14

Some common problems in medical research

> Omniscient as statisticians are, their ability to diagnose abnormality is not generally acknowledged by the medical community, and indeed they usually refrain from claiming it.
>
> Oldham (1979)

> A picture may be worth a thousand t tests.
>
> Cooper and Zangwill (1989)

14.1 INTRODUCTION

The methods of analysis described in Chapters 9 to 13 cover a high proportion of the methods used in medical research. None is specific to medical data, although survival analysis is much more common in medical research than in other fields. There are some types of medical investigation, however, that are not covered by these methods. Epidemiological studies in particular require many statistical techniques that are not used much in other fields. There are many books devoted to epidemiological methods.

This chapter covers a small miscellany of common medical problems that need a special approach – method comparison studies, observer agreement studies, diagnostic tests and the calculation of reference ranges. These methods have in common the absence of any complicated mathematics. Their difficulties lie in requiring a clear understanding of the aim of the analysis, and in the interpretation of the results. Also considered is the analysis of data that comprise a series of measurements on each subject, for which a simple approach is also recommended. Lastly, there is a brief introduction to the investigation of cyclic variation.

14.2 METHOD COMPARISON STUDIES

Most clinical measurements are not precise. Either it is not possible to measure directly the quantity of interest, such as heart volume or tumour

size, or the measurement, although direct, is difficult to make, such as arm circumference. Further, the variable may change with time, such as peak expiratory flow rate or blood pressure.

Because of these uncertainties there is usually a variety of techniques available and studies comparing two (or more) methods are common. The aim of these studies is usually to see if the methods 'agree' well enough for one method to replace the other, or perhaps for the two methods to be used interchangeably. For example, we may wish to see if a new cheap and/or quick method gives results that agree with those of an existing expensive, slow method. The same considerations apply to studies comparing two observers using one method. Note that we need to define what we mean by agreement. Also, we are concerned with the degree of agreement, so that this problem is one of *estimation* rather than hypothesis testing.

Put simply, the best approach to this type of data is to analyse the differences between the measurements by the two methods on each subject. A fuller discussion of method comparison studies is given by Bland and Altman (1986).

14.2.1 Analysis

Table 14.1 shows measurements of transmitral volumetric flow (MF) by Doppler echocardiography and left ventricular stroke volume (SV) by cross-sectional echocardiography in 21 patients without aortic valve disease. The researchers expected these measurements to be the same in such patients, but to differ in patients with aortic regurgitation. They thus first wished to see how well MF and SV agreed in patients without aortic valve disease. Figure 14.1 shows a scatter diagram of the data. If the methods agreed exactly the points would all lie on the line of equality, but of course real data never agree exactly. We can see, however, that all these data points are quite near to the line of equality. An alternative, more informative plot is shown in Figure 14.2. Here the differences between the methods (SV–MF) have been plotted against the average of the two measurements. There are several advantages of this plot. We can see the size of differences much more easily and also their distribution around zero, and we can check visually that the differences are not related to the *size* of the measurement. For this purpose the average acts as our best estimate of the unknown true value. Section 14.2.2 describes what we do when the scatter of the differences gets wider as the mean increases. Figure 14.2 shows no such problem, so we can investigate the differences further. We can construct a histogram, and can calculate the mean and standard deviation, which are $-0.24\ \text{cm}^3$ and $6.96\ \text{cm}^3$. We could use a one sample t test of the differences against zero (or, equivalently, a paired t test on the original data) to see if the mean difference is significantly different

Table 14.1 Transmitral volumetric flow (MF) and left ventricular stroke volume (SV) in 21 patients without aortic valve disease (Zhang *et al.*, 1986). Data (in cm^3) in order of MF values

Patient	MF	SV
1	47	43
2	66	70
3	68	72
4	69	81
5	70	60
6	70	67
7	73	72
8	75	72
9	79	92
10	81	76
11	85	85
12	87	82
13	87	90
14	87	96
15	90	82
16	100	100
17	104	94
18	105	98
19	112	108
20	120	131
21	132	131
Mean	86.0	85.8
SD	20.3	21.2

from zero, but it is more important to quantify the variability of the individual data points.

