Chapter 3

Operating characteristics and the Gaussian model

In Section 1.2, data from two detection sessions with different biasing instructions were described. The analysis of these data showed that the sessions differed in the observer's bias toward YES or NO responses, but not appreciably in the detectability of the signal. In these analyses, each session's data were treated independently. A more comprehensive analysis should tie the two sets of observations together in a single model of the detection task. This chapter describes this integration. It also shows how to represent data from several detection conditions in a convenient, instructive, and widely used graphical form. This representation lets one investigate the adequacy of the assumptions that underlie the Gaussian signal-detection model and fit a model in which the variances of the signal and noise distributions are unequal.

3.1 The operating characteristic

Consider the results from the two sessions of Section 1.2. Fitting an equal-variance Gaussian model to these data, as in Examples 2.2 and 2.3, gives the parameter estimates at the top of Figure 3.1. As these numbers show, the estimates of d' are very similar, but those of λ and $\log \beta$ differ considerably. These values suggest that the proper representation of the full set of observations should use the same pair of distributions in the two sessions, but change the criterion. This representation is illustrated at the bottom of Figure 3.1. Using the same distributions for both sessions implies that the characteristics of the stimuli (which are controlled by the experimenter)

Note: d' is the sensitivity (the difference between the means of the Gaussians), lambda is the criterion and beta is the criterion in the form of the ratio of likelihoods of hits and false alarms. h is the hit rate and f is the false alarm rate.

	h	f	$\widehat{d'}$	$\widehat{\lambda}$	$\widehat{\log\beta}$
First Session	0.82	0.46	1.02	0.10	−0.42
Second Session	0.55	0.19	1.00	0.88	0.38

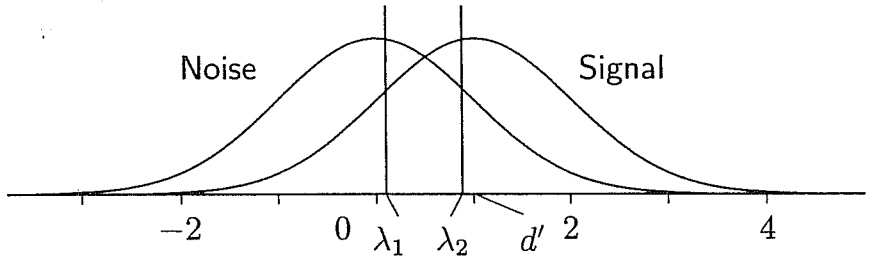

Figure 3.1: Data and parameter estimates for the two detection sessions of Section 1.2 and a representation using stimulus distributions separated by $d' = 1.01$ with specific criteria at $\lambda_1 = 0.10$ and $\lambda_2 = 0.88$.

do not change, while using session-specific criteria implies that the decision rules applied to these distributions are different.

The change in instructions that shifted the criteria in Figure 3.1 produced a trade-off between the two types of correct response, hits and correct rejections. The higher criterion produced many correct rejections but few hits, while a low criterion increased the hit rate at the cost of the correct-rejection rate. Put another way, changes in criterion cause the hit rate and the false-alarm rate to vary together. The trade-off is illustrated more directly by plotting the hit rate against the false-alarm rate, in the manner of a scatterplot, as shown in Figure 3.2. The false-alarm rate is plotted on the abscissa (horizontal axis) and the hit rate on the ordinate (vertical axis). Each session's results are represented by a point, S_1 for the first session and S_2 for the second session. Because the hit rate exceeds the false-alarm rate, these points lie above the diagonal of the square that connects the point (0, 0) to the point (1, 1). The outcome of any detection study corresponds to a point in this square, generally in the upper triangle.

The theoretical counterpart of Figure 3.2 plots the probabilities P_F and P_H generated by a particular signal-detection model. This plot is particularly valuable in illustrating how constraints on the parameters limit the predictions. Consider what is possible for an observer working according to the Gaussian model who can vary the criterion but who cannot alter d'. Figure 3.3 shows four points from such an observer, with the corresponding distributions in insets. Each point is based on the Gaussian model with $d' = 1.15$—note that the relative positions of the distributions do not change. However, the criteria differ. Bias toward NO responses gives the point at the lower left, and bias toward YES responses gives the point at the upper right.

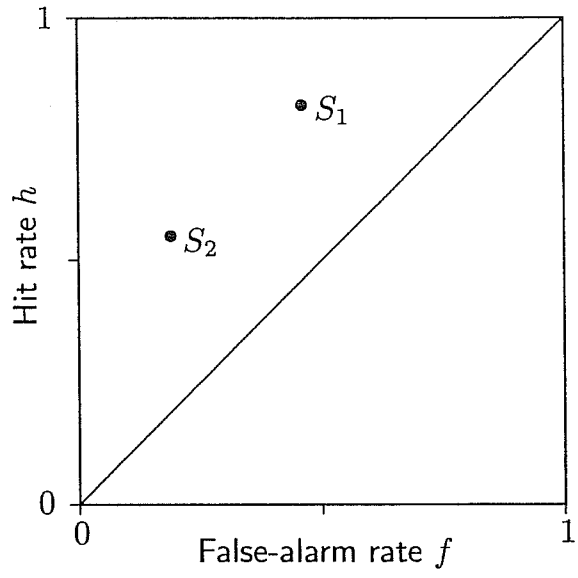

Figure 3.2: A plot of the false-alarm rate f and the hit rate h for the two detection sessions of Section 1.2.

The full range of performance available to this observer is shown by a line on this graph. When λ is very small, both P_F and P_H are large, leading to a point near the (1, 1) corner of the graph. As λ is increased, the area above the criterion decreases, as do P_F and P_H, and the point (P_F, P_H) traces out the curve shown in the figure. Eventually the prediction approaches (0, 0) when λ is so large that both distributions fall almost completely below it.

The solid line in Figure 3.3, which shows the possible range of performance as the bias is adjusted, is known as an *operating characteristic*. The original application of the statistical decision model to detection problems grew out of work on radio reception, where these curves were known as receiver operating characteristics. This name has stuck; the curve is often called a *receiver operating characteristic* or, more briefly, a *ROC curve*. Because it shows the performance possible for a fixed degree of discriminability, that is, a fixed sensitivity to the signal, the curve is also known as an *isosensitivity contour*. Both terms are used in this book.

