SOME PROBLEMS OF USING TESTS FOR DIAGNOSTIC CLASSIFICATION

1. Diagnoses as criteria

The criterion used in validation of many diagnostic psychological tests is obviously going to be diagnoses made by psychiatrists or other clinicians.

Tests of depression are validated against diagnoses of depression or its absence; tests of dementia are validated against diagnoses of dementia or its absence; and so on.

How reliable are such diagnoses?

The reliability of diagnosis used to be assessed in an intuitively appealing way. Investigators had different psychiatrists assess the same patients independently and then worked out the percentage of cases on which they agreed.

Figures for agreement on psychiatric diagnosis were typically rather low.

As an example, look at these data from the late 1950s and 1960s.

Investigation	Number of psychiatrists	Number of diagnostic categories	Percentage agreement
Schmidt and Fonda, 1956	2	11	55
Norris, 1959	2	12	58
Kreitman *et al*., 1961	2	11	63
Sandifer *et al*., 1968,	4 6 - 10	12 12	33.5 10

Kreitman, N,. Sainsbury, P., Morrissey, J., Towers, J. and Scrivener, J. (1961) The reliability of psychitric assessment: an anlysis. *Journal of mental Science*, **107**, 887 – 908
Norris, V. (1959) *Mental illmess in London*. Lomdon, Chapman and Hall
Sandifer, M. G., Hordern, A., Timbury, G. C., and Green, L. M. (1968) Psychiatric diagnosis: a comparative study in North Carolina, London and Glasgow. *British Journal of Psychiatry*, **114**, 1 – 9.
Schmidt, H. and Fonda, G. (1956) The reliability of psychiatric diagnosis: a new look. *Journal of abnormal and social Psychology*, **52**, 262 – 267.

At first glance these figures, and others like them, suggested that psychiatric diagnosis was a very poor criterion against which to validate tests. Indeed, if inter-psychiatrist agreement using the commonest categories could sink as low as 10 percent, what was the point of diagnosis?

But how should we interpret these percentage agreement figures?

Let's have a look at two possible models.

The first one we will call the hard-core model. This proposes that there are two categories of cases, a group of people who are indisputably suffering from a given particular psychiatric disorder (the hard core) and those who present such a mixed picture that diagnosis becomes largely a matter of opinion. Every psychiatrist seeing a 'hard core' case will diagnose that case the same way as any other psychiatrist. Other cases will receive different diagnoses from different psychiatrists.

Something like this model was probably in the minds of those who thought that the diagnostic enterprise was a waste of time if inter-clinician agreement could be so low.

So agreement between psychiatrists on diagnosis should equal the percentage of hard-core patients plus chance agreement. From the table above the 'hard core' value can be assumed to be the level of agreement for the 6 to 10 psychiatrists in the Sandifer *et al* study, which is 10 percent.

The formula for agreement will thus be:

$$\text{Percentage agreement} = HC + C.$$

Where

HC = the hard core value

C = chance agreement = (approximately)

$$C = \left(\frac{1}{ndc}\right)^{np-1} \times (1 - HC)$$

where:

ndc = number of diagnostic categories

np = number of psychiatrists

HC = the 'hard core' value

A quite different model is the 'Accurate Psychiatrist' model. This proposes that psychiatrist make (a) accurate and (b) inaccurate diagnoses. Further, the errors (inaccurate diagnoses) made by any given psychiatrist are random and not correlated with errors made by a different psychiatrist.

This model therefore suggests that agreement between psychiatrists will be the product of their accuracies plus chance agreement.

This leads to the formula (again a little oversimplified)

$$\text{Percentage agreement} = A^{np} + C$$

Where

A = the accuracy of a psychiatrist

and

$$C = \left(\frac{1}{ndc}\right)^{np-1} \times \left(1 - A^{np}\right)$$

where (again):

ndc = number of diagnostic categories
np = number of psychiatrists

These are of course only two of the several possible models of agreement.

Let's see how well they fit the actual data.

To make life a little easier we will assume that Sandifer's "6 to 10" psychiatrists were 8 psychiatrists, and that in all studies the number of diagnostic categories was 12.

The results of applying these models to the data reported earlier are shown below.