The question being asked relates to how well the methods agree, and there are two components to the answer. Firstly, the mean difference is an estimate of the average *bias* of one method relative to the other. Here the mean is negligible and we can say that the methods agree excellently *on average*. Secondly, it is essential to consider also how well the methods are likely to agree *for an individual*, for which purpose we use the standard deviation of the differences. Although we could simply quote the standard deviation of the differences (s_{diff}) as a measure of agreement (or disagreement), it is more useful to use the standard deviation to construct a range of values which we expect to cover the agreement between the methods for most subjects.

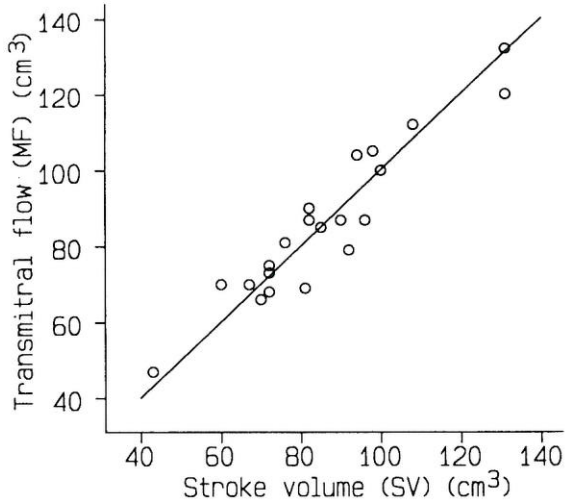

Figure 14.1 Transmitral volumetric flow (MF) and left ventricular stroke volume (SV). Data from Zhang *et al.* (1986).

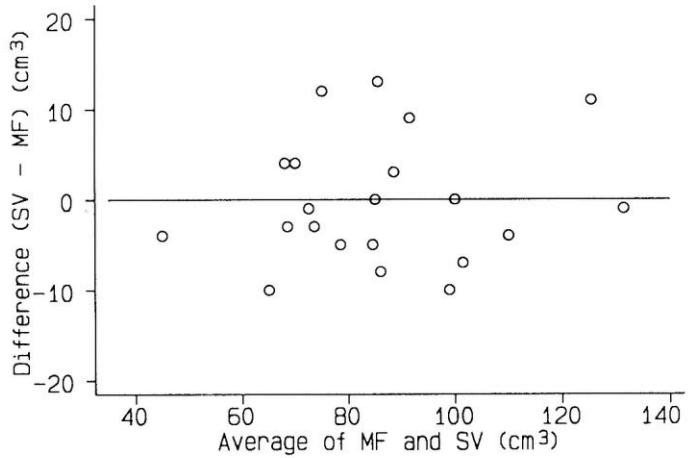

Figure 14.2 Difference between transmitral volumetric flow and left ventricular stroke volume (SV–MF) plotted against average, (MF + SV)/2.

We saw in section 3.4 that for reasonably symmetric distributions we expect the range mean ±2SD to include about 95% of the observations. For a method comparison study we can therefore take mean $\pm 2s_{diff}$ as a 95% range of agreement for individuals. This range of values defines the 95% **limits of agreement**. For the present data we get a range from

$$-0.24 - 2 \times 6.96 \quad \text{to} \quad -0.24 + 2 \times 6.96$$

which is -14.2 to $+13.7$ cm^3. In other words, for a new subject we expect the two methods to give measurements that differ by less than 14 cm^3, with any discrepancy being equally likely in either direction.

The researchers also compared MF and SV in 25 patients with aortic valve disease. Figure 14.3 compares the differences between the methods for patients with or without disease. For only two of the 25 patients with aortic valve disease was SV–MF within the 95% limits of agreement for patients without disease, supporting the researchers' expectations.

The interpretation of the mean and standard deviation of the differences must depend upon the clinical circumstances – it is not possible to use statistics to define acceptable agreement.

