**University of Birmingham** **School of Psychology**

Postgraduate research methods course - Mark Georgeson

## Sensitivity and Bias - an introduction to Signal Detection Theory

**Aim** To give a brief introduction to the central concepts of Signal Detection Theory and its application in areas of Psychophysics and Psychology that involve detection, identification, recognition and classification tasks. The common theme is that we are analyzing decision-making under conditions of uncertainty and bias, and we aim to determine how much information the decision maker is getting.

**Objectives** After this session & further reading you should:

- be acquainted with the generality and power of SDT as a framework for analyzing human performance
- grasp the distinction between sensitivity and bias, and be more aware of the danger of confusing them
- be able to distinguish between single-interval and forced-choice methods in human performance tasks
- be able to calculate sensitivity d' and criterion C from raw data

*Key references*

N A Macmillan & C D Creelman (1991) *"Detection Theory: A User's guide"* Cambridge University Press (out of print, alas)

Green DM, Swets JA (1974) *Signal Detection Theory & Psychophysics* (2nd ed.) NY: Krieger

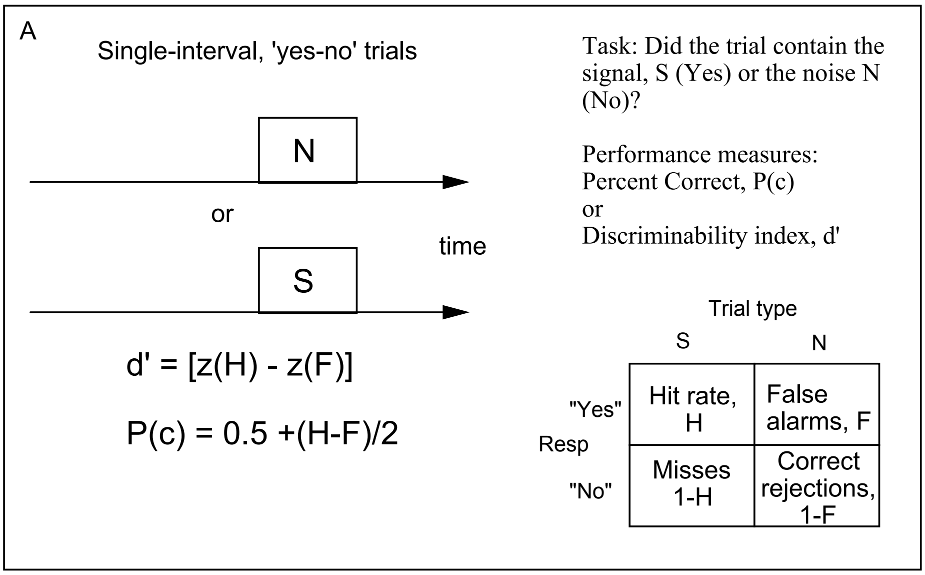
*Illustrative papers*

Azzopardi P, Cowey A (1998) Blindsight and visual awarenss. *Consciousness & Cognition* 7, 292-311.

McFall RM, Treat TA (1999) Quantifying the information value of clinical assessments with signal detection theory. *Ann. Rev. Psychol.* 50, 215-241. [ free from http://www.AnnualReviews.org ]

Single-interval and forced-choice procedures

Fig.1



$$d' = [z(H) - z(F)]$$

$$P(c) = 0.5 + (H-F)/2$$

Task: Did the trial contain the signal, S (Yes) or the noise N (No)?

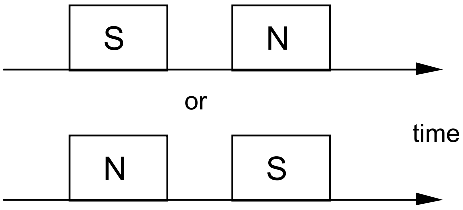
Performance measures:
Percent Correct, P(c)
or
Discriminability index, d'

Trial type

| Resp | S | N |
|---|---|---|
| "Yes" | Hit rate, H | False alarms, F |
| "No" | Misses 1-H | Correct rejections, 1-F |

B

2 alternative forced-choice trials



Task: Which interval contained the signal, S ?

