ancillary advantage, in that it provides information about the extent to which the subjects or sessions vary. This information is necessary for the statistical analysis that is discussed in Section 11.7.

Example 4.5: Five subjects are given fifty trials on a recognition test in which half the items are old and half are new. The left-hand columns in Table 4.1 give the number of YES responses made by each subject to new and old items. What is the average recognition performance for these subjects?

Solution: The first step is to calculate the detection statistics for each individual subject. Dividing the frequencies of OLD responses by 25 gives the false-alarm and hit rates, and a standard equal-variance analysis gives the detection statistics in the final columns. These values are averaged to give the mean values of $\overline{d'} = 0.64$ and $\overline{A}_z = 0.67$.

Pooling the responses before applying the model is less appropriate. Doing so gives mean response rates of

$$\bar{f} = \frac{70}{125} = 0.560 \qquad \text{and} \qquad \bar{h} = \frac{93}{125} = 0.744.$$

	New	Old	f	h	$\widehat{d'}$	$\widehat{A_z}$
S_1	10	15	0.40	0.60	0.507	0.640
S_2	12	20	0.48	0.80	1.095	0.781
S_3	14	19	0.52	0.76	0.792	0.712
S_4	16	23	0.64	0.92	1.047	0.770
S_5	18	16	0.72	0.64	−0.224	0.437
Mean					0.643	0.668

Table 4.1: Recognition responses for new and old items from 5 subjects on a 50-item recognition test with equal numbers of old and new items. See Example 4.5.

Detection statistics derived from these values are $d' = 0.505$ and $A_z = 0.639$, somewhat less than the values obtained by the preferred method.

When the number of trials is small or the task particularly hard, some subjects may have a false-alarm rate that exceeds the hit rate. For example, subject S_5 in Example 4.5 has $f = 0.72$ and $h = 0.64$, for which the standard estimating equations give $\widehat{d'} = -0.22$. Presumably, d' for this subject is not actually negative, although it must be small—the best guess is $d' = 0$. The negative value occurred because sampling accidents happened to increase f and decrease h. It is tempting to replace the negative value by zero or to drop the subject from the analysis, but doing so biases the mean. Sampling accidents that went the other way, producing a too-large value of sensitivity would not be altered or censored. The net effect of systematically removing any negative deviations, is to inflate the sample mean relative to its population counterpart. Thus, it is important to keep the negative value in the analysis when averaging the detection statistics, as has been done in Example 4.5.

The use of individual detection statistics places demands on the study. There must be a sufficient number of observations available from each entity (subject or session) to make a reasonably accurate estimate of d', A_z, $\log \beta$, or whatever measure is desired. A researcher who is planning to use these techniques should ensure that enough data are collected in each condition to calculate stable values. For example, the tests given to the subjects in the memory study must include a sufficient number of items to give at least rough estimates of the individual hit and false-alarm rates. A study in which, say, five items of each type were used would be unsatisfactory.

Finding the average of the individual measures can encounter difficulties when the number of observations is small enough that false-alarm rates of $f = 0$ or hit rates of $h = 1$ occur. As noted in Section 2.3, the detection measures in such cases are indeterminate, and adjustments to give values for distance measures such as d' are arbitrary. When many such entities are present, it is usually advisable to use a bounded detection statistic such as A_z that is less affected by how the extreme scores are treated. One should also verify that any conclusions are not dependent on one's choice of adjustment procedure.

Reference notes

The material in this chapter draws on the same references that were cited for Chapter 2. For the material on the choice of measures, see Swets (1986b) and Swets and Pickett (1986). Examples of operating characteristics from a variety of domains are in Swets (1986a). The measure A' was first described by Pollack and Norman (1964); see Smith (1995) and Macmillan and Creelman (1996) for historical and theoretical discussion.

Exercises

4.1. The following data come from three detection experiments that use the same stimuli, but in which the proportion of signal and noise trials was varied to induce the observer to shift the criterion.

Signal proportion	f	h
High	0.04	0.40
Middle	0.31	0.70
Low	0.67	0.83

a. Plot these data and qualitatively evaluate the assumptions of the signal-detection model.
b. Sketch the distributions of noise and signal events and locate the criteria.
c. Estimate the four detection statistics Δm, d_e, d_a, and A_z.
d. Estimate the bias for each of the three conditions.

4.2. Find A_z for the data in Problem 2.6.

4.3. Calculate the detection statistics d_e, d_a, and A_z for the data in Problem 3.3. Estimate the bias for conditions A, D, and F.

4.4. A single observer performs simple detection of a signal in three conditions. In one condition, the observer is asked to minimize false alarms, in another to minimize misses, and in the third to roughly balance the two types of error. In each condition 200 signal trials and 200 noise trials are given. The results are as follows:

Low F.A.

	NO	YES
N	189	11
S	114	86

Balanced

	NO	YES
N	151	49
S	88	122

Low miss

	NO	YES
N	92	108
S	25	175

a. Which of the Gaussian models is the most reasonable choice to describe these data?

b. Estimate the area under the Gaussian operating characteristic.

c. Sketch the underlying distributions of evidence for the noise and signal conditions based on the model that you selected.

4.5. Suppose that the equal-variance Gaussian model with $d' = 1.5$ is a correct description of a single detection task. Determine the area A_z under the operating characteristic. Now consider a condition with $z_F = 0.2$. Find the associated hit rate from the Gaussian model and use z_F and z_H to calculate the trapezoidal area A_{trap}. Make similar calculations for conditions with $z_F = 0.5$ and $z_F = 0.75$. Compare the results to A_z. Illustrate what is going on with a diagram.