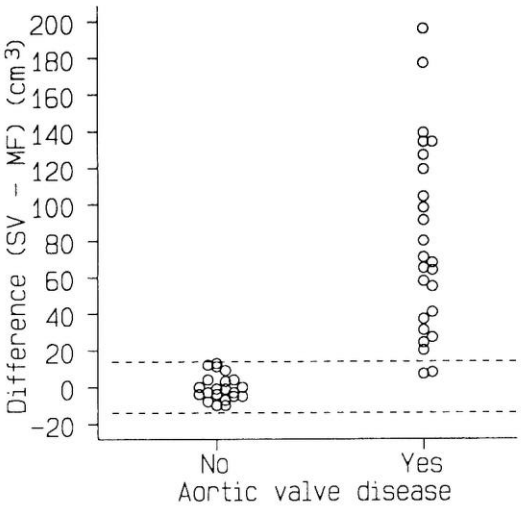

Figure 14.3 Differences between SV and MF for patients with or without aortic valve disease, showing 95% limits of agreement for patients without disease.

14.2.2 Variable agreement (relation between difference and mean)

Sometimes a plot of the differences between two methods against the average shows that there is a wider scatter as the average increases. In other words, the standard deviation of the differences increases. Although the approach given in the previous section may not be unreasonable, a better analysis is often obtained by taking logs of the data before calculating the limits of agreement. Here we are implicitly considering the differences between methods to be an approximately constant *proportion*

of the size of the measurement. As with other uses of the log transformation described in previous chapters, we perform the usual analysis on the logs of the data and then back-transform the results. Antilogs of the limits of agreement thus give us a range of *proportional* agreement between the methods. For example, we may conclude that for a new subject method A will be likely to give a value between 80% and 130% of that obtained by Method B. Bland and Altman (1986) discuss this type of analysis, and give a worked example.

14.2.3 Repeatability

An important aspect of method comparison is the comparison of the repeatability of each method. If we have two (or more) measurements of the same subjects by each method then we can assess the similarity of the duplicate measurements made using the same technique. For paired observations we simply calculate the standard deviation of the differences between the pairs of measurements using the same method. We can then compare the standard deviations to see which method is more repeatable. Each standard deviation can also be used to calculate limits within which we expect the differences between two measurements by the same method to lie. Bland and Altman (1986) give a worked example.

Replicate measurements are rarely made in method comparison studies, so that an important aspect of comparability is often overlooked. A method with poor repeatability will never agree well with another method.

14.2.4 Erroneous analyses

Method comparison studies are frequently mis-analysed. In particular, the correlation between the values by the two methods is often calculated, with a high value of r interpreted as an indication of good agreement. There are several reasons why correlation is an inappropriate analysis. Firstly, the correlation coefficient is a measure of the strength of *linear association* between two variables, which is not the same as a measure of *agreement*. As we have seen, agreement should be assessed in terms directly related to the measurements. It is not possible to interpret, say, $r = 0.92$ in the same way as the limits of agreement. Secondly, we may have a high degree of correlation when the agreement is clinically poor. For example, in a study of the variability of knee circumference measurements Kirwan *et al.* (1979) found that the repeatability of measurements made 15 cm above the patella by two observers was far too poor for the measurement to be clinically valuable. Nevertheless, there was a correlation of 0.99 between the observers' readings. A high value of r can be obtained because, as in their study, there is large variation between subjects. It is clearly not reasonable to assess agreement by a statistical method that is highly sensitive to the

choice of the sample of subjects. Similar criticisms can be levelled at the use of regression analysis for assessing agreement.

Another common incorrect analysis is the comparison of means by a hypothesis test, often a paired t test. We cannot deduce that methods agree well because they are not significantly different. Indeed a high scatter of differences may well lead to an important difference in means (bias) being non-significant. Using this approach worse agreement decreases the chance of finding a significant difference and so increases the chance that the methods will appear to agree!

14.2.5 Presentation

Comparing methods of measurement is very simple and informative using the approach of section 14.2.1. The mean difference and limits of agreement give an excellent summary of the data. It is useful to have one or two plots as well, especially one showing the difference against the mean, on which the other values can be superimposed as three horizontal lines. A plot of the raw data, such as in Figure 14.1, should be square and should show the line of equality.

14.2.6 Discussion

We should remember the limitations of this type of analysis. We cannot tell which method is nearer to the 'truth' because we do not usually know the true values. Nor for unreplicated studies can we compare the repeatability of different methods of measurement. It is important to realize that if one method is either inaccurate or has poor repeatability (or both) comparison with any other method will inevitably show poor agreement, however good the second method is. Thus we should not infer from poor agreement that *both* methods are poor. In contrast, good agreement is most unlikely unless we have two methods that are both accurate and repeatable.