Performance measures:
Percent Correct, P(c)
or
Discriminability index, d'

$$d' = [z(H) - z(F)]/\sqrt{2}$$

$$P(c) = 0.5 + (H-F)/2$$

| | | Trial order | |
|---|---|---|---|
| | | S-N | N-S |
| Resp | "1st" | Hit rate, H | False alarms, F |
| | "2nd" | 1-H | 1-F |

## 1. Introduction

*Example 1* Suppose I'm interested in knowing whether people can detect motion to the right better than to the left. I set up an experiment where faint dots move left or right at random on different trials. Each observer does lots of trials responding 'right' or 'left' on each trial, and I tally the results. I find that people are 95% correct on rightward trials (they say 'right' on 95% of trials when motion was rightward) but only 60% correct on leftward trials. The difference is significant by some suitable test. Am I justified in concluding that people really are better at rightward motion? If not, why not?

*Example 2* Suppose I have invented a fancy computerized method of recognizing tumours in X-ray plates. I want to know whether the method is better than doctors can do by intuition and experience. I create a series of test plates, 100 with tumours, 100 without, and then test the doctors and my machine. The doctors get 80% correct for plates with tumours, and 80% correct without. The machine gets 98% correct with tumours, and 62% correct without. Thus average performance is 80% correct for doctors and for my gizmo. Does this mean both methods equally good ? Or is the machine better because it hardly misses any tumours? Or is it worse because it gives more false positives (38% to the doctors' 20%), which may be alarming to patients and cause unnecessary surgery ?

*Table 1*

Doctors' performance

| | Signal | |
|---|---|---|
| | Present | Absent |
| "Yes" | 80 | 20 |
| "No" | 20 | 80 |
| | p(Hit) | p(FA) |
| | 0.800 | 0.200 |
| | z(Hit) | z(FA) |
| | 0.842 | -0.842 |
| Sensitivity, d' = | 1.683 | |
| Criterion, C = | 0.000 | |
| P(correct)= | 0.800 | |

Automated recognition

| | Signal | |
|---|---|---|
| | Present | Absent |
| "Yes" | 98 | 38 |
| "No" | 2 | 62 |
| | p(Hit) | p(FA) |
| | 0.980 | 0.380 |
| | z(Hit) | z(FA) |
| | 2.054 | -0.305 |
| Sensitivity, d' = | 2.359 | |
| Criterion, C = | -0.874 | |
| P(correct)= | 0.800 | |

We may have views on the relative importance of 'hits' (correct 'yes' responses), 'misses' (saying 'no' when it should be 'yes') and 'false alarms' (incorrect 'yes' responses), and this may vary with the context of our problem. But can we characterize the information value of the two methods independently of these value judgements ? Signal Detection Theory (SDT) offers a framework and method for doing this, and in general for distinguishing between the sensitivity or discriminability (d') of the observer and their response bias or decision criterion (C) in the task.