Number of psychiatrists	Obtained agreement	Agreement predicted by hard core model	Agreement predicted by the Accurate Psychiatrist model
2	59%	17%	60%
4	33%	10%	32%
8	10%	10%	10%
(Table based on: Ley, P. (1972) The reliability of psychiatric diagnosis: some new thoughts. *British Journal of Psychiatry*, **121**, 41 – 43)			

The 'Accurate Psychiatrist' is a clear winner.

So, what was the value assumed for the accuracy of a psychiatrist in preparing this table? It was .75. In other words the data suggest that when a psychiatrist made a diagnosis in those days the diagnosis would be correct 75 percent of the time (see below for how this figure was arrived at.).

This is not a perfect performance, but it is probably a much higher accuracy rate than the crude percentage agreement figure suggests when we first read them.

So percentage agreement can be a misleading guide to the validity of judgements made.

(***Mini optional appendix.***

If we fill in values in the Accurate Psychiatrist equation for 8 psychiatrists seeing the same patients we get:

$$.10 = A^8 + \left(\left(\frac{1}{12}\right)^7 \times \left(1 - A^8\right)\right)$$

As $\left(\frac{1}{12}\right)^7$ equals next to nothing (0.0000000279), the value of

$\left(\frac{1}{12}\right)^7 \times \left(1 - A^8\right)$ can be treated as zero.

Therefore:

$A^8 = .10$, and, thus, A = .75.)

2. Kappa – an improved measure of agreement

Kappa (*k*) is nowadays the most frequently used measure of agreement between diagnosticians.

Suppose we had two clinicians independently seeing the same 80 patients and each independently diagnosing then as schizophrenic or not schizophrenic.

We could summarise the results of their diagnoses in a 2 x 2 table like this.

		Rater B		
		Schizophrenic	Not Schizophrenic	Total
Rater A	Schizophrenic	A = 40	B = 10	Row1 n = 50
	Not schizophrenic	C = 10	D = 20	Row 2 n = 30
	Total	Column 1 n = 50	Column 2 n = 30	N = 80

The cells containing the number of cases on which the clinicians agreed are cells A and D.

The old measure of agreement simply used to add cells A and D, and express the answer as a percentage of N. This was the percentage agreement measure.

$$Percentage\,agreement = 100 \times \frac{(A+D)}{N}$$

So in the present case we have 75% agreement between the diagnosticians.

The main trouble with this measure was that it made no allowance for chance agreement.

We can work out what chance agreement would be expected in the same way as we work out expected values for cells when using χ^2. We simply multiply the row total for the row in which a given cell lies by the column total for the column in which that cell lies, and divide the answer by N.

So how many agreements would we expect by chance in the table above?

If we do this for Cell A we get $\frac{50 \times 50}{80} = 31.25,$

and for cell D we get $\frac{30 \times 30}{80} = 11.25$

Adding the two values together we find that we would expect 42.5 agreements by chance alone. If we divide this number by the total number of cases we obtain the proportion of cases on which the raters would agree by chance, about .53.

Kappa (κ) is a statistic which attempts to allow for chance agreement by working out how much of the non-chance agreement is ascribable to agreement between the clinicians or other raters.

Its verbal formula is:

$$\frac{observed\ agreement - chance\ agreement}{1 - chance\ agreement}$$

where observed and chance agreement are expressed as **proportions** of the total number of cases, i.e.,

$$\kappa = \frac{p_o - p_c}{1 - p_c}$$

where:

p_o = observed agreement

p_c = chance agreement

For our example the value of kappa is:

$$\kappa = \frac{.75 - .531}{1 - .531} = .467$$

This is the proportion of non-chance agreement accounted for by agreement between the diagnosticians.

Kappa is often interpreted according to this scale.

Value of kappa	Description of degree of agreement
0.00	poor
.01 - .20	slight
.21 - .40	fair
.41 - .60	moderate
.61 - .80	substantial
.81 – 1.00	perfect
After: Landis, J. R. and Kock, G. C. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 1089 - 91	

In our example the level of inter-rater agreement would be classed as moderate.

Identical proportions of percentage agreement will not necessarily, of course, produce identical values for kappa.

For example, in the following table, percentage agreement between psychiatrists is 60 percent, but kappa varies in value from .14 to .29.