Care should be taken with the design of method comparison studies. The sample size should be large enough to allow the limits of agreement to be estimated well. We can calculate confidence intervals for the limits of agreement, and these will be wide in small samples. Thus a sample size of at least 50, but preferably rather larger, is desirable for a method comparison study. It is definitely valuable to take two measurements on each subject by each method, so that the repeatability of the two methods can be compared. The analysis can then be based on the average of the two replicates, but a correction must then be made to the standard deviation of the differences to allow for this fact (Bland and Altman, 1986). It is most undesirable for the two techniques being compared to be carried out by different observers. Any systematic variation between

observers (a common phenomenon) will be inseparable from any difference between methods. This may be necessary, however, when the techniques involve considerable skill and experience.

As indicated by the knee circumference example, we can use the same statistical approach for studies of observer comparability. We cannot, though, use this method when comparing assessments in categories as opposed to measurements. Section 14.3 considers such problems, which usually arise in observer comparisons rather than method comparisons.

14.3 INTER-RATER AGREEMENT

Agreement between categorical assessments is usually considered as a problem of comparing the ability of different raters (observers) to classify subjects into one of several groups. The approach outlined below does, however, also apply to studies that compare two alternative categorization schemes, that is, a method comparison study for categorical data. I shall consider an example of each.

Table 14.2 shows the classification by two radiologists of 85 xeromammograms as 'Normal', 'Benign disease', 'Suspicion of cancer' or 'Cancer'. The data come from a larger study of nine radiologists (Boyd *et al.*, 1982). As with the comparison of continuous data discussed in the previous section, we require some measure of agreement rather than association. Thus we do not use the χ^2 test, both because we do not wish to assess association and also because this is not a hypothesis testing problem. (Further, the data are paired).

Table 14.2 Assessments of 85 xeromammograms by two radiologists (Boyd *et al.*, 1982)

	Radiologist B				
Radiologist A	Normal	Benign	Suspected cancer	Cancer	Total
Normal	21	12	0	0	33
Benign	4	17	1	0	22
Suspected cancer	3	9	15	2	29
Cancer	0	0	0	1	1
Total	28	38	16	3	85

14.3.1 Measuring agreement

The simplest approach to assessing agreement is simply to see how many exact agreements were observed, which here is $21 + 17 + 15 + 1 = 54$.

There is thus agreement for 54/85 = 0.64 (64%) of the films. There are two weaknesses of this simple calculation. Firstly, it takes no account of where in the table the agreement was, and secondly, we would expect some agreement between the radiologists by chance even if they were guessing. We can get a more reasonable answer by considering the agreement in excess of the amount of agreement that we would expect by chance.

We saw in section 10.3 that the expected frequency in a cell of a frequency table (under the null hypothesis of no association) is the product of the total of the relevant column and the total of the relevant row divided by the grand total. Thus the expected frequencies along the diagonal in Table 14.2 are

Normal	$33 \times 28/85 = 10.87$
Benign disease	$22 \times 38/85 = 9.84$
Suspected cancer	$29 \times 16/85 = 5.46$
Cancer	$1 \times 3/85 = 0.04$
Total	26.20

So the number of agreements expected just by chance is 26.2, which as a proportion of the total is 26.2/85 = 0.31. The question, therefore, is how much better were the radiologists than 0.31. The maximum agreement is 1.00, so we can express the radiologists' agreement as a proportion of the possible scope for doing better than chance, which is 1.00 − 0.31. We thus calculate the agreement as

$$\frac{0.64 - 0.31}{1.00 - 0.31} = 0.47.$$

The name for this measure of agreement is **kappa**, written κ. It has a maximum of 1.00 when agreement is perfect, a value of zero indicates no agreement better than chance, and negative values show worse than chance agreement, which is unlikely in this context.

How do we interpret values between 0 and 1, such as 0.47? While no absolute definitions are possible the following guidelines (slightly adapted from Landis and Koch, 1977) should help:

Value of κ	Strength of agreement
< 0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good

We can thus say that there was moderate agreement between the radiologists. It is of some interest that these two observers showed the best agreement of any pair of observers in the study.

