J Clin Pathol 1989;**42**:225–230

Occasional article
Statistics on microcomputers: A non-algebraic guide to the appropriate use of statistical packages in biomedical research and pathology laboratory practice

6 Statistical methods for diagnostic tests

R A BROWN, J SWANSON BECK*

*From the Departments of Mathematical Sciences and *Pathology, University of Dundee*

Reference ranges

Most tests used in clinical medicine give a numerical result on a continuous measurement scale. Pathologists or clinicians attempt to interpret the result by comparing it with a "reference range" previously calculated from a study of people who do not have the disease in question. By current convention, the reference range includes all but the top and bottom 2·5% of the results expected from a population of healthy people, so that 5% of the "normal" healthy population will have test values falling outside the reference range. Consequently, the fact that the test result for an individual subject is outside the reference range does not necessarily imply that the individual is abnormal—this is one reason why the older term "normal range" is becoming obsolete.

A reference range may be determined from test values obtained from a sample of healthy subjects provided: (i) the subjects constitute a random sample from the healthy portion of the population; and (ii) the sample size is sufficiently large for it to be representative of the population and for the sample mean and standard deviation to be precise estimates of the population mean and standard deviation.

If the population distribution of the test results is normal, then 95% of all values will lie within the range population mean ± 1·96 (population standard deviation) (Article 2). The sample mean and standard deviation are estimates of the unknown population mean and standard deviation and by convention the *reference range* is taken to be:

sample mean ± 1·96 (sample standard deviation)

The sample mean and standard deviation, however, are subject to sampling variation—that is, each random sample gives rise to different sample statistics—and so the limits of the reference range will be imprecise. Confidence intervals can be calculated for the upper and lower reference limits to indicate the imprecision of the reference range. If the sample size is large a 95% confidence interval for each end of the reference range is:

reference limit ± 1·96 $\sqrt{3}$ (SE mean).

The method for calculation can be illustrated by data

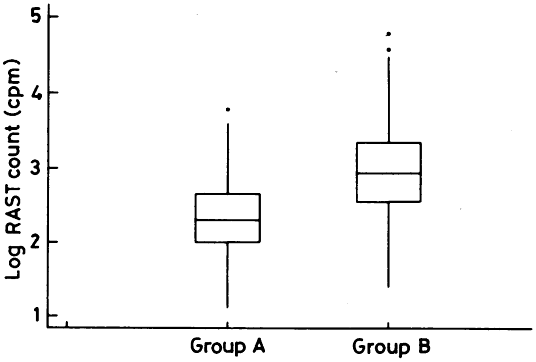

Fig 1 *Box and whisker plots comparing titre of IgE antibodies to* M tuberculosis *using radioallergosorbent test (RAST) in 350 healthy Indonesian factory workers (group A) and 350 Indonesian patients with tuberculosis (group B). Both distributions are symmetrical on a logarithmic scale.*
Group A: mean of log counts = 2·31. SD of log counts = 0·45.
Group B: mean of log counts = 2·96. SD of log counts = 0·58.

Accepted for publication 12 December 1988

obtained in a study of the prevalence and titre of IgE antibodies to *Mycobacterium tuberculosis* in healthy Indonesian factory workers (group A in fig 1). It is clear that the distribution of the counts in the radioallergosorbent test (RAST) is not symmetrical but that it becomes so after logarithmic transformation. These log-transformed results have a reference range of 1·42 to 3·20: the 95% confidence interval calculation indicates that the upper limit for the upper end of the range is 3·28 and the lower limit for the lower end point is 1·34. If these limits are back-transformed (using antilogs) this results in a reference range of 22 to 1905 cpm.

The concept of a reference range is not restricted to medical diagnosis but appears in the guise of a "tolerance interval" in industrial quality control. Comprehensive information on the internationally accepted definition of a statistical tolerance interval is contained in ISO Standard 3207.[1]

The validity of the calculations of the reference range and the confidence intervals for the reference limits depends very critically on the assumption that the test values are normally distributed so the sample must be assessed for normality and possible outliers using methods described in Article 2 before starting the reference range calculation. If the data are not normally distributed or contain inexplicable outliers the sample mean and standard deviation method cannot be used to determine the reference range. If the data are skewed to the right, a square root or a logarithmic transformation (as used with the data of fig 1) may bring about symmetry and perhaps normality thus permitting calculation by the above method.