Example 1		**Psychiatrist B**	
		dementia	**not dementia**
Psychiatrist A	**dementia**	30	20
	not dementia	20	30
		kappa =	**.20**

Example 2		**Psychiatrist B**	
		dementia	**not dementia**
Psychiatrist A	**dementia**	40	0
	not dementia	40	20
		kappa =	**.29**

Example 3		**Psychiatrist B**	
		dementia	**not dementia**
Psychiatrist A	**dementia**	10	35
	not dementia	5	50
		kappa =	**.14**

3. The Base Rate problem

3.1. Some jargon – the nomenclature

Psychological test norms have traditionally been presented in a form that allows you to see what percentage of the abnormal group and what percentage of the control group(s) score above a given cut-off.

In fact test norms have usually centred on the presenting the 'true positive' (the percent of the target group who obtain an abnormal score) and the 'false positive' rates (the percent of those not in the target group who obtain an abnormal score).

For most purposes, and for most base rate calculations, this information about the test is sufficient.

But for some purposes a more complicated terminology is used.

The full standardisation data for a test will often refer to true positives, false positives, true negatives, and false negatives.

The terms 'positive' and 'negative' refer to test results. An abnormal score being called 'positive', and a normal score being called 'negative'.

If the test gives a positive score and the person who gets that score in fact has the abnormal condition, that person is called a '**true positive**' (cell A of the table below).

If the test gives a positive score but the person does not have the abnormal condition, that person is called a '**false positive**' (Cell B below)

Similarly somebody who gets a negative score on the test and who does not have the condition is called a '**true negative**' (Cell D)

Finally, someone who gets a negative score, but who does have the condition, is called a '**false negative'** (Cell C)

		'True Diagnosis'	
		Depression	No Depression
Test Result	Depression	A	B
	Not Depression	C	D

		'True Diagnosis'	
		Depression	No Depression
Test Result	Depression	**True positives**	**False positives**
	Not Depression	**False negatives**	**True negatives**

Further terms include the following.

Sensitivity = True positives/(true positives + false negatives)

$$= A/ (A + C)$$

Thus it is the proportion of those with the condition who get a positive score

Specificity = True negatives/ (true negatives plus false positives)

$$= D/ (B + D)$$

Thus it is the proportion of those without the condition who get a negative score.

Positive predictive value = true positives/(true positives + false positives)

$$= A/ (A + B)$$

It is the proportion of those who get a positive score who have the condition.

Negative predictive value = true negatives / (true negatives + false negatives)

$$= D/ (C + D)$$

It is the proportion of those who get a negative score who do not have the condition.

This alternative terminology is summarized below.

The alternative terminology summarised

		Condition		
		Present	**Absent**	
Test result	**Positive**	*a* True Positive	*b* False Positive	*a* + *b*
	Negative	*c* False Negative	*d* True Negative	*c* + *d*
		a + *c*	*b* + *d*	N

Sensitivity $= a/(a+c)$

Specificity $= d/(b+d)$

Positive predictive value $= a/(a+b)$

Negative predictive value $= d/(c+d)$

Positive likelihood ratio $= \dfrac{a/(a+c)}{b/(b+d)}$

Negative likelihood ratio $= \dfrac{c/(a+c)}{d/(b+d)}$

3.2. The base rate problem.

When you use a test you really want to know the probability that the test result you get is accurate. For example, if the test says that somebody is psychotic, what are the chances that they really are?

Unfortunately the standardisation data for diagnostic tests do not give you this information.

Instead, test standardisation data give you the probability that a person with a given diagnosis will get a given test score.

Let's take a hypothetical test designed to diagnose PTSD.

The validation data for this test are:

		'True Diagnosis'	
		PTSD	**Not PTSD**
Test Result	**PTSD**	90	40
	Not PTSD	10	60

As you can see, the test correctly classifies 90 percent of those with PTSD (a true- positive rate of .90), and misclassifies 40 percent of those without this disorder (a false-positive rate of 0.4).

So we know that the probability of somebody with PTSD getting an abnormal score is .90, but this unfortunately doesn't tell us what we really want to know when we use the test.

What we want to know is the probability that a person has PTSD if they get an abnormal score on our test.

To find this probability we would need to do the following sum:

> (Number of people referred to us and tested as possible cases of PTSD who (a) have PTSD and (b) obtain a PTSD score on the test)
>
> Divided by:

(The total number of people referred to us and tested as possible cases of PTSD who (whether they have PTSD or not) obtain a PTSD score on the test)

The answer will vary depending on what percentage of people referred to us actually do have PTSD.