The reduction of the data to a single number inevitably yields an answer that is not terribly meaningful without examination of the table of frequencies. In practice, any value of κ much below 0.5 will indicate poor agreement, although the degree of acceptable agreement must depend upon circumstances. There is no substitute for inspecting the table of frequencies, because many different tables will yield similar values of κ.

An example of the comparison of alternative methods of categorical assessment is given by the data in Table 14.3. The aim of the study was to compare a radioallergosorbent (RAST) test and a multi-RAST (MAST) test on sera for specific IgE as a test of allergy in subjects for whom prick tests cannot be used. The MAST was a new, simpler and cheaper method.

As Table 14.3 shows, there was considerable disagreement between the methods, with some samples in nearly all the cells of the table. The value of κ for Table 14.3 is 0.32, confirming the visual impression.

Section 14.3.4 shows the mathematical expression for calculating κ.

Table 14.3 Comparison of RAST and MAST methods of testing serum for allergies (Brostoff *et al.*, 1984)

	RAST					
MAST	Negative 1	Weak 2	Moderate 3	High 4	Very high 5	Total
Negative (1)	86	3	14	0	2	105
Weak (2)	26	0	10	4	0	40
Moderate (3)	20	2	22	4	1	49
High (4)	11	1	37	16	14	79
Very high (5)	3	0	15	24	48	90
Total	146	6	98	48	65	363

14.3.2 Confidence interval

We can obtain a standard error for κ, and thus a confidence interval. In general this is not all that useful because unless the sample is small the confidence interval will be narrow and thus will not allow for much variation in interpretation. For the radiologists' assessments we had $\kappa = 0.47$ and can calculate $se(\kappa) = 0.07$, so that a 95% confidence interval for κ is given by 0.33 to 0.61. For the rather larger MAST/RAST study κ was 0.32 with a 95% confidence interval from 0.26 to 0.38. The method of calculation is given in section 14.3.4.

14.3.3 Weighted kappa

A weakness of the kappa statistic is that it takes no account of the degree of disagreement – all disagreements are treated equally. Where the categories are ordered, as is often the case, it may be preferable to give different **weights** to disagreements according to the magnitude of the discrepancy. Here observations near to the diagonal, representing a difference of only one category, are considered less serious than those where the discrepancy is two or three categories.

We can build this idea into the calculation of κ to get a quantity called **weighted kappa**. For the MAST-RAST study weighted kappa is $\kappa_w = 0.56$, somewhat better than the unweighted $\kappa = 0.32$. Similarly, weighted kappa for the radiologists' assessments is $\kappa_w = 0.57$ compared with unweighted $\kappa = 0.47$. Weighted kappa is usually higher than unweighted kappa because disagreements are more likely to be by only one category than by several categories.

14.3.4 Mathematics for kappa

(This section can be omitted without loss of continuity.)

Kappa is calculated from the observed and expected frequencies on the diagonal of a square table of frequencies. If there are n observations in g categories, then the observed proportional agreement is

$$p_o = \sum_{i=1}^{g} f_{ii}/n$$

where f_{ii} is the number of agreements for category i. The expected proportion of agreements by chance is given by

$$p_e = \sum_{i=1}^{g} r_i c_i / n^2$$

where r_i and c_i are the row and column totals for the ith category. The index of agreement, kappa, is given by

$$\kappa = \frac{p_o - p_e}{1 - p_e}.$$

The approximate standard error of κ is

$$se(\kappa) = \sqrt{\frac{p_o(1 - p_o)}{n(1 - p_e)^2}}$$

so that a 95% confidence interval for the population value of κ is given by

$$\kappa - 1.96\, se(\kappa) \quad \text{to} \quad \kappa + 1.96\, se(\kappa).$$

Weighted kappa is obtained by giving weights to the frequencies in each cell of the table according to their distance from the diagonal that indicates agreement. For the cell in row i and column j, with observed frequency f_{ij}, a weight is calculated as

$$w_{ij} = 1 - \frac{|i - j|}{g - 1}.$$

Thus we give cells on the diagonal a weight of 1, while those where the difference is by one category get a weight of $1 - 1/(g - 1)$. For the MAST-RAST data weights for discrepancies of 0, 1, 2, 3 and 4 are thus 1, 0.75, 0.5, 0.25 and 0 respectively.