Figure 3.3 shows the operating characteristic for a signal with detectability $d' = 1.15$. By varying the detectability of the signal—say, by changing its intensity—one obtains a family of curves. Figure 3.4 shows several members of this family. These are drawn for the equal-variance Gaussian model (using the methods described in the next section) and are labeled with values of d'. When $d' = 0$ the hit rate and the false-alarm rates are identical and the operating characteristic lies along the diagonal of the square. For positive values of d', the curves lie above this diagonal, moving close and closer to the upper left-hand corner as d' increases. When d' is

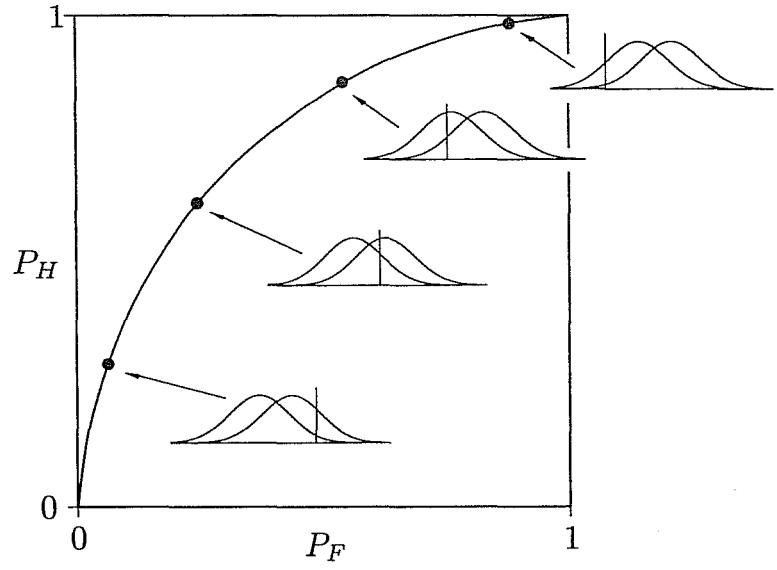

Figure 3.3: The operating characteristic or isosensitivity contour for a detection task derived from the equal-variance Gaussian model with $d' = 1.15$. The inset diagrams show the distributions of the random variables X_n and X_s and the criterion λ at four typical points.

very large, the operating characteristic is almost identical to the two lines at right angles that make up the left and top sides of the square. This operating characteristic describes performance when the signal is so strong that the observer can almost completely avoid making errors.

In an important sense, the operating characteristics are what describe the sensitivity of an observer to a particular signal. Together, the strength of that signal and the receptivity of the observer determine how detectable that signal will be. The combination selects the operating characteristics associated with that d', and the outcome of any study falls on that line. The point on that line that is actually observed is determined by the observer's bias.

3.2 Isocriterion and isobias contours

Aspects of the signal-detection model other than the bias can be varied to produce contours in the (P_F, P_H) space. *Isocriterion contours* are created by holding λ constant and varying the sensitivity. Because the criterion and the false-alarm rate determine each other (remember that P_F is the probability that $X_n > \lambda$), the isocriterion contours are vertical lines at constant values of P_F (Figure 3.5, top). Contours such as these might describe the results of a study in which signals of different strengths were

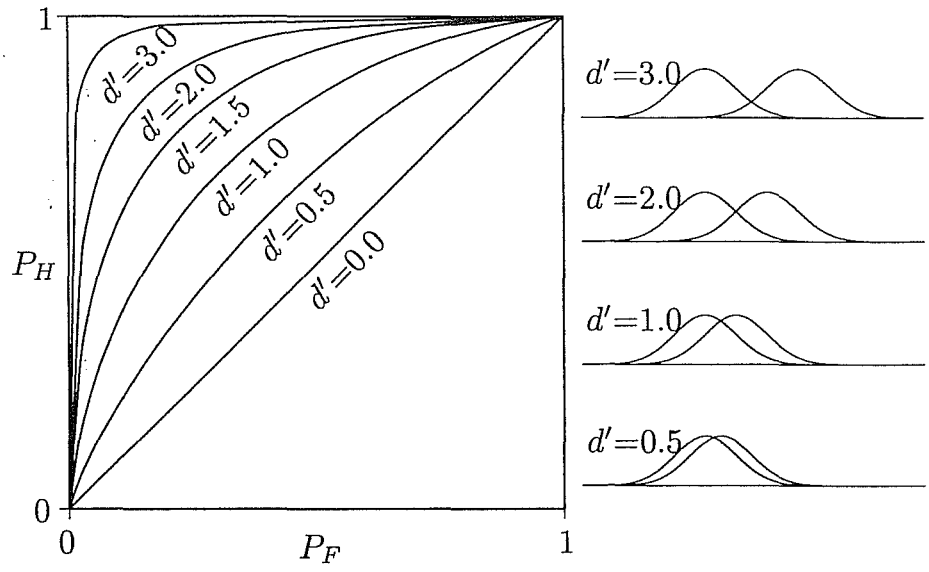

Figure 3.4: The family of isosensitivity contours determined by different values of d' for the equal-variance Gaussian model. At the side are shown the noise and signal distributions for four of the contours.

mixed with simple noise events, thus giving one false-alarm rate and several hit rates (one for each strength), with decisions based on a single criterion.

More interesting are the *isobias contours* created by holding one of the bias indices constant while varying the sensitivity. The lower two panels of Figure 3.5 show the lines of constant β or $\log\beta$ (left) and of constant (right). Both sets of isobias contours end in the upper left corner of the plot, a consequence of the fact that with sufficiently strong signals it is possible to attain near-perfect performance regardless of the bias. The curves otherwise have substantially different shapes, reflecting the different properties of the two measures. Each curve of constant likelihood ratio starts for very weak signals either at the point (0, 0) (for positive values of $\log\beta$) or the point (1, 1) (for negative values). In contrast, each curve of constant λ_{center} starts on a different point on the chance diagonal.

The isobias contours of either type cut across the isosensitivity contours of Figure 3.4. Thus, any point (P_F, P_H) in the space determines one level of sensitivity (a line from Figure 3.4) and one level of bias (a line from Figure 3.5). Similarly, any combination of sensitivity and bias determine, by the intersection of one line from each of the figures, a false-alarm rate and a hit rate.