Usually it will be safer to determine reference limits by a method which does not require normality of the raw test results. A universally applicable method is based on the percentiles of the observed distribution of the sample values. Any given percentile is that value below which a specified percentage of the distribution lies; so 2·5% of the sample values do not exceed the 2·5 percentile and 97·5% of sample values do not exceed the 97·5 percentile (this is analogous to the quartiles introduced in Article 1). On condition that the sample size is sufficiently large, this provides a very simple method of determining reference limits. Statgraphics will automatically calculate any specified percentiles of a set of results, and for the test results of the healthy workers shown in fig 1 the 2·5 and 97·5 percentiles of the untransformed RAST counts are 26 and 1518. This technique will only produce reliable results if the sample size is sufficiently large for the sample to be representative of the variability in the population (greater than 100 to be safe); for small samples the same percentiles will be very imprecise estimates of the corresponding population values. A 95% confidence interval for the 2·5 percentile may be found as follows:

$$0{\cdot}025\ (\text{sample size}) \pm 1{\cdot}96\sqrt{(0{\cdot}025 \times 0{\cdot}975 \times \text{sample size})}$$

this will give two values which will not usually be integers: then round the values up to the next integer.[2] For example, if the sample size is 350 then the above calculation gives the values 3·48 and 14·02 which are rounded up to 4 and 15. This means that the 95% confidence interval for the 2·5 percentile stretches from the fourth smallest up to the fifteenth smallest test value (17 to 32). The corresponding interval for the 97·5 percentile is determined by counting down the same number of values from the largest test value (giving 1180 to 1995). Thus the reference range based on percentiles and allowing for uncertainty in the percentiles is 19 to 1995, which is not very different from that found by using the previous method based on the assumption of normality of log RAST count.

There are several sources of variation which are likely to affect the test value produced by any well monitored and accurately calibrated analytical technique for a given subject: for example, (a) inherent random error in the analysis of the sample, usually monitored by quality control methods; (b) time-related variation; (c) variation associated with physiological changes; (d) variation associated with external factors such as diet, tobacco, or alcohol consumption, exercise, posture and so on before taking the test specimen. Such factors will tend to increase the variability of results and, in the absence of information about the magnitude of their effects on variability, it may be difficult to interpret a result which lies "just outside" the reference range; thus a healthy subject who has a "true" value close to the upper reference limit may give a test result above the limit. Similarly, a diseased subject who has a "true" value above the upper reference limit could be within the reference range when tested.

Assessment of diagnostic tests

In view of the uncertainties inherent in comparing a test result with a reference range it is worth while examining the criteria that should be used in assessing and choosing diagnostic tests.[3,4] To assess a particular diagnostic method results must first be obtained from a group of persons who have the fully developed disease and a group in which the disease is absent. In practice, diagnostic tests are often used to assess patients who are in the early stages of disease; this raises the question of whether comparison of the above groups addresses the correct question. The criteria for determining the presence and absence of the disease in these two groups must be carefully stated and should be in accord with standard well understood

		Disease		
		Present	Absent	Total
Test Result	Positive Negative	a c	b d	a + b c + d
Total		a + c	b + d	a + b + c + d

Fig 2 *For any specified cut off point the results of a single test on a group of patients with the disease and a group of controls without the disease may be classified into a 2 × 2 table.*

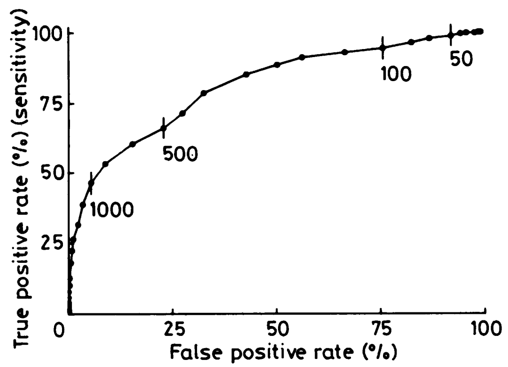

Fig 4 *Receiver-operator characteristic curve (ROC) for a hypothetical diagnostic test based on the RAST counts of fig 1. Selected cutoff points are marked.*

and accepted methodology (a "gold" standard). The criteria for what constitutes a positive test result must be carefully and unambiguously stated. For example, if the test result is determined by comparison of peaks on a chart then the precise way in which the peaks are located and exactly how they are compared should be explained unambiguously.

If the initial studies on a proposed new diagnostic method were based on 350 healthy subjects and 350 patients with pulmonary tuberculosis who had been diagnosed as having the disease on the basis of some independent, well defined, and accepted criterion (in this case chest *x*-ray pictures), the test results might be as shown in fig 1. The investigator might perform a Mann-Whitney U test (Article 3) as the test results are not normally distributed, the resulting p value being less than 0·0001.