The weighted observed and expected proportional agreement are obtained as

$$p_{o(w)} = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{g} w_{ij} f_{ij}$$

and

$$p_{e(w)} = \frac{1}{n^2} \sum_{i=1}^{g} \sum_{j=1}^{g} w_{ij} r_i c_j$$

and weighted kappa is given by

$$\kappa_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}.$$

Fleiss (1981, p. 223) shows how to calculate the standard error of weighted kappa.

14.3.5 Discussion

As with other methods of looking at small, square frequency tables, there are difficulties associated with the use and interpretation of kappa. The most often cited problem is that the value of kappa depends upon the proportion of subjects (prevalence) in each category. This can be seen most clearly using a simple artificial example, where we have only two categories. Table 14.4 shows two tables with the same proportional agreement of 0.8, but with different proportions in the two categories (+ and −) and with markedly different values of κ. The reason for the difference is that the chance expected frequencies are very different, as shown in Table 14.5. The consequence of this property of κ is that it is misleading to compare values of κ from different studies where the prevalences of the categories differ. For larger tables the same is true, but it is even more complicated to judge comparability.

Table 14.4 Comparison of two observers' diagnoses with different prevalences in the two categories

(a)

		Observer 1 +	Observer 1 −	Total
Observer 2	+	70	10	80
	−	10	10	20
	Total	80	20	100

$\kappa = 0.38$

(b)

		Observer 1 +	Observer 1 −	Total
Observer 2	+	40	10	50
	−	10	40	50
	Total	50	50	100

$\kappa = 0.60$

Table 14.5 Expected frequencies corresponding to the data in Table 14.4

(a)

		Observer 1 +	Observer 1 −	Total
Observer 2	+	64	16	80
	−	16	4	20
	Total	80	20	100

(b)

		Observer 1 +	Observer 1 −	Total
Observer 2	+	25	25	50
	−	25	25	50
	Total	50	50	100

Another problem is that κ depends on the *number* of categories. The data in Table 14.3 can be grouped into three rather than five categories; 0, 1 or 2, 3 or 4. For the resulting 3×3 table we find $\kappa = 0.42$, compared with $\kappa = 0.32$ for the full 5×5 table. If we consider that the methods are really only going to be used to categorize samples as negative (0) or positive (1, 2, 3 or 4) we can collapse the data into a 2×2 table, for which $\kappa = 0.53$, not wonderful but better than $\kappa = 0.32$.

Despite these shortcomings, the use of kappa is becoming common for data like the examples discussed. It is undoubtedly the right type of approach. Incorrect analyses of such data are still common, however. The MAST-RAST data were analysed by calculating the correlation coefficient (Brostoff *et al.*, 1984). The authors concluded from the value of $r = 0.72$ that the methods gave similar results and recommended the use of the simpler and cheaper MAST methods. Not only is Pearson's correlation coefficient unsuitable for ordinal data but, as we saw in section 14.2.4, it is an inappropriate approach to judge agreement. Nor is their conclusion compatible with the data shown in Table 14.3. Similarly, it would be incorrect to judge agreement by a χ^2 test, which is also a test of association. The kappa statistic, which may be interpreted as the *chance-corrected proportional agreement*, is the best approach to this type of problem, but it is important to show the raw data if at all possible. Acceptable agreement depends upon the circumstances. There is no value of kappa that can be regarded universally as indicating good agreement – statistics cannot provide a simple substitute for clinical judgement.

14.4 DIAGNOSTIC TESTS

Diagnosis is an essential part of clinical practice, and much medical research is carried out to try to improve methods of diagnosis. The statistical analysis of these studies is fairly simple, but causes difficulty because of unfamiliar and confusing terminology.

The simplest case to consider is that where patients can be classified into two groups according to the results of an investigation, perhaps an X-ray or biopsy, or the presence or absence of a symptom or sign. An example is given in Table 14.6, which shows the relation between the results of liver scans and diagnosis based on either autopsy, biopsy or surgical inspection. The question of interest here is how good is the liver scan at diagnosis of abnormal pathology. While we could simply calculate the agreement between the two classifications using the methods described in section 14.3, this problem is different because of the asymmetry of the relation between the two classifications. We wish to describe the ability of the scan to diagnose the true patient status. In practice we rarely know the truth, and so evaluate the test in relation to the diagnosis. This distinction is considered further in section 14.4.7.