Isocriterion and isobias curves are usually of less practical interest than are the isosensitivity curves. One reason is that it is easier to find situations in which one expects the sensitivity to remain constant while the bias changes than situations where one expects the criterion or the bias to remain constant while sensitivity changes. Another reason is that bias is a

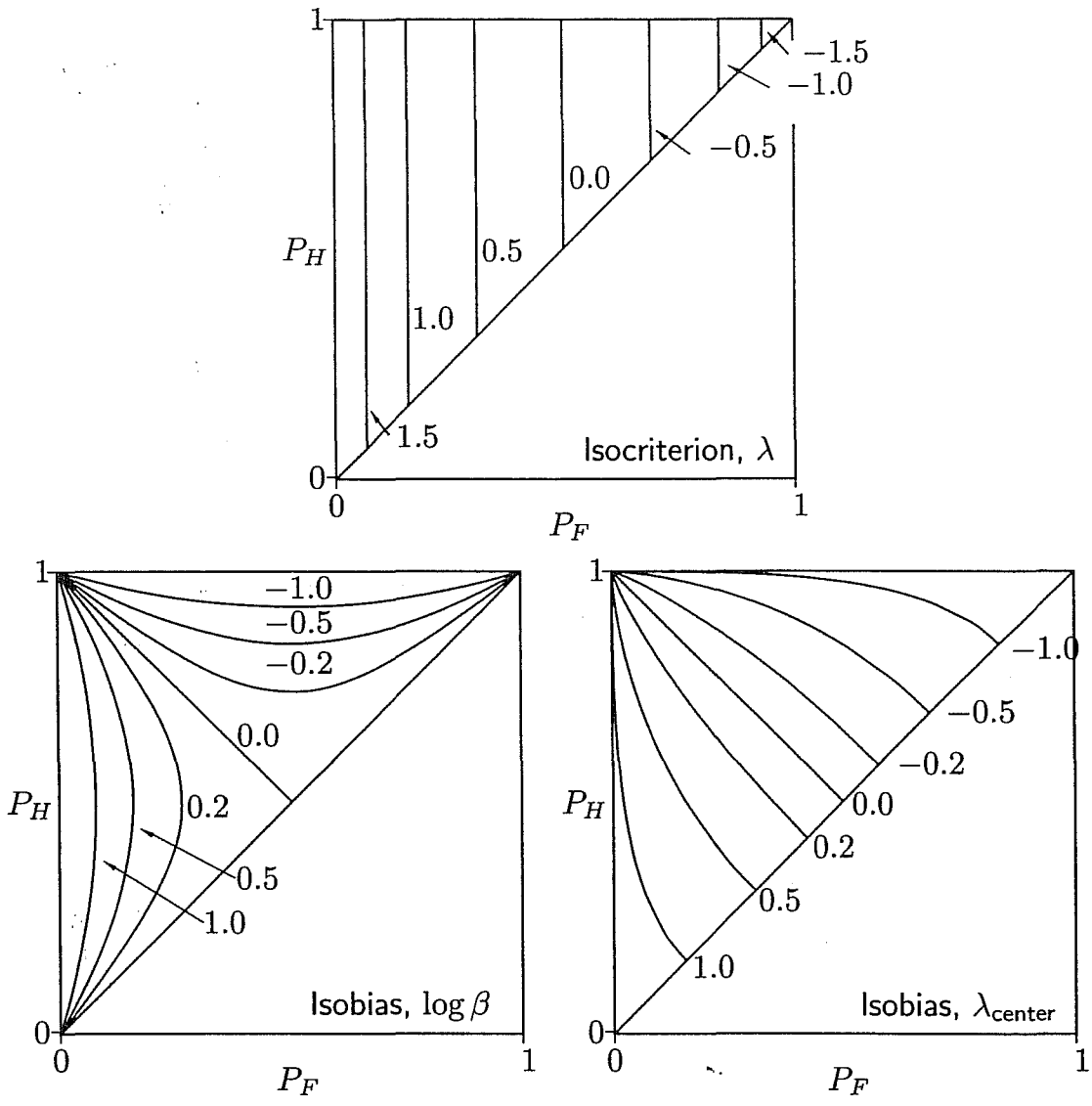

Figure 3.5: Families of isocriterion curves (top) and isobias curves (bottom) for the equal-variance Gaussian model. Isobias curves are shown for β or $\log\beta$ (left) and λ_{center} (right).

more complex psychological construct than is sensitivity. It is easy to understand how the sensitivity of an observer to a particular physical stimulus could remain constant, but it is much harder to assert that bias, by any particular definition, should not vary as the signal changes. Finally, most research that uses detection theory is concerned with factors that influence detectability, not with those affecting the decision process. Differences in response bias in these studies are a nuisance to be removed, not a process to be studied. Determining on which isosensitivity curve the performance falls is most important.

3.3 The equal-variance Gaussian operating characteristic

The actual shape of the operating characteristics is derived theoretically from the distributions of the random variables X_s and X_n. First consider the process in its most general form. Suppose that $f_s(x)$ and $f_n(x)$ are the density functions of these random variables. The hit rate and the false-alarm rate are areas under these curves above the criterion point and are calculated as integrals of the corresponding density functions above those points:

$$P_F = \int_\lambda^\infty f_n(x)\,dx \qquad \text{and} \qquad P_H = \int_\lambda^\infty f_s(x)\,dx \tag{3.1}$$

(Equations 1.1 and 1.2). Denote these quantities, as functions of λ, by $P_F = F(\lambda)$ and $P_H = H(\lambda)$, respectively—here F and H stand for false alarms and hits, respectively, and are not cumulative distribution functions. If one could write these integrals as simple expressions, then one could solve $F(\lambda)$ for λ and substitute it in $H(\lambda)$ to get the isosensitivity curve. Working purely formally:[1]

$$\begin{aligned} \lambda &= F^{-1}(P_F), \\ P_H &= H(\lambda) = H[F^{-1}(P_F)]. \end{aligned} \tag{3.2}$$

Although this procedure is simple in the abstract, a practical difficulty arises with the Gaussian distribution. For it, Equations 3.1 cannot be written as simple expressions, and they cannot be solved algebraically to give a formula for Equation 3.2. Gaussian operating characteristics, such as those in Figures 3.3 and 3.4 are constructed using tables or numerical methods.

