Chapter 5

Confidence ratings

The last two chapters have argued that the most robust measures of the detectability of a signal are those calculated from the isosensitivity contour. The largest difficulty in using them routinely is the demands that they place on the observer. Enough data must be collected at each of several levels of bias to get accurate estimates of the hit rate and the false-alarm rate. These data are not hard to obtain from a practiced observer who can devote the required time to the task and who can maintain a stable performance level as the bias is changed. They are harder to get from observers who have limited availability or from inexperienced observers who cannot maintain consistent performance throughout the study. These difficulties, for example, almost preclude using bias manipulations in the typical learning or memory experiment.

One way to retain the advantages of an analysis based on the isosensitivity contour while alleviating some of these problems is to increase the amount of data collected on each trial. That can be done by combining each detection response with an indication of the observer's confidence that the response is correct. The confidence is taken as providing evidence about where on the underlying decision axis the effect of that stimulus falls. By using these ratings, information that bears on several points of the isosensitivity contour can be collected on each trial. The analysis of such experiments is the topic of this chapter.

5.1 The rating experiment

A study that uses confidence ratings is conducted exactly like a yes/no detection study in all aspects except the response. In the rating task, each trial is made more informative by asking the observer to provide a

	Response					
	N3	N2	N1	Y1	Y2	Y3
Noise	166	161	138	128	63	43
Signal	47	65	66	92	136	294

Table 5.1: Three-level confidence ratings from a single practiced observer detecting a weak visual pattern. The data come from an experiment by L. A. Olzak and P. Kramer, and I thank Lynn Olzak for their use.

confidence rating along with the response. For example, an observer could follow the response of YES or NO by assigning one of three categories: SURE, UNCERTAIN, or GUESSING. The result is an ordered range of responses indicating increasing certainty about the presence of the signal. At one end of this *rating scale* is the NO-SURE response, then NO-UNCERTAIN and NO-GUESSING, then the YES responses starting with from YES-GUESSING and ending with YES-SURE. In effect, six response categories are obtained from the observer instead of two.

It is not necessary to collect the yes/no response and the confidence rating as two separate entities. One could equally well have the observer give a single number that ranges between 1, which indicates high confidence that no signal was present, and 6, which indicates high confidence that it was. The same six ordered levels are produced.

Table 5.1 shows an example of some data collected from a rating-scale experiment. In this study a well-practiced observer attempted to detect a faint regular brightness modulation of a uniform visual field (a CRT screen). The Noise condition corresponded to an unmodulated field; the Signal condition, to a pattern of faint vertical bands of light and dark strips produced by a sinusoidal brightness modulation along the horizontal axis of the display. A three-level rating scale, like the one just described, was conjoined with the yes/no decision to produced six categories. In Table 5.1 the detection responses are denoted by Y or N and the ratings by a number from 1 to 3.

A quick examination of Table 5.1 reveals three things. First, the response frequencies for the two stimuli clearly differ, so the signal was to some degree detectable. Second, the rating scale provides information beyond that in a response of YES or NO. An N3 rating is more likely to be from a noise trial than is an N1 rating. A similar gradient is present for the YES responses. Third, the observer's performance was imperfect. All types of responses were made to both stimuli, and even high-confidence NO responses to signals and high-confidence YES responses to noise are moderately frequent. This pattern is consistent with a moderately difficult detection task.

Although the data in Table 5.1 are substantially richer than simple yes/no data, they can be reduced to these two categories. Ignoring the rating, there are 166 + 161 + 138 = 465 responses of NO to the blank field (noise) stimulus. Pooling the three levels of confidence for the other responses and for the signal stimulus gives a table classified only by YES and NO:

	NO	YES
Noise	465	234
Signal	178	522

(5.1)

The detectability of the signal is evident in the greater number of YES responses to it than to the noise. A simple analysis like that in Chapter 2 gives $\widehat{d'} = 1.09$ with $\widehat{\lambda} = 0.43$. However, by collapsing this way, the grading of the rating data is lost. It is this graded character that gives the rating experiment its power.

5.2 The detection model for rating experiments

The analysis of rating-scale data is based on the same representation of the stimuli that was used to describe the yes/no experiment. There is a single evidence dimension X, and the observation of a stimulus produces a value on that dimension distributed according to the random variables X_n and X_s. In most applications, these random variables are given Gaussian distributions with different means and possibly different variances:

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

The only change in the model is in the response process, which must be able to describe the confidence ratings. Instead of a single criterion that separates the response NO from the response YES, there is a series of criteria that separate the confidence levels. Figure 5.1 shows the model for six-level responses. Five criterion lines divide the range of X into six parts. Each section of the abscissa is associated with a different compound response. The response to any value of X below the criterion λ_3 is NO and that to any value above λ_3 is YES. Within the NO range, the criteria λ_1 and λ_2 divide the confidence levels. Observations below λ_1 are given a rating of 3, those between λ_1 and λ_2 are given a rating of 2, and those between λ_2 and λ_3 are given a rating of 1. Similarly, the criteria λ_4 and λ_5 divide the