Table 14.6 Relation between results of liver scan and diagnosis in 344 patients (Drum and Christacapoulos, 1972)

	Pathology		
Liver scan	Abnormal (+)	Normal (−)	Total
Abnormal (+)	231	32	263
Normal (−)	27	54	81
Total	258	86	344

14.4.1 Sensitivity and specificity

One approach is to calculate the proportions of patients with normal and abnormal liver scans who are likewise 'diagnosed' by the scan. The terms **positive** and **negative** refer to the presence or absence of the condition of interest, here abnormal pathology. Thus there are 258 positives and 86 negatives. The proportions of these two groups that have correct diagnoses based on the scan are thus 231/258 = 0.90 and 54/86 = 0.63 respectively. These two proportions have confusingly similar names which are formally defined as follows:

Sensitivity is the proportion of positives that are correctly identified by the test;

Specificity is the proportion of negatives that are correctly identified by the test.

We can thus say that, based on the sample studied, we would expect 90% of patients with abnormal pathology to have abnormal (positive) liver scans, while 63% of those with normal pathology would have normal (negative) liver scans.

At first sight these simple calculations appear to have answered the question posed, but there is more to these problems than meets the eye. We have answered the question from one direction only. In clinical practice the test result is all that is known, so we want to know how good the test is at predicting abnormality. In other words, what proportion of patients with abnormal test results are truly abnormal?

14.4.2 Positive and negative predictive values

The whole point of a diagnostic test is to use it to make a diagnosis, so we need to know what the probability is of the test giving the correct diagnosis, whether it is positive or negative. The sensitivity and specificity do not give us this information. Instead we must approach the data from

the direction of the test results. Of the 263 patients with abnormal liver scans 231 had abnormal pathology, giving the proportion of correct diagnoses as 231/263 = 0.88. Similarly, among the 81 patients with normal liver scans the proportion of correct diagnoses was 54/81 = 0.67. These two proportions are given more sensible names, which are formally defined as follows:

Positive predictive value is the proportion of patients with positive test results who are correctly diagnosed;

Negative predictive value is the proportion of patients with negative test results who are correctly diagnosed.

The positive and negative predictive values give a direct assessment of the usefulness of the test in practice. Unfortunately, we still cannot stop the analysis because there is another essential aspect of the analysis to consider, which is invisible in the above calculations, and that is the **prevalence of abnormality**.

14.4.3 The effect of prevalence

The disadvantage of the sensitivity and specificity is that they do not assess the accuracy of the test in a clinically useful way. They do have the advantage, however, that they are not affected by the proportion of subjects with the abnormality, which we call the **prevalence**. It is assumed here that we know the patients' true status. See section 14.4.7 for further comment on this point.

The predictive values, in contrast, are clinically useful but depend very strongly on the prevalence. In the liver scan study the prevalence of abnormality was very high, being 258/344 = 0.75; that is, exactly three-quarters. In different clinical settings the prevalence of abnormality will vary greatly. Using the data in Table 14.6 I constructed Table 14.7 to show the results we would expect in a group of patients where the prevalence of abnormality is 0.25. Table 14.8 shows the analyses of the data for these

Table 14.7 Predicted effect on liver scan results of a prevalence of abnormality of 0.25, based on data in Table 14.6

	Pathology		
Liver scan	Abnormal (+)	Normal (−)	Total
Abnormal (+)	77	96	173
Normal (−)	9	162	171
Total	86	258	344

Table 14.8 Analysis of liver scan data with prevalences of abnormality of 0.75 and 0.25

	Prevalence	
	0.75	0.25
Sensitivity	0.90	0.90
Specificity	0.63	0.63
Positive predictive value	0.88	0.45
Negative predictive value	0.67	0.95
Total correct predictions	0.83	0.69

two prevalences. As noted, the sensitivity and specificity are unchanged: these calculations are made on the columns of the table, and are not affected by the proportion of patients in each column. In contrast the predictive values of the test are based on the rows, and have changed a lot because they are affected by the prevalence of abnormality. The contrast between the data in Tables 14.6 and 14.7 is illustrated in Figure 14.4.