The relationship between P_F and P_H implied by the operating characteristic is simplified by transforming the probabilities before they are plotted. The appropriate transformation here is the inverse Gaussian function $Z(p)$. After applying it, the operating characteristic is a straight line. In more detail, first recall (from Equations 2.2 on page 21) that the response probabilities under the Gaussian model with $\sigma_s^2 = 1$ are

$$\begin{aligned} P_F &= 1 - \Phi(\lambda) = \Phi(-\lambda), \\ P_H &= 1 - \Phi(\lambda - d') = \Phi(d' - \lambda). \end{aligned}$$

[1] A function raised to a negative power is the inverse function: if $y = f(x)$, then $x = f^{-1}(y)$. This use of the exponent differs from the notation for positive powers, for which $f^2(x) = [f(x)]^2$ (see footnote 4 on page 30).

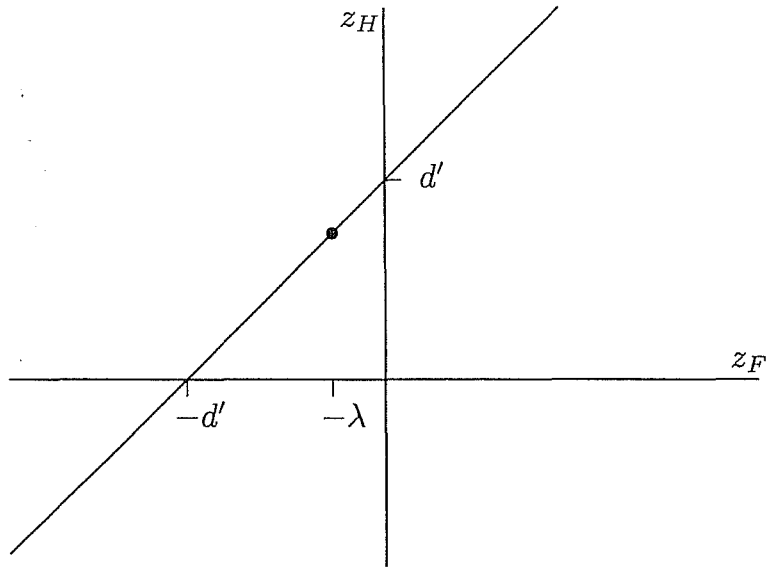

Figure 3.6: The isosensitivity function for the equal-variance Gaussian model plotted in Gaussian coordinates. The solid point corresponds to performance with a criterion of λ.

The formula for the hit rate takes the nonzero mean of X_s into account. Now transform these probabilities to $z_F = Z(P_F)$ and $z_H = Z(P_H)$. This transformation undoes the function $\Phi(z)$:

$$z_F = Z(P_F) = Z[\Phi(-\lambda)] = -\lambda,$$
$$z_H = Z(P_H) = Z[\Phi(d' - \lambda)] = d' - \lambda.$$

Eliminating λ from these equations gives

$$z_H = z_F + d'. \tag{3.3}$$

This equation describes the isosensitivity function when it is plotted on Gaussian transformed axes, that is, in *Gaussian coordinates*.

Figure 3.6 shows the operating characteristic plotted in these coordinates. Four things about it are important to notice:

- The function is linear. Plotted in Gaussian-tranformed coordinates, the isosensitivity function is a straight line. This fact is a consequence of the choice of the Gaussian form for X_n and X_s.
- The function has a 45° slope. This slope is a consequence of the equal-variance assumption. As will be seen in the next section, models with unequal variances have different slopes.
- The line crosses the axes at $-d'$ and d'. Thus, having drawn the line, one can easily read off the value of d' from either intercept or, having d', can easily draw the line.

Probabilities		Gaussian scores	
f	h	$Z(f)$	$Z(h)$
0.12	0.47	−1.18	−0.08
0.18	0.72	−0.92	0.58
0.20	0.58	−0.84	0.20
0.38	0.78	−0.31	0.77
0.51	0.77	0.02	0.74
0.66	0.92	0.41	1.40
0.77	0.96	0.74	1.75

Table 3.1: Hit and false-alarm data obtained by varying the bias at without changing the signal.

- Isocriterion contours are vertical lines at $-\lambda$. So the point corresponding to a model with a particular criterion λ lies at the point of the operating characteristic with a value of $-\lambda$ on the z_F axis.

Because the isosensitivity function in Gaussian coordinates is a straight line, the model is easy to apply to data from several bias conditions. First convert the observed hit and false-alarm proportions to Gaussian coordinates using either tables or a computer program. Then draw a straight line at 45° through these points. The intercept on the ordinate is an estimate of d'. This procedure is discussed in more detail in Section 3.5; for the moment an example will illustrate it.

Example 3.1: Table 3.1 shows the proportions that might be obtained from a seven-level manipulation of bias. Fit the equal-variance Gaussian model to these data and draw the operating characteristic.

Solution: The observed proportions f and h are plotted directly on the left in Figure 3.7. The points roughly trace out an isosensitivity curve, but there is sufficient scatter that an operating characteristic cannot be constructed by connecting them. The resulting line is neither smooth nor monotonic. The scatter is not surprising and is inevitable unless an extremely large number of observations have been obtained at each point, a condition fulfilled only by a few psychophysical experiments.

To fit the function the proportions are converted to Gaussian scores in the second part of Table 3.1 and plotted on the right in Figure 3.7. It is now easy to draw a line at 45° through the midst of them. The points are adequately fitted by this line—certainly they have no systematic curvature—and from the place where this line crosses the axes, $\widehat{d'}$