The investigator would be wrong to conclude from this that the assay has high diagnostic value because the Mann-Whitney U test is designed to answer the question: "do the data provide evidence of a difference in the medians of the population distributions for healthy and diseased individuals?" This is largely irrelevant for judging the clinical value of a diagnostic test, where the critical decision *must* involve a notion of a cut off point to assist in decision taking. In practice, values above a certain critical limit will be regarded as indicating that the patient has the disease, and lower values that it is unlikely that the patient has the disease. The pertinent question must concern not merely the separation between the medians (or means) of the corresponding population distributions for the healthy and diseased groups but separation of the distributions *as a whole*. It is relatively common for the distributions to overlap, and this raises the possibility that a healthy subject will be classified as diseased or that a diseased subject may be classified as healthy. Assessment of a diagnostic test must concentrate on such issues and look at the risks of misclassification.

As the initial assessment of the value of the proposed new test depends on its application to a group of patients with clearly established disease and to people who are definitely healthy the results for any selected cut off point can be summarised in a 2 × 2 table (fig 2). The *sensitivity* of a test is defined as the proportion of the patients who have the disease in whom the test result is positive (a/(a + c)), and this is a measure of the probability that a person who has the disease will give a positive test result. The *specificity* of a test is defined as the proportion of healthy subjects in whom the test result is negative (d/b + d)), and this measures the probability that a healthy person will give a negative test result. The *false positive rate* is the proportion of the healthy subjects who give a positive test result (b/b + d)), and the *false negative rate* is the proportion of those subjects who have the disease but who give a negative test result (c/(a + c)). Thus in this example with the cut off point set at a RAST count of 320 the test identified correctly 71·4% of those with the disease but failed to detect the disease in 28·6% of cases. It also correctly identified 72·6% of those who

		Disease		
		Present	Absent	Total
Test Result	Positive Negative	250 100	96 254	346 354
Total		350	350	700

Fig 3 *The observed incidence of positive and negative test results based on the data of fig 1 when a positive result is defined as a RAST count exceeding 320 cpm.*

		Disease		
		Present	Absent	Total
Test Result	Positive Negative	714 286	274 726	988 1022
Total		1000	1000	2000

Fig 5 *Expected numbers of subjects in the four categories in a group of 2000 subjects if a positive test result is defined to be a RAST count exceeding 320 cpm, when the prior estimate of the disease probability is 0·5.*

were disease-free but was positive in 27·4% of cases where the disease was absent (fig 3).

Very few diagnostic tests have 100% sensitivity and 100% specificity, and consequently it should be realised that although knowledge of these operating characteristics of a test is very important,this is not in itself sufficient to determine the probability that the disease is present when the test result is positive, or of absence when the test proves negative.

When selecting a test it is necessary to compare the sensitivity and specificity of the available tests. As a general rule the test with the highest sensitivity should be selected for clinical diagnosis unless its specificity is unacceptably low. When a test is to be applied for screening purposes it can be argued that it is important that as large a proportion as possible of those who give a positive test result do, in fact, have the disease the screening program is meant to detect, and so the sensitivity should be high. A test used to confirm a diagnosis should have a high specificity—that is, it should produce a negative result in a high proportion of those who do not have the disease.

The sensitivity and specificity of a test can be changed by altering the criterion for positivity. When the test gives numerical values on a continuous scale a cut off point is used to define the boundary between positive and negative results. For example, in fig 1 a cut off point at 500 (log 500 = 1·699) results in a sensitivity of 66% and a specificity of 77%. By moving the cut off point down the scale a greater proportion of diseased subjects will give a positive tests result so the sensitivity increases, but this occurs at the cost of fewer subjects without the disease giving a negative result—that is, decreasing specificity. The *receiver-operator characteristic (ROC) curve* of a test is a graph which is constructed by plotting for a series of cut off points, the sensitivity (true positive rate) against the false positive rate (100-specificity). This curve is shown in fig 4 for the data displayed in fig 1, and it is apparent that the test can not provide a very high true positive rate coupled with a low false positive rate. The ROC may be used to determine the optimal cut off point for the test according to the costs and benefits (in the widest sense) resulting from correct and incorrect diagnoses. Competing tests may be assessed by comparing their ROC curves.