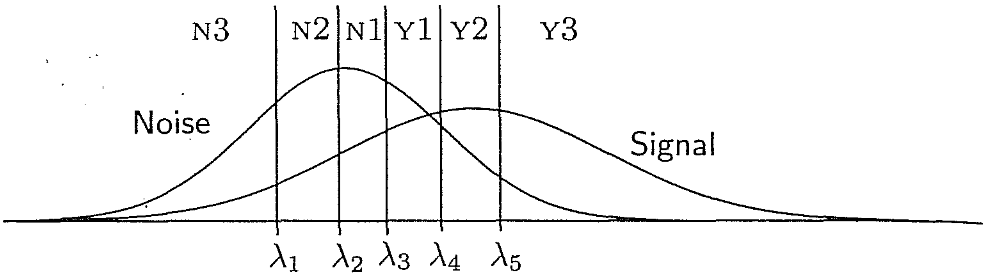

Figure 5.1: Signal and noise distributions with five criteria separating the six response categories generated by three-level confidence ratings in a detection task.

YES responses into three confidence levels. The most confident responses are made when large or small values of X are observed—those below λ_1 or above λ_5—and the least confident responses are made at intermediate values between λ_2 and λ_4.

This description is easily written in formal notation. Suppose that the rating experiment divides the responses into J categories. Let R be a discrete random variable that denotes the response that the observer makes. Thus, for Table 5.1, $J = 6$ and R takes values 1, 2, 3, 4, 5, or 6, corresponding to the responses from N3 through Y3. Assign end criteria that are very far to the left or right, $\lambda_0 = -\infty$ and $\lambda_J = \infty$. Now the probability of response $R = j$ is the area under the appropriate density function between the criteria λ_{j-1} and λ_j:

$$\begin{aligned} P(R{=}j|\text{noise}) &= \int_{\lambda_{j-1}}^{\lambda_j} f_n(x)\,dx, \\ P(R{=}j|\text{signal}) &= \int_{\lambda_{j-1}}^{\lambda_j} f_s(x)\,dx. \end{aligned} \tag{5.2}$$

This representation differs from the one used for the yes/no task, only in that the integrals are bounded on both sides (compare Equations 5.2 to Equations 1.1 and 1.2).

For the Gaussian model, the integrals are replaced by differences in the Gaussian cumulative distribution function $\Phi(z)$. Using the parameters of the model or the slopes of the isosensitivity contour, Equations 5.2 become

$$P(R{=}j|\text{noise}) = \Phi(\lambda_j) - \Phi(\lambda_{j-1}), \tag{5.3}$$

$$P(R{=}j|\text{signal}) = \Phi\left(\frac{\lambda_j - \mu_s'}{\sigma_s}\right) - \Phi\left(\frac{\lambda_{j-1} - \mu_s'}{\sigma_s}\right) \tag{5.4}$$

$$= \Phi(b\lambda_j - a) - \Phi(b\lambda_{j-1} - a). \tag{5.5}$$

From these equations, one can calculate response probabilities for any set of parameters.

The collection of criteria in Figure 5.1 suggests one way to analyze the experiment. The two-by-two table of Display 5.1 was obtained by pooling the responses corresponding to the areas on either side of the criterion λ_3. Responses N3 through N1 were combined in the NO category, and responses Y1 through Y3 were combined in the YES category. Similar two-by-two tables are created by pooling the observations on either side of any other division between categories. In this way, five two-by-two tables are obtained from the six-level categorization, one corresponding to each criterion in Figure 5.1. Table 5.2 shows these tables, with their hit rates and false-alarm rates. One can then treat these five tables as if they were the results of five different bias manipulations, giving five (f, h) pairs. These points are plotted as a five-point operating characteristic in Figure 5.2, using both probability and Gaussian coordinates.

Two facts immediately emerge from an inspection of Figure 5.2. First, the transformed points fall very close to a straight line, which implies that a Gaussian model is satisfactory. Second, the slope of this line is less than 45°, which implies that the unequal-variance model is necessary and that $\sigma_s^2 > 1$. The inadequacy of the unequal-variance model is also evident in the asymmetry of the probability operating characteristic.