For example, let's suppose that the percentage of those with PTSD amongst patients referred to us for assessment of this problem is 60 percent.

We know from the standardisation data that 90 percent of these people will get a PTSD score, so as a proportion of our total referral population this will equal 0.9 x 0.6 = 0.54.

And, of course, it follows that if 60 percent of the people referred for this problem have PTSD, then 40 percent do not. So, again consulting the standardisation data, we see that 40 percent of these will get a PTSD score. So in our sample we would expect the proportion of those who do not have PTSD but who obtain a PTSD score to be 0.4 x 0.4 = 0.16.

Therefore, in our sample, we would expect a proportion of (0.54 + 0.16) = .70 of all those tested to get a PTSD score.

Thus, the proportion of those who get a PTSD score who will actually have PTSD will be:

$$\frac{.54}{.70} = .77$$

So, in these circumstances, we can say that the probability of someone obtains a PTSD score actually has PTSD is .77.

But the answer would have been very different if we had used the test as a general screening test for all comers. Suppose that actual cases of PTSD made up only 5 percent of the people referred to us.

The proportion of all referrals who had PTSD and got a PTSD score on the test would be .05 x 0.9 = 0.045.

The proportion of all our referrals who did not have PTSD but who nevertheless got a PTSD score would be 0.95 x 0.4 = 0.38.

So, altogether, the test would say that 42.5 percent of our referrals suffered from PTSD.

Unfortunately, the test would be wrong much more often than it was right. The probability of someone actually having PTSD when the test said they had would only be:

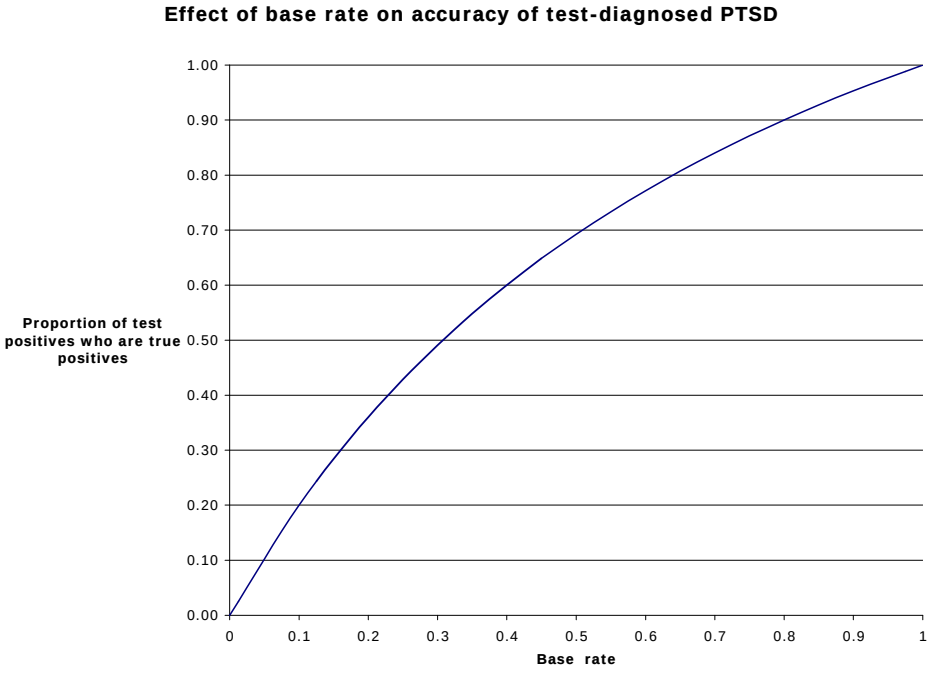
$$\frac{0.045}{0.425} = 0.106$$

The moral of this tale is that the usefulness of a test can vary greatly depending on the proportion of the people we use it on who have the condition we are trying to diagnose.

This proportion is known as the 'base rate' for that condition in that population.

And, the base rate has tremendous impact on the usefulness of a given test.

For our hypothetical test, the following graph shows the probability that a PTSD score indicates the existence of PTSD for different base rates for PTSD,



In case you find tables easier than graphs to deal with, here are the data in tabular form.