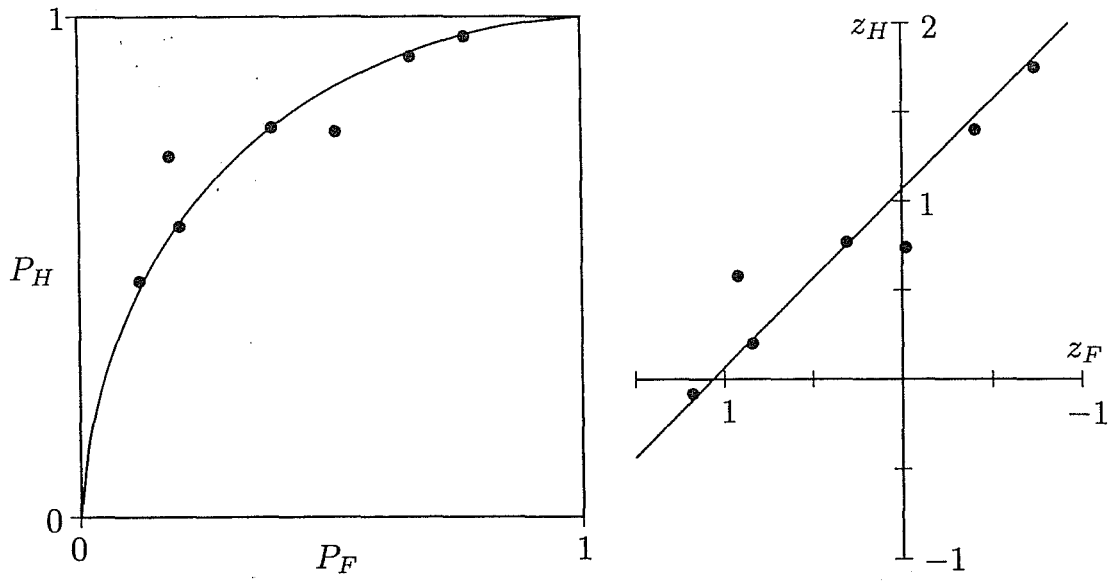

Figure 3.7: Plots of the data in Table 3.1 with fitted operating characteristics in probability coordinates (left) and Gaussian coordinates (right).

is apparently slightly greater than one. A more accurate estimate, obtained from a computer program, is $\widehat{d'} = 1.065$. To plot the isosensitivity function in its conventional form, points on the line in Gaussian coordinates are reconverted to probability coordinates using the transformation $p = \Phi(z)$ giving the theoretical line in the probability plot in Figure 3.7.

3.4 The unequal-variance Gaussian model

In the general Gaussian signal-detection model, the distributions of X_n and X_s can differ in their variance as well as their means. With the constraints placed on parameters of the noise distribution to give the model unique values (as discussed in Section 2.1), the distributions of the random variables associated with the two stimulus events are

$$X_n \sim \mathcal{N}(0,1) \qquad \text{and} \qquad X_s \sim \mathcal{N}(\mu_s, \sigma_s^2).$$

The resulting model is more flexible than the equal-variance version, and when data from several conditions are available, it often provides a superior fit. Figure 3.8 shows an example in which the signal distribution has greater variance than the noise distribution.

Unequal variability of the signal and noise events arises quite naturally in a number of situations. One explanation is based on the rules for the addition of random variables. Consider the detection of a pure tone of

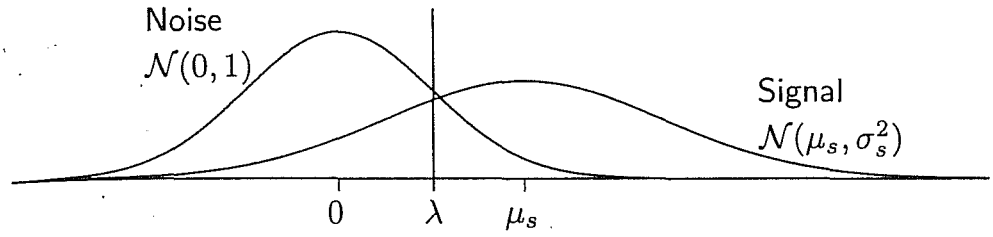

Figure 3.8: Distributions of signal and noise under a Gaussian model in which the standard deviation of the signal distribution is 1½ times that of the noise distribution.

known frequency ν in a white noise background. Suppose that the observer bases the detection on the output of a tuned detector responding to the intensity of the stimulation at the frequency ν. The "noise" here consists of the variation in the extent to which the white noise near the frequency ν excites the detector. This variation creates the variability of X_n. When the tone is added to the background, creating the signal-plus-noise condition, the background variability does not vanish. The added signal increases the response of the detector above that in the noise conditions, say, by an amount S:

$$X_s = X_n + S.$$

If S is a pure constant, without any variability itself, then the random variable X_s is simply a displaced version of X_n, and their variances are the same. However, if the signal has variability, then S is a random variable and its variation combines with that of the noise background. A reasonable assumption here is that S and X_n are independent. The variance of the signal-plus-noise distribution is now the sum of the variances of its parts (Equation A.33 on page 235):

$$\text{var}(X_s) = \text{var}(X_n) + \text{var}(S) \geq \text{var}(X_n).$$

Any variation in the signal or in its detection makes $\text{var}(X_s) > \text{var}(X_n)$. Equality of variance holds only when the signal is a fixed nonrandom quantity.

Another cause of differences in signal and noise variance is a direct consequence of the mechanism by which observations of X_n or X_s are generated. The mean and variance of many random processes are related, so that random variables with larger means also have larger variances. If the response to the stimulus is generated by such a process, then the larger mean of X_s also gives it a greater variance. As a specific example, suppose that X is observed by counting the number of discrete events (perhaps

neural responses) during a fixed interval of time. The rate at which these events occur is larger when the signal is present than when it is not, so the count is, on average, bigger when the signal is present. *Counting processes* of this type have been widely studied. For many of them, the variance of the number of events is approximately proportional to the mean.[2] Although counting processes do not produce true Gaussian distributions, they are usually closely approximated by Gaussian distributions whenever the number of counts is large. Because of the relationship between the mean and the variance, the unequal-variance model must be used.

In other situations one might expect the noise variance to exceed that of the signal. Sometimes the signal event acts to reduce the diversity of an original distribution. Consider a word recognition experiment in which the subject identifies words that were presented in the first part of the experiment, and suppose that the subject does this by estimating how recently he or she has heard the target word. The old, or signal, words have all been presented recently and so have a relatively tight distribution of ages, but the new, or noise, words have a great range of ages. Some words have been seen or heard only a few hours ago, while other words are days, months, or years old. When the words are studied, values from this highly variable distribution are replaced by much more similar values that refer to the experimental presentation.

The unequal-variance Gaussian model depends on three parameters: the two distributional parameters, μ_s and σ_s^2, and the criterion λ. Thus, its parameters cannot be determined by a single yes/no detection study, which yields but two independent results (f and h). Fitting the unequal-variance model requires either several conditions varying in bias or the rating-scale experiment of Chapter 5.