The points on the isosensitivity contour obtained from a rating-scale study differ from those created by bias manipulations in several respects. Each point in a bias study derives from its own set of observations, while those in a rating study all come from the same observations. This difference has several consequences. The most obvious of these is that the rating-scale observations usually appear more stable than those derived from separate conditions. The way that the rating points are created by cumulating responses means that the empirical operating characteristic cannot have the sort of random nonmonotonic leaps and drops that appeared with the bias data in Figure 3.7 on page 48. It is easier to fit a straight line to rating data than to bias data, particularly by eye. Another nice property of rating data is that the points are not differentially affected by practice or fatigue effects. Even when there is some drift in the observer's performance, it affects all the points equally, not just those recorded later in the study.

The common origin of the points in the rating isosensitivity contour does mean that they are not statistically independent. For example, if it happened that the number of noise stimuli given N3 responses was unusually high in a particular study, then all points on the contour are pushed to the right. In contrast, the points of a bias-manipulated operating characteristic are independent. The principal effect of this dependence is technical: it must be accommodated in the algorithms used to fit the model. It also

	N3	N2-Y3	
Noise	166	533	$f = 0.762$
Signal	47	653	$h = 0.933$

	N3-N2	N1-Y3	
Noise	327	372	$f = 0.532$
Signal	122	588	$h = 0.840$

	N3-N1	Y1-Y3	
Noise	465	234	$f = 0.335$
Signal	178	522	$h = 0.746$

	N3-Y1	Y2-Y3	
Noise	593	106	$f = 0.152$
Signal	270	430	$h = 0.614$

	N3-Y2	Y3	
Noise	656	43	$f = 0.061$
Signal	406	294	$h = 0.420$

Table 5.2: Five two-by-two tables created by collapsing the rating data in Table 5.1.

increases the sampling variability of the parameter estimates over those for an operating characteristic based on bias manipulation with the same number of observations per point (although, of course, the bias study will require more observations overall).

5.3 Fitting the rating model

Because the underlying representation of the stimuli as two distributions is the same for the yes/no and the rating-scale models, the same measures of detectability are used with each. All the discussion in Chapter 4 applies. When the observations are regular enough to fit an operating characteristic by eye, then the measures are calculated as before. A computerized fitting algorithm for rating-scale data must take account of the fact that the points are not independent of each other, as they would be for a bias manipulation.

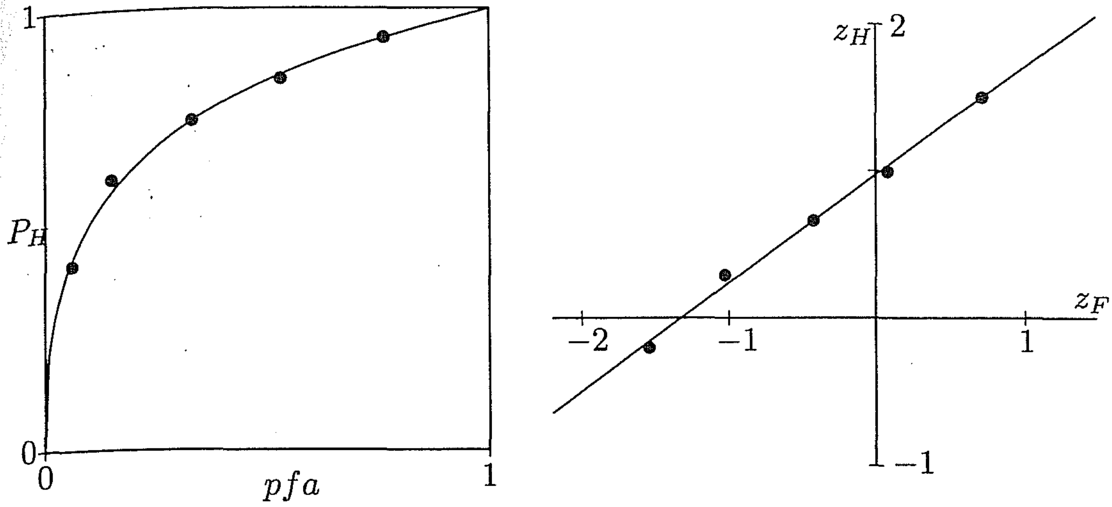

Figure 5.2: Points corresponding to the five two-by-two tables in Table 5.2 that were created by collapsing the data of Table 5.1. The isosensitivity contours are derived from the Gaussian model fitted below.

Programs of both types are available, and one should be sure to use the correct program for one's type of data.

Table 5.3 shows how the data in Table 5.1 is fitted in a form that accommodates the structure of a rating-scale study. The original response frequencies are entered in the columns labeled x_{nj} and x_{sj}. The frequencies of ratings higher than the category are cumulated in the columns $\sum x_{nk}$ and $\sum x_{sk}$ (the limits on these sums are from $j+1$ to 6). These entries are easily built up from the bottom: each is obtained by summing the two entries on the line below. So, $63 + 43 = 106$ and $128 + 106 = 234$. At the top of the table, this procedure gives the total number of trials of that type. In the next columns of Table 5.3, these sums are divided by the total frequencies to give proportions corresponding to false alarms f_j and hits h_j. These calculations repeat those in Table 5.2 in more compact form. The proportions are then converted to Gaussian coordinates, which are the values plotted in Figure 5.2. Besides being more compact, a display like Table 5.3 makes clear the dependence of the rows on each other, which the two-by-two configurations of Table 5.2 do not.