# Rudiments of signal detection theory (SDT)

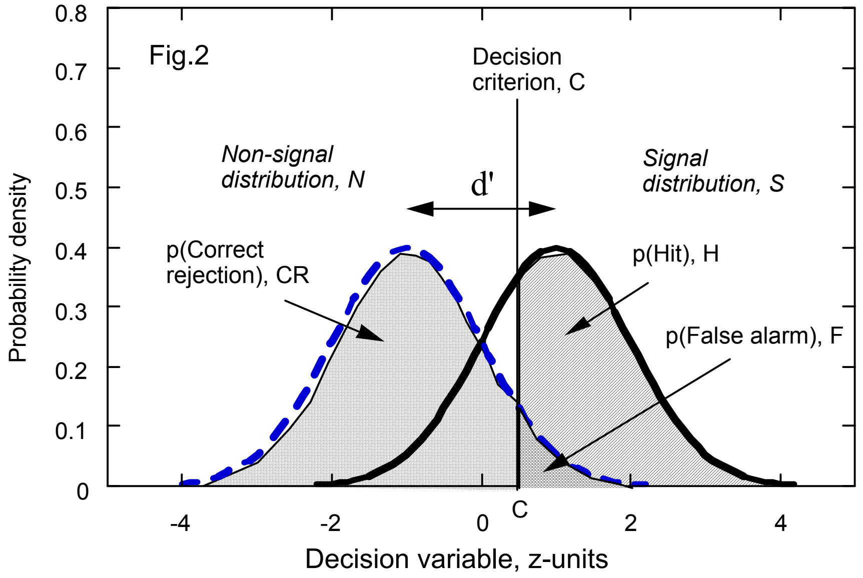


Fig. 2

## 2. Rudiments of Signal Detection Theory

Examples 1 and 2 above illustrate the 'single-interval task' (Fig.1). Only one stimulus 'event' is presented per trial (signal, S, or non-signal, N) and the task is to classify the event as S or N. Hence the data fall into a 2x2 contingency table (Fig. 1). SDT envisages that stimulus events generate internal responses (X) that vary from occasion to occasion. The responses to S and N have different mean values (Fig. 2) and standard SDT supposes that both are normally distributed with the same variance ("the equal variance assumption"). This may not be so, but it's a nice simple model to start with. The variance will depend on both external and internal noise factors.

The variable X is the decision variable that forms the basis for the observer's decision on each trial. The observer has a statistical decision to make: given a response value X, was it more likely to have arisen from the N or S distribution? The reliability of performance on this task will depend on how separate the 2 distributions are. Much overlap => poor discrimination; little overlap => good discrimination. The discriminability (or 'sensitivity') can be quantified by **d'** - defined as the separation between the two means expressed in units of their common standard deviation (z-units).

## 3. Estimating d'

SDT may so far sound rather abstract - but the power of SDT arises when we see how sensitivity d' can be estimated from experimental data on Hit rate and False alarm rate (Fig. 1). First we need to grasp how these response rates (probabilities) are converted into a z-score (Fig.3) and then see how the z-scores are used to give us d' (Fig.4).

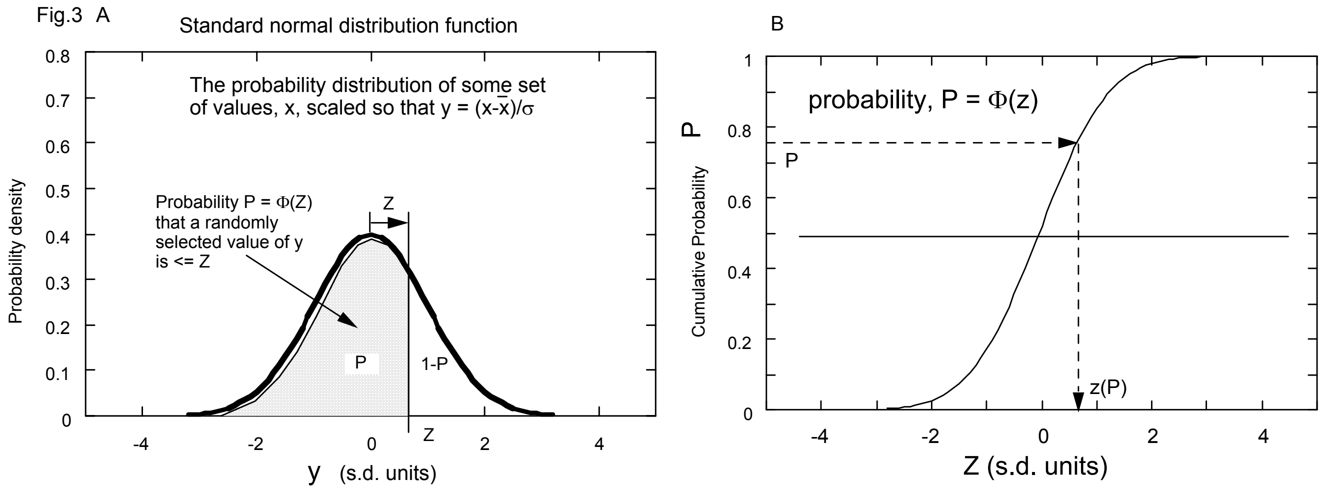


Fig. 3. Note in (B) that z(P) is a simple, but nonlinear transformation of the probability P. Note also from the symmetry of the functions that z(1-P) = -z(P).