The theoretical operating characteristic for the unequal-variance model is constructed by converting the response rates to Gaussian coordinates as in Section 3.2. As in the equal-variance model, the false-alarm rate depends directly on the criterion:

$$P_F = \Phi(-\lambda) \qquad \text{and} \qquad z_F = -\lambda.$$

The signal distribution now involves its variance. Using the rule for finding area under a normal distribution with nonunit variance (Equation A.46 on page 240),

$$P_H = P(X_s > \lambda) = 1 - \Phi\left(\frac{\lambda - \mu_s}{\sigma_s}\right) = \Phi\left(\frac{\mu_s - \lambda}{\sigma_s}\right).$$

[2]The simplest of the counting processes is the *Poisson process*, in which the counted events are postulated to occur independently at a constant rate. It gives rise (not surprisingly) to Poisson distributions for X_n and X_s. The mean and variance of a Poisson random variable are equal.

Inverting this relationship gives the ordinate of the operating characteristic:

$$z_H = Z(P_H) = \frac{\mu_s - \lambda}{\sigma_s}.$$

Finally, substituting $-z_F$ for λ gives the equation of the isosensitivity function:

$$z_H = \frac{1}{\sigma_s} z_F + \frac{\mu_s}{\sigma_s}. \tag{3.4}$$

This function is linear, with a slope that is the reciprocal of the signal standard deviation. In the case depicted in Figure 3.8, where $\sigma_s = 1.5\sigma_n$, the slope is $1/1.5 = 2/3$.

When the isosensitivity function for a model with unequal variance is translated from Gaussian coordinates back to probability coordinates, the resulting operating characteristic is not symmetric about the minor diagonal of the unit square (the line from lower right to upper left), as it was in the equal-variance case. If $\sigma_s > 1$, then the function has a form like that shown in Figure 3.9. The line rises sharply from (0, 0) as the criterion drops through the signal distribution without reaching much of the noise distribution, then turns more slowly toward the right as the noise distribution is passed while there is still appreciable area under the signal distribution. Eventually the curve approaches (1, 1) when the criterion is below both distributions.

The operating characteristic in Figure 3.9 has several disconcerting features. The most obvious of these is that it dips below the diagonal at the upper right. This dip suggests that for certain criterion positions the false-alarm rate exceeds the hit rate. More subtle is the fact that the slope of the operating characteristic goes from shallower to steeper in this region, a condition that is necessary for the dip to occur. Operating characteristics in which the slope changes nonmonotonically like this are said to be *improper*. Those with curves that progress from (0, 0) to (1, 1) with ever-decreasing slope are said to be *proper*. An improper operating characteristic is a sign that the observer is using a response rule that does not make optimal use of the available information. In the case of the unequal-variance Gaussian model, the improper operating characteristic indicates that a simple criterion applied to the axis is not the best way to make the response. A superior response rule will be described in Section 9.3.

On first being introduced to the unequal-variance Gaussian model, one is inclined to worry about the dip below the diagonal and the apparently perverse behavior that it implies. However, to fret much about it is to take the model too seriously. The model is an useful description of detection behavior, but cannot be taken as mathematical truth. There are many

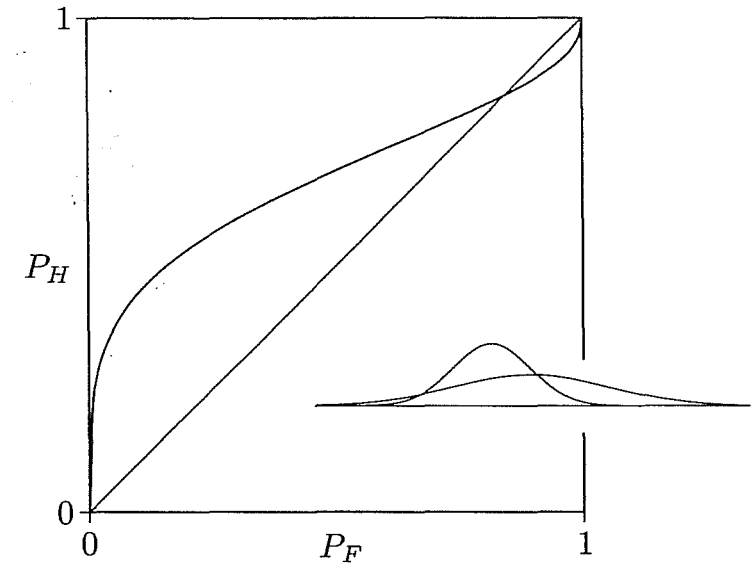

Figure 3.9: The isosensitivity function for the unequal-variance Gaussian model with $X_n \sim \mathcal{N}(0,1)$ and $X_s \sim \mathcal{N}(1,4)$. The inset shows the distributions of X_n and X_s.

circumstances in which the nonunit slope of a transformed operating characteristic implies that an unequal-variance representation is appropriate, but where no evidence of the dip is present. In fact, to observe the dip in real data is almost impossible. It would occur only for conditions where the observer was so biased toward YES responses that both the hit rate and the false-alarm rate were almost, but not quite, equal to one. An accurate determination of such an extreme point would require an enormous sample of data. It is doubtful that an observer could be induced to perform in a stable way in such extreme conditions for the requisite number of observations.

3.5 Fitting an empirical operating characteristic

Fitting the unequal-variance signal-detection model requires a set of conditions that differ in bias but not in the detectability of the signal. These conditions can be created by keeping the signal and noise events the same while inducing the observer to alter the criterion. There are several ways to produce this shift. Practiced observers can respond to instructions to change their detection performance in a way that increases or decreases the hit rate or false-alarm rate. Less trained observers will change the criterion in response to the proportion of signals. The analysis of the ideal observer in Section 2.5 indicated that bias should be shifted with the signal frequency

to maximize correct responses (Equation 2.15). Real observers shift their performance in this direction, although usually not to the optimal extent. When signals are rare, fewer YES responses are made, lowering both the hit rate and the false-alarm rate. When signals are common, the bias is shifted to increases the number of YES responses. Thus, running sessions with different proportions of signals will give different points on the operating characteristic. Bias can also be manipulated by assigning payoffs to the various types of responses, as in the example of Section 1.2. Each of these manipulations concerns the observer's response behavior only and can typically be made without substantially altering the sensitivity.