Base Rate for PTSD	Probability that someone with a PTSD test score actually has PTSD
.1	.20
.2	.36
.3	.49
.4	.60
.5	.69
.6	.77
.7	.84
.8	.90
.9	.95

In both formats it should be clear that the usefulness of a test depends heavily on the base rate for the condition of interest.

3.3. Bayes Theorem

The formula applied to finding the probability that an abnormal test score indicates the existence of the abnormal condition is Bayes Theorem.

This theorem states:

$$p\text{D/T} = (p\text{D} \times p\text{T/D}) / ((p\text{D} \times p\text{T/D}) + (p\text{N} \times p\text{T/N}))$$

Where:

D is the condition

N is the absence of the condition

T is a test result indicating the presence of the condition

*p*D is the base-rate for the condition

*p*N equals (1 - *p*D)

*p*T/D is the true positive rate

*p*T/N is the false positive rate

To use Bayes Theorem to work out the probability that a test result indicates a given condition you use the formula:

$$pD/T = \frac{BR \times sensitivity}{(BR \times sensitivity) + ((1 - BR) \times (1 - specificity))}$$

where:

pD/T = probability that test positive is true positive

BR = base rate

3.4. Using base rates for several alternative diagnoses

As noted earlier, psychological test norms have traditionally been presented in a form that allows you to see what percentage of the abnormal group and what percentage of the control group(s) score above a given cut-off.

For example, norms for a test of thought disorder might look like this:

Diagnosis	Percent of diagnostic group scoring above cut-off
Schizophrenia	70%
Other psychosis	20%
Neurosis	7%
Normal	5%

The first row gives the **true-positive rate** for schizophrenia, and the other rows give the **false positive rates** for schizophrenia for people in each of the other diagnostic groups.

False negative and true-negative rates can be found by simple subtraction.

Thus, the false negative rate for those with schizophrenia will be
(100 – 70)%; the true negative rate for normal people will be (100 – 5)% and so on.

If we give this test to someone suspected of suffering from schizophrenia, and we know that, amongst people we assess for this problem, 60 percent have schizophrenia; 20 percent suffer from other psychoses; 15 percent have a neurotic disorder; and 5 percent are normal; what is the probability that somebody with a score above the cut-off suffers from schizophrenia?

First we will need to work out the proportions of each group obtaining an abnormal. (We have these in the table above)

Then we will have to multiply these proportions by the by the relevant frequencies with which the various conditions amongst those we test.

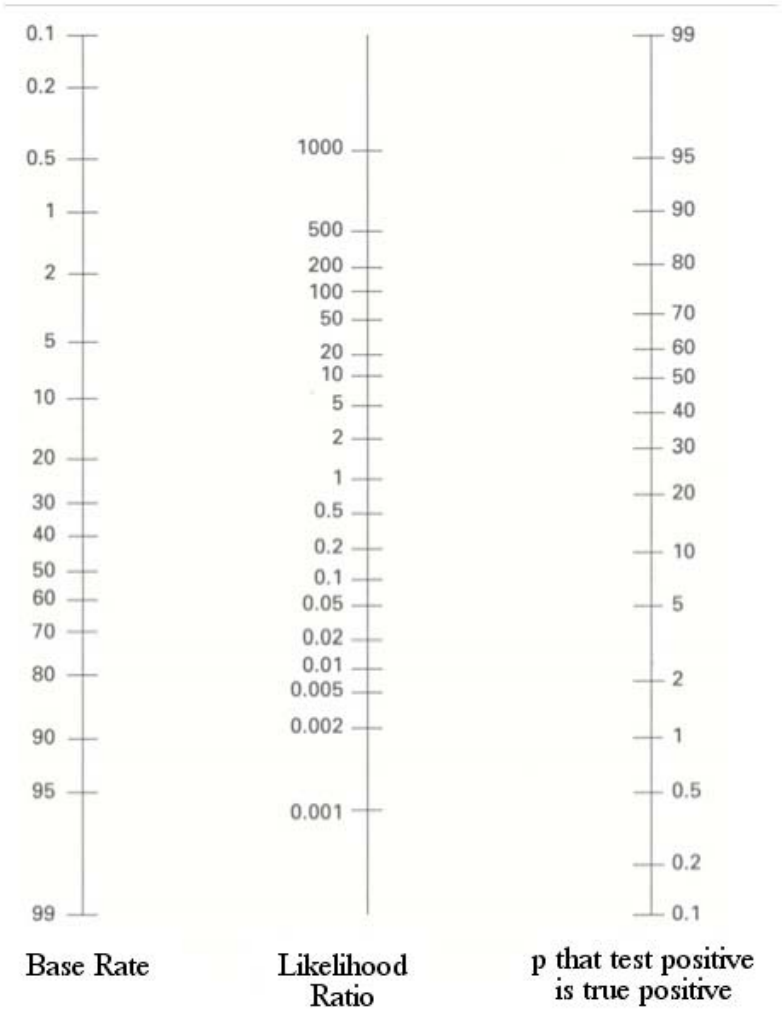
One easy way of doing this is to use a table like this one.