Note from fig. 4A that : d' = z(CR) + z(H). Also: CR + F =1, hence CR = 1-F, and so z(CR) = z(1-F). From fig. 3 we have z(1-F) = -z(F), therefore z(CR) = -z(F). Hence: **d' = z(H) - z(F)**

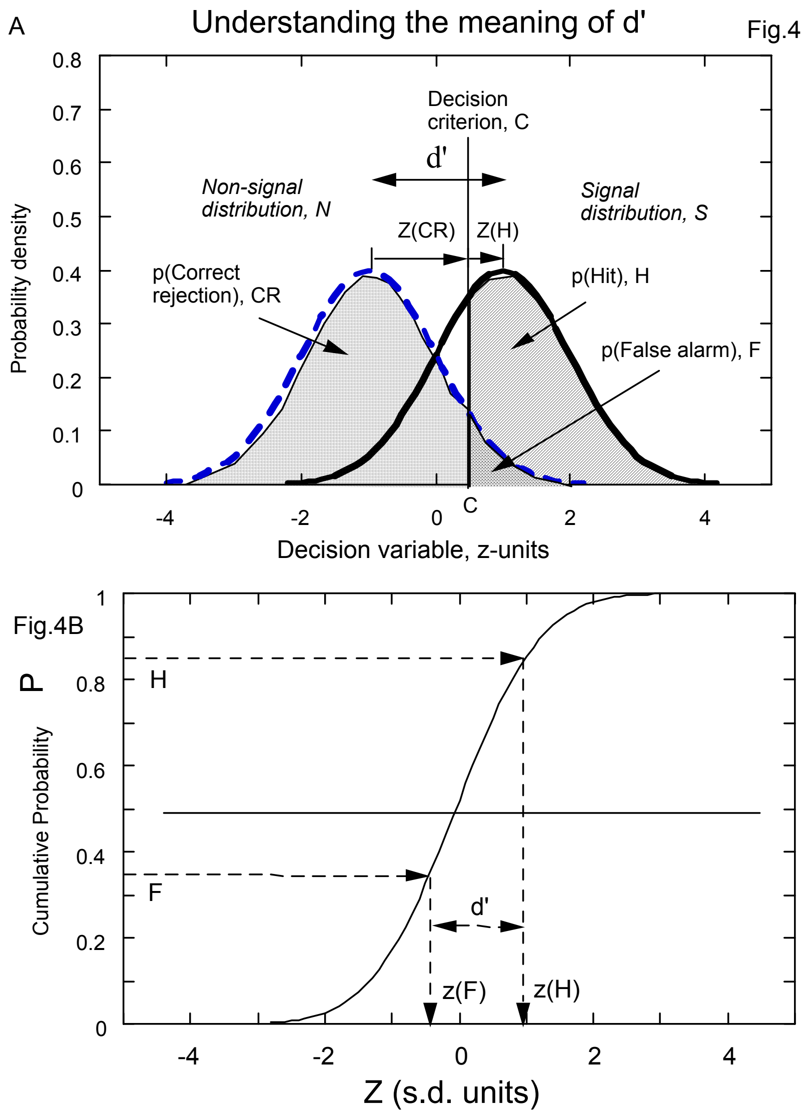


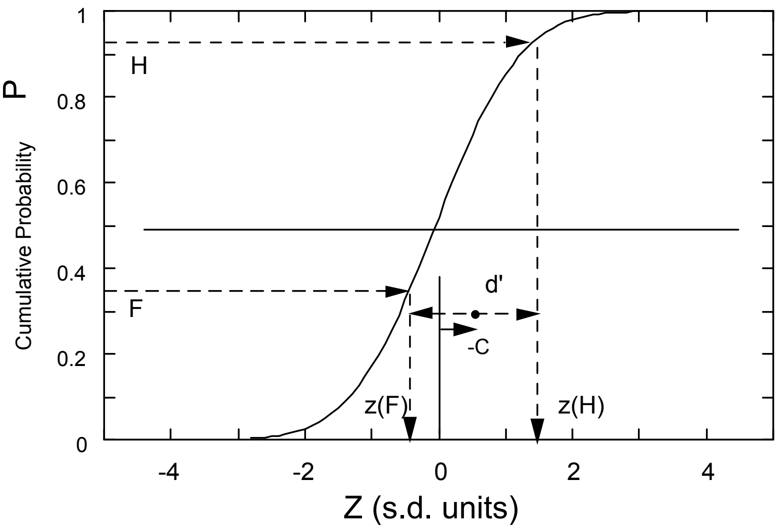
Fig.4

Understanding the meaning of d'

Thus d' is the difference between the z-transformed probabilities of hits and false alarms. It is also the sum of z-transformed probabilities of hits and correct rejections. It is NOT the hit rate, nor z(Hits), nor z(P(c)). All these vary with criterion; d' doesn't. This is so central I'll repeat it: **d' = z(H) - z(F).**

If z(H) increases while z(F) goes down, this means sensitivity (d') is increasing, e.g because stimulus intensity has been increased (or subject has learned to do better on the task).

Fig.5 Understanding the criterion C in SDT



**4. The decision criterion C**

If z(F) and z(H) shift up or down together equally, then their separation (d') clearly stays constant; the common change in z(F) and z(H) reflects a criterion shift, given by the position of the midpoint between z(F) and z(H) (Fig. 5). Thus: **$C = - [z(H) + z(F)]/2$**

An increase in z(H) and z(F) reflects a lower, more relaxed criterion for saying 'Yes'; the midpoint shifts to the right; C <0. If the observer uses a stricter criterion the midpoint shifts to the left; C>0. When C=0 the criterion is midway between the S and N distributions of Fig. 2. Here the observer is said to be 'unbiassed'. Table 1 shows calculations for example 2. The (imaginary) doctors are unbiassed, but my gizmo is biassed in favour of 'yes' responses. Note that P(correct) = 0.8 in both cases, but d' is higher for the machine. How come? SDT implies that if we use P(c) as our measure of sensitivity we will always under-estimate the true sensitivity (d') when bias is present. This can be quite gross if bias is large (Fig. 6).