The first thing to do with such data is to look them over for anomalies or inconsistencies that could indicate something wrong with the study. Plot the hit rate against the false-alarm rate and verify that the general form of an operating characteristic is obtained. Problems are indicated by points for which the false-alarm rate exceeds the hit rate or by a failure of the order of the conditions on the operating characteristic to match the manipulation used to shift the bias. The sampling fluctuation associated with any set of empirical observations can lead to some reversals or to points where $h < f$, particularly when the amount of data is small, but large discrepancies are unlikely. When they happen, it is possible that the observer misunderstood the task or was behaving perversely, but that is unlikely in a well-run study. A more common cause of these irregularities is a mistake by the researcher in recording the data or in entering them into a computer program for analysis. Such errors should be ruled out before taking an irregular point too seriously.

When the data have been deemed satisfactory, turn to the analysis. It is helpful to organize the analysis in five steps:

1. Convert the data from hit and false-alarm rates h and f to the transformed values $Z(h)$ and $Z(f)$, using the table of the Gaussian distribution on page 250 or an equivalent computer program. Plot these points as an operating characteristic in Gaussian coordinates. Be accurate—it helps to use a full sheet of graph paper with closely spaced grid lines.
2. Evaluate the Gaussian model. If the points fall in a straight line (except for what can be deemed sampling error), then the use of this model is justified. If they curve, then consider a model based on another distribution. Under most circumstances a visual evaluation is sufficient (formal statistical tests are covered in Section 11.5). Usually, the Gaussian model will be adequate.
3. Fit a straight line to the set of points. Be sure to try the line at 45° that corresponds to the equal-variance model. When there is little

scatter in the data, the line can be drawn by eye without serious error. When the data are more scattered, some form of statistical fitting procedure is better, although rough estimates still can be made by eye. A transparent 45° drafting triangle is very useful here. Decide whether the 45° is adequate (some statistical procedures are discussed in Section 11.6). Otherwise conclude that the variances are unequal and that a line of some other slope is needed.

4. Estimate the parameters of the signal distribution from the fitted line. If the chosen line has a slope of one, then estimate d' by the intercept of the function, as described in Section 3.2. If the line does not lie at 45°, then find the slope b and intercept a of the equation

$$Z(h) = bZ(f) + a.$$

Match the slope and intercept to Equation 3.4 to estimate the Gaussian model's parameters:

$$\widehat{\mu}_s = a/b \quad \text{and} \quad \widehat{\sigma}_s = 1/b. \tag{3.5}$$

When working from a graph on which the line has been drawn, it is easier to forget about the slope and intercept and instead note the points x_0 and y_0 where the line crosses the horizontal and vertical axes, respectively. The parameters of the line are

$$a = y_0 \quad \text{and} \quad b = -y_0/x_0, \tag{3.6}$$

and those of the detection model are

$$\widehat{\mu}_s = -x_0 \quad \text{and} \quad \widehat{\sigma}_s = -x_0/y_0. \tag{3.7}$$

5. To estimate the criterion λ for a condition, the observed point must be translated to one on the fitted line. A full analysis here takes account of the accuracy with which each coordinate is observed. However, for most purposes it is sufficient to take the bias from the point on the operating characteristic that is closest to the observation. Frequently this can be done by sketching a line perpendicular to the line that goes through the point. Details and formulae are given in Section 4.5. Once the point is chosen, $\widehat{\lambda}$ is minus the abscissa of the point.

Example 3.2: Three carefully measured detection conditions give the pairs of false-alarm and hit rates (0.12, 0.43), (0.30, 0.76), and (0.43, 0.89). Does a Gaussian model fit these data? If it does, then estimate its parameters.

Solution: First sketch the data in probability coordinates (not shown) and look for irregularities. Here there are none. Then transform the probabilities to Gaussian coordinates:

f	h	$Z(f)$	$Z(h)$
0.12	0.43	−1.17	−0.18
0.30	0.76	−0.52	0.71
0.43	0.89	−0.18	1.23

The transformed points are plotted in Figure 3.10. A line at 45° (dashed in the figure) is clearly unsatisfactory, so the equal-variance Gaussian model does not apply. However, they lie so close to another straight line that it can be drawn through them by eye (the solid line in the figure). Evidently, the unequal-variance Gaussian model fits these data. Because the slope is greater than one, the signal distribution has a smaller variance than the noise distribution. Reading the graph carefully, the line crosses the horizontal axis at $x_0 = -1.04$ and the vertical axis at $y_0 = 1.46$. The slope and intercept of the line are found from these crossing points (Equations 3.6):

$$a = y_0 = 1.46 \quad \text{and} \quad b = \frac{y_0}{-x_0} = \frac{1.46}{1.04} = 1.40.$$

These values can be converted to estimates of the parameters of the signal distribution using Equations 3.5, or they can be found directly from the crossing points (Equations 3.7):

$$\widehat{\mu}_s = -x_0 = 1.04 \quad \text{and} \quad \widehat{\sigma}_s = \frac{-x_0}{y_0} = \frac{1.04}{1.46} = 0.71.$$

Finding the criteria associated with the conditions is simple here. The line fits so accurately that its closest approaches to the three points are at the abscissas already calculated as $Z(f)$. Accordingly, the criteria are 1.17, 0.52, and 0.18.

The problem of finding the best-fitting line deserves comment. The fact that one is fitting a straight line to a group of points makes it tempting to use the simple fitting equations of ordinary linear regression. One would write z_H as a linear function of z_F and calculate the slope and intercept from the usual regression equations. Unfortunately, this procedure gives the wrong line and systematically biased estimates of the parameters. The difficulty lies in the way that the regression line is defined. The model for linear regression treats one variable x as a predictor and the other variable y as an outcome to be predicted from x. The outcome variable y is represented theoretically as a linear function of the predictor plus a random error e:

$$y = \alpha + \beta x + e.$$

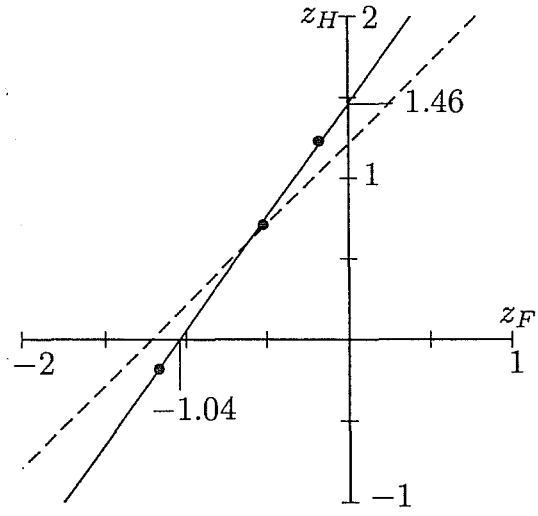

Figure 3.10: A three-point operating characteristic for Example 3.2 plotted in Gaussian coordinates. The dashed line at 45° is unsatisfactory, but the solid line from the unequal-variance model fits well.