1 Diagnosis	2 Proportion scoring above the cut-off	3 Proportion of people with the diagnosis, shown in Column 1, amongst those tested (the base rate for that diagnosis)	4 Proportion of those tested who (a) have the diagnosis shown in Column 1, and (b) score above the cut-off (Column 2 x Column 3)	5 Probability that score above cut off indicates the diagnosis amongst those tested (Column 4 / Value 6)
Schizophrenia	.70	.60	.42	.89
Other psychosis	.20	.20	.04	.08
Neurosis	.07	.15	.01	.02
Normal	.05	.05	.003	.01
		Value 6 Total proportion	.473	

Thus, with this cut-off and these base rates, given a score above the cut-off, the probability that the person has schizophrenia is .89, the probability that the person has a different psychosis is .08, the probability that the person has a neurosis is .02, and the probability that the person is normal is .01.

3.5. A nomogram for estimating the probability that a positive score indicates a true positive

If you are addicted to sensitivity and specificity as terms of that ilk, you might be interested in this nomogram. It was first suggested by T J Fagan in a letter to the NewEngland Journal of Medicine in 1975, and it has appeared in various forms since then. Note that the base rate is expressed as a percent – not a proportion



To use it, you will need to know the base rate, and the positive likelihood ratio. All you need to do is find the point on the first scale which corresponds to the base rate, and then find the point on the second scale which corresponds to the positive likelihood ratio. Draw a straight line between these two points, and continue it to the third scale. The point where it crosses the third scale gives you the probability that a positive test score is in fact a true positive.

Let's see what happens when we use the formula and the nomogram on the same data. Suppose we have a test for dementia which we are using in a situation where about 60 percent of those referred to us for assessment are in fact demented. This gives a base rate of .6.

Suppose, further , that the test correctly diagnoses 70 percent of those with the condition, and has a false positive rate of 10 percent.

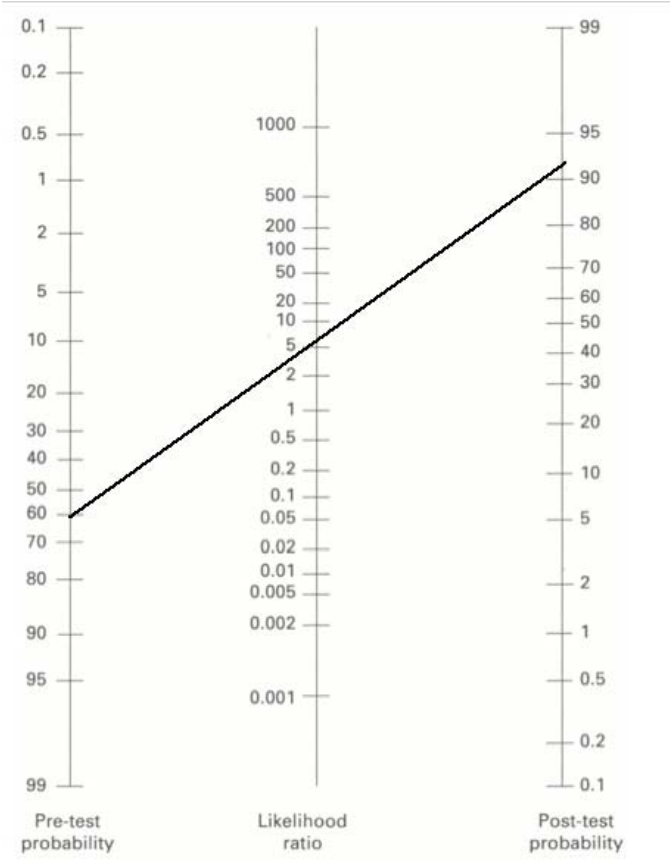
Putting these values as proportions into our formula we get the following:

$$pD/T = \frac{.6 \times .7}{(.6 \times .7) + (.4 \times .1)}$$ which equals .91.