**5. Discussion - some general points about single-interval data & interpretation**

(i) Hit rate (proportion of correct Yes responses) is a poor guide to psychophysical sensitivity, because it confounds sensitivity (d') and criterion (C). Azzopardi & Cowey (1998) give an interesting, critical discussion of this in relation to the clinical observation of 'blindsight' after damage to visual cortex, and the problem of assessing 'awareness'. Asking "were you aware of it?" is a biassed yes-no task.

(ii) Estimating sensitivity in a single-interval experiment requires the combination of two performance measures - Hits (H; correct yes responses) and False Alarms (F; incorrect yes responses), or equivalently Hits and Correct rejections (Fig. 4A).

(iii) "Percent Correct" (average of H and CR) is not a bad index of sensitivity if bias (C) is not too extreme. In symbols: 2.z[P(c)] = 2.z[(H+CR)/2] = d' (approximately, or exactly if C=0.)

(iv) If the criterion is centrally placed (C=0; no bias) then even the hit rate is OK, because z(H) = -z(F) in this case; hence d' = 2.z(H). But how do we know C=0 if we don't analyze it properly?

(v) Quite often, single-interval experiments are mistakenly thought to be 2AFC. E.g. in my example 1 on each trial a stimulus either moves left or right, and the observer has to report the direction in a 'forced-choice' between left or right. This is not a 2AFC design, because only a single stimulus interval is presented per trial. It is a yes-no experiment in disguise (eg Right = Yes, Left = No). The good news is that if there is little or no directional bias in the responses, as may be the case in a simple detection experiment, then P(c) and d' measure the same thing (see iii).