The probability that a patient with a positive test result has the disease depends not only on the sensitivity and specificity of the test but also on the pretest or prior estimate of the probability that a patient presenting with a particular combination of signs and symptoms has the disease. Note that an investigation designed to assess a test and determine its ROC curve does not provide this information as the assessment will be based on roughly equal numbers of healthy and diseased persons. The prior estimate of disease probability must be based on past experience of the prevalence of the disease among patients with the specified signs and symptoms.

Гable

Prior estimate of disease probability	*Positive predictive value*	*Predictive value of a negative result*
0·1	0·225	0·958
0·3	0·528	0·856
0·5	0·723	0·717
0·7	0·859	0·521
0·9	0·959	0·220

If the prior estimate was, for example, 50%—that is, in 2000 patients presenting one would expect 1000 to have the disease—then the data from the test evaluation may be combined with this information to give the results shown in fig 5. Of the 1000 patients with the disease, 714 will give a positive test result, and 274 of the 1000 patients who do not have the disease will also be positive. Consequently one should expect 988 positive test results and the proportion of diseased patients among these to be 714/988 or 72·3. Thus a positive test result converts a 50% chance that the patient has the disease into a 72% chance.

The *positive predictive value* (PPV) of a positive test result is the probability that a patient who gives a positive test has the disease and the *predictive value of a negative result* (PVN) is the corresponding probability that a patient with a negative result does not have the disease. The table shows the positive and negative predictive value for prior estimates ranging from 10% to 90% when the sensitivity and specificity of the test are 71·4% and 72·6%. In this case the results are not particularly impressive, which is a reflection of the fact that the RAST count is not a good diagnostic indicator in this particular situation.

It must be borne in mind that the results of an assessment of a diagnostic test which compares those without the disease and those with fully developed disease are likely to overstate the predictive value of a positive or a negative result when the test is applied to patients who do not have the fully developed disease as the distribution of test results in such patients is likely to be closer to that of the disease-free group than the distribution of test results from patients with fully developed disease.

Discriminant analysis

There are many clinical situations in which a single diagnostic test may not be sufficiently specific or sensitive to diagnose a disease and it is usual to perform two or more different tests and base the diagnosis on the combined results. The assessment of such situations leads naturally to investigation of the

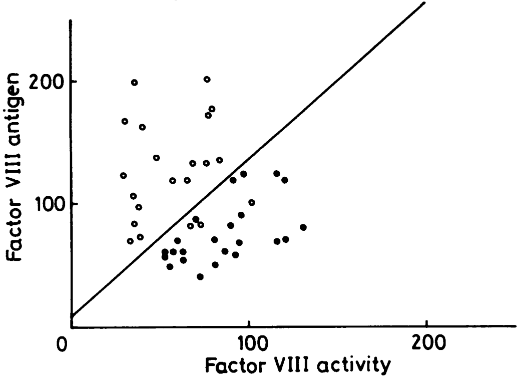

Fig 6 *Scatterplot of results of radioimmunoassay (antigen) and bioassay (activity) of serum factor VIII in 21 confirmed female carriers of haemophilia and 23 normal women. The straight line indicates linear discriminant function.*

sensitivity and specificity of the combination of two or more tests. *Discriminant analysis* is the method generally chosen for determining objectively the cut off region between controls and patients. For example, assays for serum factor VIII based on radioimmunoassay and bioassay were compared for their diagnostic value in detecting women who were carriers of haemophilia.[5] The results for each test taken separately showed substantial overlap between women who were carriers and controls. When, however, the results were displayed in a scatterplot it was clear that the two groups could be separated almost completely by an imaginary diagonal line (fig 6). On this basis it would be possible to set out the rule: "classify the subject as a carrier if the results of the pair of tests lie above the line". For this cut off line the sensitivity is 86% and the specificity is 100%, so the combination of the two tests is an immense improvement over the use of either test alone.

Discriminant analysis[6 7] is the name given to the statistical technique for determining objectively the position of the line which best discriminates between the two groups, in the sense of producing as few misclassifications as possible in the data from the original investigation. The line separating the two groups defines the *linear discriminant function* which is merely a formula for combining both test results into a single *discriminant score* which may then be used to assess sensitivity and specificity in the usual way. The Statgraphics output for the data of fig 6 is shown in fig 7; the "unstandardised discriminant function coefficients" are interpreted as meaning that a negative test result is one for which the discriminant score, 3·47 (factor VIII antigen) −2·72 (factor VIII activity) exceeds −23·5. The discriminant scores for both groups are plotted in fig 8 and show almost complete separation of the groups.

```
Discriminant Analysis for Factor VIII

Unstandardised Discriminant Function Coefficients
                          1
activity                  0·0347
antigen                   0·0272
CONSTANT                  0·2350
```