To use the nomogram we need to calculate the positive likelihoo ratio. The positive prediction rate for this test will be :

sensitivity divided by (1 – specificity) = .70/.10 = 7.0

Turning to the nomogram and filling in these values we get the following:



3.6. Why aren't base rates taken into account more often in psychological assessment practice?

As base rates have such a strong effect on the validity of classification it is perhaps surprising that in many psychological assessments they are not taken into account.

A major reason for this is that base-rate data are seldom readily available. This is especially true as base rates will vary from setting to setting, so the relevant base rate cannot usually be obtained from national statistics.

For example while the national prevalence of schizophrenia (about one percent of the population) might provide the relevant base rate for assessing the performance of a screening test for use in a population survey, it would be an unwise clinical psychologist who assumed that only one percent of the cases likely to be seen were suffering from schizophrenia.

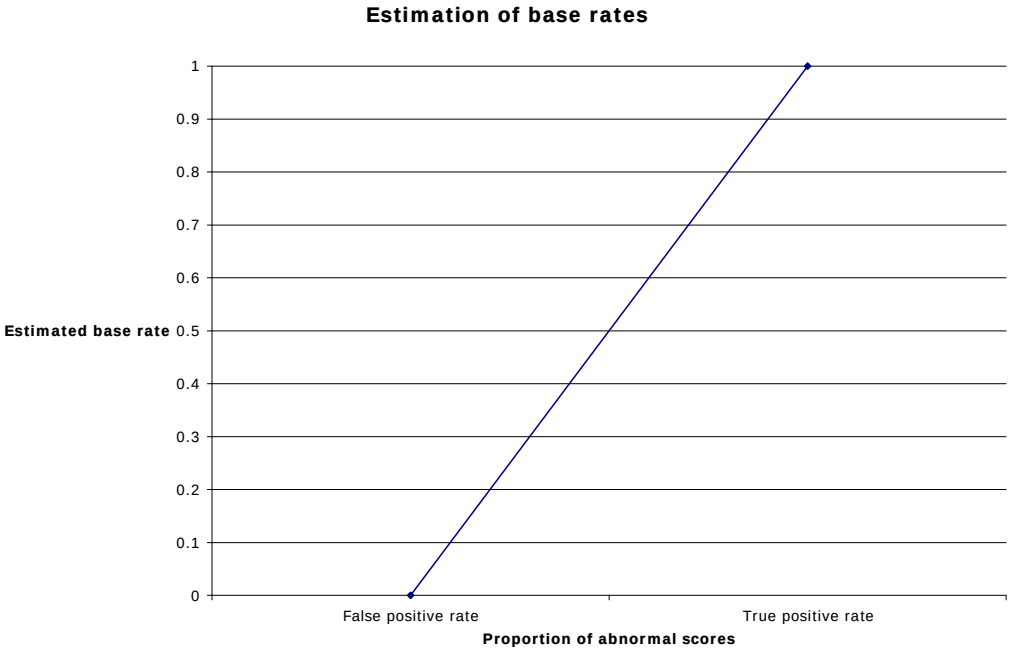
Indeed, for prediction and classification by assessment devices, the base rate required is that pertaining in a particular setting to the clients and patients on whom the assessment is to be made.

So what is needed is a simple method for working out these 'local' base rates.