But some experiments may induce large biases, eg by presenting a visibly moving priming stimulus before the test interval. This was done by Raymond & Isaak (1998, Vision Res. 38, 579-589, "Successive episodes produce direction contrast effects in motion perception"). They found that a same-direction prime greatly increased the 'threshold' for detecting motion coherence in dynamic random-dot displays, while an opposite-direction prime greatly decreased it. But unfortunately their measure of threshold performance was "71% correct", ie. hit rate H of 0.71 in a single interval task, not P(c) = 0.71. The most likely account of the symmetrical increases and decreases of "threshold" produced by the primes is in terms of criterion shift, not sensitivity change. As H and F go up, so 'correct rejections' (CR = 1-F) go down; the two kinds of correct-response rates move in opposite directions as the criterion changes. Thus if separate thresholds are (inappropriately) derived from H and from CR then they will go up and down symmetrically, as found by Raymond & Isaak.

The solution to problems of this kind is to use a genuine 2-interval forced-choice design, and measure P(c) or d', as in Fig. 1B

(vi) Sensory psychophysics has tended to concentrate on d' as the real measure of interest, while treating criterion effects a psychological nuisance. However, SDT makes no prescription that C is psychological rather than sensory. Michel Treisman has done much in recent years to restore interest in C by developing a 'criterion-setting theory' recently applied to spatial frequency discrimination by Lages and Treisman (1998, Vision Res. 38, 557-572; see refs therein).

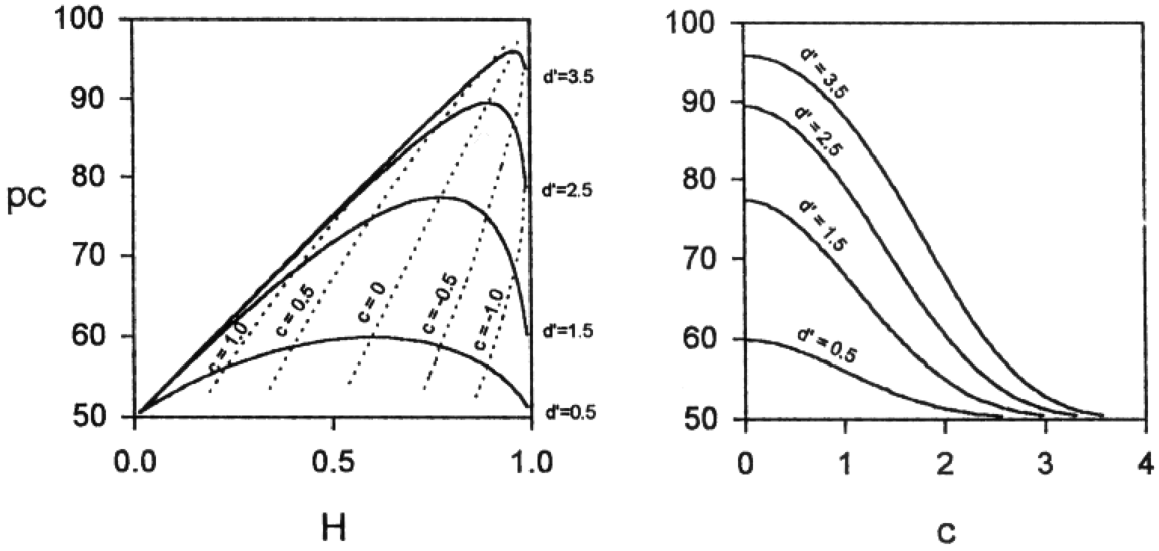


FIG. 2. The relation between percent correct (pc), hit rate (H), and response criterion (c) at fixed sensitivities, calculated from signal detection theory for the case when S+ and S− are presented with equal probability. The latter graph is symmetrical about c = 0. Both graphs indicate that a subject's performance could vary between 50 and 95% correct depending on his response criterion, even though his sensitivity is constant. (Adapted from Azzopardi & Cowey, 1997a.)

Fig. 6 (from Azzopardi & Cowey, 1998)