Fig 7 *Statgraphics output for linear discriminant analysis of the data of fig 6.*

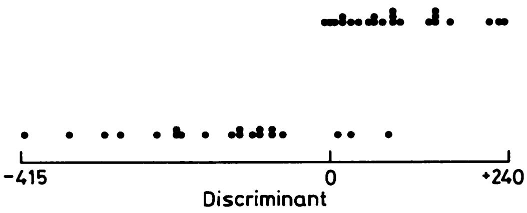

Fig 8 *Plot of the discriminant scores for the data of fig 6.*

The sensitivity and specificity of the pair of tests may be examined by calculating the misclassification rates for other cut off points than that resulting from the discriminant analysis. This is done by replacing the value −23·5 by a range of alternatives. The data manipulation facilities of Statgraphics can be used to calculate the discriminant score 3·47 (factor VIII antigen −2·72 (factor VII activity) for each person in both the control and the carrier group, and the tabulation facilities can be used to calculate how many persons in either group have discriminant score less than each alternative cut off value. From this the sensitivity and specificity at each alternative value can be determined which enables the ROC curve to be plotted.

Discriminant analysis is not limited to pairs of measurements; it may be applied to situations in which several different measurements have been made on each of a number of subjects in different groups with the intention of finding linear discriminant functions which allow newly observed subjects to be allocated to one of the groups. It should be noted, however, that the method will be reliable only if the subjects used in the original investigation were allocated to the groups by some well understood and reliable method.

The method of discriminant analysis used by Statgraphics and similar packages can not cope successfully with data which consist of a mixture of measurements (continuous variables) and categorical or ordinal data. Advice from your local statistical guru should be sought if your data include both types of variables.

Reprise

This series of articles has presented several standard techniques of statistical analysis and has sought to emphasise the importance of using statistical analysis to answer the *right* questions. It will be apparent from the discussion of the misuse of correlation and regression in comparison of laboratory methods and in the discussion on the evaluation of diagnostic tests in this article that it is sometimes easy to lose sight of the questions central to the scientist's investigations in the search for a convenient statistical technique. The investigator should always remember that it is stupid to use the wrong technique to answer an irrelevant question. This situation can be avoided by asking a number of questions at the beginning of the investigation: (a) what is the precise aim of the investigation? (b) if a very large amount of a particular type of data was available, would it provide the answer to the questions being asked? (c) is the investigation concerned with differences or similarities between groups? (d) is the investigation large enough to detect clinically important differences?

Finally, one should attempt to clear one's mind of the idea that all that is important is the calculation of some result (any result) which has a p value small enough to allow the result to be declared "significant". Many statistical text books with a heavy emphasis on significance testing do not help in this respect. At the risk of boring the reader we will reiterate our view that significance should always be related to clinical or biological significance, and a full measure of common sense (with clinical or scientific insight) should be exercised in planning an investigation and interpreting its results.

We thank Dr J H Gibbs and Mr R W Fawkes for the preparation of the diagrams and Professor CD Forbes of the department of medicine, Ninewells Medical School, for the generous gifts of the data on which fig 6 is based. Statgraphics is a proprietary statistical analysis package produced by the Statistical Graphics Corporation. Minitab is a proprietary statistical analysis package produced by Minitab Inc.

References

1 International Standards Organisation. *Statistical tolerance intervals, ISO standards 3207*. Geneva: ISO, 1980.

2 Conover WJ. *Practical nonparametric statistics*. New York: John Wiley and Sons, 1980:111–6.

3 Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures: principles and applications. *Ann Int Med* 1981;**94**:553–600.

4 Anonymous. When is a test diagnostic? [Editorial.] *Hum Pathol* 1985;**16**:325.

5 Prentice CRM, Forbes CD, Morrice S, McLaren AD. Calculation of predictive odds for possible carriers of haemophilia. *Thromb Diath Haem* 1975;**34**:740–7.

6 Fisher RA. The use of multiple measurements in taxonomomic problems. *Ann Eugenics* 1936;**7**:179–88.

7 Morrison DF. *Multivariate statistical methods*. Tokyo: Kogakusha. McGraw-Hill, 1976:230–45.

Requests for reprints to: Professor J Swanson Beck, Department of Pathology, Ninewells Hospital and Medical School, Dundee, DD1 9SY Scotland.