In this regression model, x is an exact number, and all the sampling uncertainty is attributed to y. This model is a unsatisfactory representation of detection data, for which both $Z(h)$ and $Z(f)$ are subject to sampling error.

The use of the regression model here would not create problems if it did not bias the estimates of the detection parameters. The problem is a phenomenon known as *regression to the mean*. Geometrically, variability in the data acts to flatten the regression line, so that the best prediction of y is nearer (in standard deviation units) to the mean of that variable than the predictor x is to its mean. The greater the variability of the data, the more the line is flattened. Shrinking the predictions toward the mean is appropriate for the asymmetric measurement structure of regression analysis, but is incorrect for detection data. The line regressing $Z(h)$ on $Z(f)$ has a smaller slope than does a line that treats both $Z(h)$ and $Z(f)$ as subject to sampling error. Unless the data are almost error free, a regression analysis will give too small a value for b and consequently overestimate $\widehat{\sigma}_s$. When there is considerable scatter in the data, this bias can lead the equal-variance model to be rejected when in fact it is appropriate.

3.6 Computer programs

The amount of calculation involved in fitting the signal-detection theory model makes it a good candidate for computerized calculation. The calculations for a single detection condition are easy to implement. Algorithms

for the functions $Z(f)$ and $Z(h)$ are well established, and their values are directly available in some higher level languages. Using them, Equations 2.3 and 2.4 can be calculated directly.

When three or more conditions with different bias are to be fitted, the task is considerably harder. It is no longer possible to convert the proportions directly into parameter estimates, but a program must, in effect, find the operating characteristic that fits the data best, even though it may not pass exactly through any of the observed points. However, because the probabilistic structure of the signal-detection model is well defined, it can be fitted with the standard statistical technique known as maximum-likelihood estimation. Several programs that make these calculations have been published (see reference notes). These usually report most or all of the statistics that will be discussed in Chapter 4. One of the great advantages of the maximum-likelihood procedure is that it gives values for the standard errors of the parameter estimates. The use of these estimates will be discussed in Chapter 11.

When using one of these programs, it helps to know a little about what they are doing. There are no equations that directly give estimates of the parameters of the Gaussian model in terms of the data (e.g., as there are in multiple regression). Instead, the signal-detection model is fitted by an iterative algorithm. The programs start by making some guess at the parameter values—not necessarily a very good one. Then it adjusts these values to improve the fit. This process of adjustment, or iteration, is repeated (generally out of sight of the user) until the estimates cease to improve, at which time the results are reported and the values of any summary measures are calculated. Many programs report the number of iterations required to complete the process or the criterion of change used to decide when to stop.

For most sets of data, this procedure runs successfully and delivers good estimates. However, with certain very irregular sets of data, the algorithm can break down and fail to find a solution—a failure to converge, as it is called. Exactly how the program handles this contingency depends on the implementation, and some versions of the algorithm are more robust than others. Failure of the estimates to converge is uncommon with signal-detection data. It most often occurs when the observations are very much inconsistent with the signal-detection models, for example, with points lying very far from a line in Gaussian space. Thus, when a program fails to converge, it is advisable to review the data and see if they have been entered incorrectly or if they are sufficiently at odds with the signal-detection model that any parameters obtained by fitting that model will be meaningless. If it is necessary find estimates for such data nonetheless, it may be possible to revise the starting point for the search to one from which it will converge.

Fitting a line to the data by eye may give a good place to start. It is also worth trying another program, as minor differences in the way the program selects its starting point or implements the iterative algorithm give them different sensitivities.

Reference notes

Some approximations for calculating $\Phi(z)$ and $Z(p)$, sufficiently accurate for calculation of d' and the like, are given by Zelen and Severo (1964). The standard maximum-likelihood estimation algorithm used to fit the signal-detection model to several conditions was originally published by Dorfman and Alf (1968a, 1968b). The background to these methods are found are discussed in most advanced statistics texts. Several implementations of the procedure have been published, and a summary of programs is given in Swets (1996). I will make my version of these programs available on the web site mentioned in the preface.

Exercises

3.1. Suppose that $\mu_s = \sigma_s = 1.5$ in the Gaussian signal-detection model.

a. Draw the isosensitivity function in Gaussian coordinates.

b. Convert this function to an operating characteristic in probability coordinates.

3.2. Two detection conditions give estimates of $\widehat{d}'_1 = 1.05$, $\widehat{\lambda}_1 = 1.03$ and $\widehat{d}'_2 = 1.70$, $\widehat{\lambda}_2 = 0.38$. Use these results to fit an unequal-variance model to the pair of conditions. Draw the operating characteristic.

3.3. Six detection conditions (A to F) are run with the same signal and observer but with the bias manipulated. The response frequencies observed are

	Noise		Signal	
	YES	NO	YES	NO
A	264	36	294	6
B	168	132	273	27
C	102	198	252	48
D	30	270	198	102
E	17	283	171	129
F	2	298	108	192

a. Estimate d' and $\log \beta$ for each condition, based on the equal-variance Gaussian model. Is there a pattern to these values? Why is this analysis unsatisfactory?

b. Sketch (by eye) an isosensitivity function for these data in standard coordinates.

c. Plot the data in Gaussian coordinates and draw a straight line by eye through them.

d. Use your line to decide whether $\sigma_s^2 = \sigma_n^2$.

e. Estimate the parameters of whichever model you chose in the previous step.

3.4. Use a program to fit the data from Problem 3.3. Compare the results to the estimates made in this problem.

3.5. Suppose that the hit rate and false-alarm rate for detection of visual stimulus are collected from five different subjects. Explain why it would be inappropriate to use the five points to create an operating characteristic as described in this chapter.