3.7. A method for estimating base rates

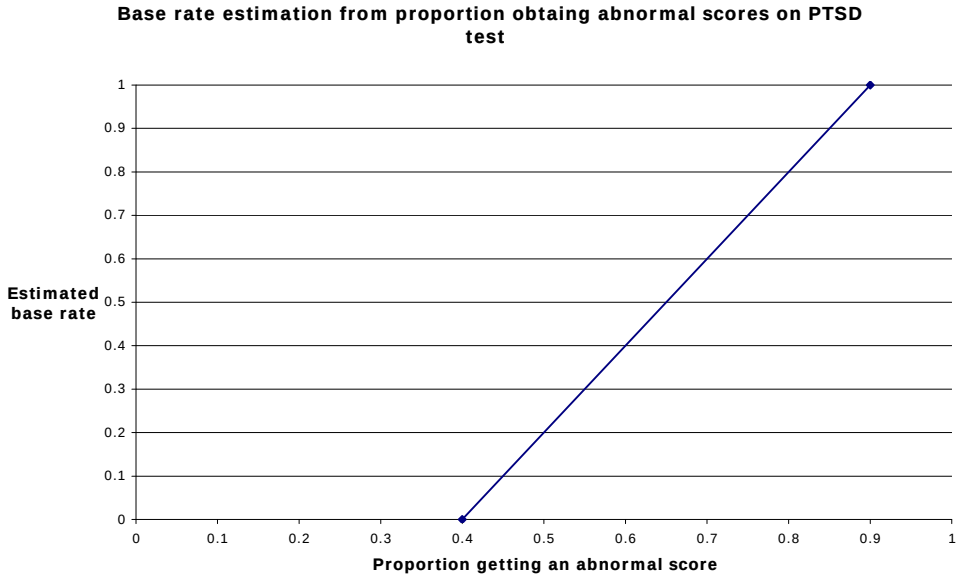
One such method consists of drawing a simple graph. The vertical axis running from 0 to 1 will represent the base-rate (expressed as a proportion), The horizontal axis, ranging from the proportion of false-positives to the proportion of true-positives will represent the proportion of those tested who obtain abnormal scores on the test.

The graph consists of a straight line joining two points, running from a base rate of zero at the false positive rate to base rate of 1.0 at the true positive rate. this graph in its general form is shown here.



The graph thus allows estimation of base rates from the proportion of a tested who obtain an abnormal score. The base-rate obtained by this method can then be used to estimate the probability that a test positive is a true positive.

This graph for our hypothetical PTSD Inventory data is shown in the next figure.



The method requires having a number of test results available from which to calculate the proportion scoring 'abnormally'. The easiest method of obtaining the required data is to use case records to calculate the proportion of all of those to whom a given test has been administered who obtained an abnormal score.

Alternatively, in some circumstances, a sample of potential testees could be given the test. and the base rate derived from the proportion of the sample who obtain an abnormal score. This method would be highly suitable for use with a screening test to be administered to all comers.

Rationale

The upper limit for the proportion of those tested who obtain an abnormal score will be given by the true positive rate. If everybody in the tested population was a PTSD case then the proportion tested who would obtain an abnormal score would equal the true positive rate. Similarly the lower limit for the proportion of those assessed who obtain an abnormal score will equal the false positive rate. - the proportion of those in the standardization population who did not have the abnormality but who nevertheless obtained an abnormal score.

The base rate will of course range between zero and one. It will be zero when the proportion of those tested who obtain an abnormal result equals the false positive rate and 1.0 when the proportion of those tested equals the true positive rate. Further, the base rate will rise in linear fashion as the proportion of those tested who obtain an abnormal score increases above the false positive rate.

Thus the base rate (BR) will rise in accordance with the following formula as the proportion of abnormal scores rises above the false positive rate. In effect the formula simply converts a point of a scale ranging from FP to TP to the equivalent point on a scale ranging from zero to one.

$$BR = (AS - FP) / (TP - FP)$$

Where:

AS = proportion obtaining abnormal score

FP = false positive rate (as proportion)

TP - true positive rate (as proportion)

Obviously it is possible to use the formula above rather than a graph to estimate the base rate, but, arguably, the graphical method is easier to remember, and, once drawn, it can be used in different settings, It also provides more information. For example it highlights the often frequent discrepancy between the percentage of abnormal score and the actual percentage of abnormality in a given population.

More generally, the method can be used to estimate prevalence rates, For example, suppose a psychologist working in a large general practice wished to estimate the prevalence of abnormal depression in pregnant women. All that would be needed would be to administer a test of depression to a sample of the relevant women, calculate the proportion obtaining an abnormal score, and use the outlined method. This prevalence estimate would be superior to simply reporting the percentage of those obtaining an abnormal score, as it will have corrected for the presence of false positives.

The assumptions necessary for the use of the graph, and equivalent formulas, are essentially those involved in using the particular test in the particular situation anyway. The test and its norms should be relevant, and validity should have been demonstrated in validation and cross validation studies. After all, in any given assessment or prediction situation why would any psychologist be using an irrelevant test which was not adequately validated and cross-validated.