Example 5.1: Fit the rating model to the data in Table 5.1.

Solution: The transformed points in Figure 5.2 fall so close to a straight line that a fit by eye is nearly as accurate as a computer program. The straight-line isosensitivity contour in Gaussian coordinates is

$$z_H = 0.735 z_F + 0.974.$$

The criteria are calculated from the abscissas of this line at the fitted

	Noise				Signal			
j	x_{nj}	$\sum x_{nk}$	f_j	$Z(f_j)$	x_{sj}	$\sum x_{sk}$	h_j	$Z(h_j)$
		699				700		
1	166	533	0.762	0.714	47	653	0.933	1.498
2	161	372	0.532	0.081	65	588	0.840	0.994
3	138	234	0.335	−0.427	66	522	0.746	0.661
4	128	106	0.152	−1.030	92	430	0.614	0.290
5	63	43	0.061	−1.542	136	294	0.420	−0.202
6	43				294			

Table 5.3: Calculation of Gaussian isosensitivity coordinates for the data in Table 5.1.

points, as in Figure 4.7 on page 77. The five criterion values, λ_1 to λ_5, are

$$-0.714, \qquad -0.067, \qquad 0.423, \qquad 0.979, \qquad \text{and} \qquad 1.590.$$

From here on, the analysis is like that for a set of independent conditions. Following Equations 3.5, the intercept and slope are converted to the parameters of the Gaussian model:

$$\widehat{\mu}_s = \frac{a}{b} = \frac{0.974}{0.735} = 1.325,$$
$$\widehat{\sigma}_s = \frac{1}{b} = \frac{1}{0.735} = 1.360.$$

The four summary detection measures discussed in Chapter 4 are

$$\Delta m = \widehat{\mu}_s = \frac{a}{b} = 1.325,$$
$$d_e = \frac{2a}{1{+}b} = 1.123,$$
$$d_a = \frac{\sqrt{2}\,a}{\sqrt{1{+}b^2}} = 1.110,$$
$$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) = \Phi(d_a/\sqrt{2}) = \Phi(0.784) = 0.784$$

(by happenstance, the value of $\Phi(z)$ and its argument are identical here). These statistics can be used to compare this result to those measured in other conditions.

When the rating scale has more than three levels, there are more free observations in the data than are required to estimate the parameters. Each observed point of the isosensitivity contour (or each line of Table 5.3) reflects two data values. So, the six-level distributions in Table 5.1, which yield five points, have 10 free values. There are seven parameters in the model, μ_s, σ_s^2, and the five criteria λ_1 through λ_5. The excess of data over parameters makes it possible to test the fit of the model. Both the adequacy of the Gaussian assumption and the need for unequal variance can be investigated. These tests are described in Section 11.5 (and specifically for these data in Examples 11.8 on page 214 and 11.9 on page 218). They show that, up to the limits of the power of the test, the description of these data by the Gaussian model is satisfactory.

Exercises

5.1. A rating experiment is run with 150 noise trials and the same number of signal trials. A five-level rating scale is used for the responses. The numbers of responses of each type are

	1	2	3	4	5
Noise	42	49	20	24	15
Signal	10	20	15	39	66

a. Plot the operating characteristic, fit a line to it by eye.
b. Using this line, decide whether the Gaussian model acceptable. Can the variances of X_n and X_s be assumed to be the same?
c. Estimate the area A_z under the Gaussian operating characteristic.
d. If you have a computer program for fitting rating data, apply it and compare the results to your estimates.

5.2. A recognition memory study is run as follows. Subjects read a short (two page) story and are asked questions about its content. After half an hour of irrelevant activity, they are given a sheet containing 80 words. Half of these words had appeared in the story; the other half had not been used in any part of the experiment. The subjects are asked to circle those words that they remembered having been used and to put a check mark next to those words (either circled or not) for which they were very sure of their answer. The study involved various between-subject conditions, but these are not the concern of this problem, which deals with how to summarize each subject's data. The responses of one subject in this study are

	NO ✓	NO	YES	YES ✓	
New	15	13	7	5	40
Old	4	6	7	23	40

a. Plot the operating characteristics implied by these data.

b. Do these data support the use of a Gaussian model? Specifically, the equal-variance model?

c. Summarize each subject's recognition performance by a number between 0 and 1.