The effect of a lower prevalence is much as we would expect: the more uncommon is true abnormality the more sure we can be that a negative test indicates no abnormality, and the less sure that a positive result really

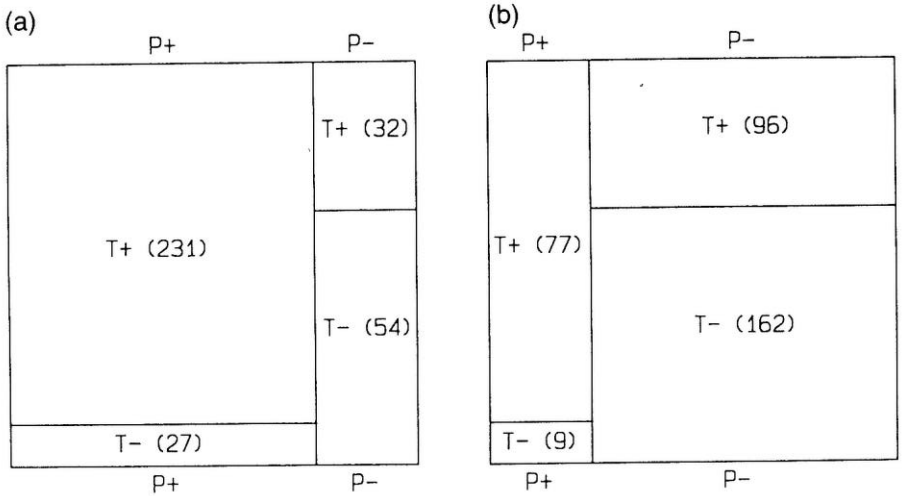

Figure 14.4 Graphical illustration of (a) Table 14.6 and (b) Table 14.7. P indicates the pathology and T indicates the test. The sensitivity is depicted by the proportion of the area P+ that is labelled T+, and is the same in both figures. Likewise the specificity is the proportion of the area P− that is labelled T−, and this is the same in both figures. Conversely, the PPV is the proportion of the area labelled T+ that is P+, and is markedly different for the two figures. The same applies to the NPV.

indicates an abnormal patient. The predictive values of a test thus depend upon the prevalence of the abnormality in the patients being tested, which may not be known. **We should not take the predictive values observed in the sample as applying universally.**

14.4.4 Diagnosis based on a continuous measurement

So far I have considered the case where we wish to determine the presence or absence of some abnormality on the basis of the presence or absence of some symptom or test result. Another common situation arises when the diagnosis is to be made using a continuous measurement. I exclude here conditions such as hypertension, anaemia and perhaps obesity, which are *defined* by the value of a continuous measurement. We may have a single measurement or a score derived from combining two or more different measurements. Here the distinction between discriminant analysis based on logistic regression (section 12.5.2) and the methodology of diagnostic tests becomes decidedly blurred, as does that between diagnosis and prognosis.

Table 14.9 shows results of an HTLV-III (now HIV) antibody assay among patients with AIDS and healthy blood donors. If we wish to use the test to diagnose HIV seropositivity then we need to choose an appropriate cut-off. For each possible cut-off we can calculate the sensitivity and specificity of the test, and we can also calculate the positive and negative predictive values for any prevalence of seropositivity. The method for this last calculation is given in section 14.4.5.

Table 14.10 shows these calculations for the HTLV-III antibody assay results. Predictive values have been calculated assuming the prevalence of AIDS to be either 10% or 1% to illustrate the effect of the prevalence on

Table 14.9 Results of enzyme-linked immunosorbent assay (ELISA) for HTLV-III among patients with AIDS and healthy blood donors (Weiss *et al.*, 1985). (Results expressed as the ratio of the mean absorbance of a pair of test samples divided by the mean absorbance of eight negative control wells)

Ratio	Healthy blood donors	Patients with AIDS
< 2.0	202 (68%)	0 (0%)
2.0–2.99	73 (25%)	2 (2%)
3.0–3.99	15 (5%)	7 (8%)
4.0–4.99	3 (1%)	7 (8%)
5.0–5.99	2 (1%)	15 (17%)
6.0–11.99	2 (1%)	36 (41%)
12.0 +	0 (0%)	21 (24%)
Total	297 (100%)	88 (100%)

Table 14.10 Calculations of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for data in Table 14.9

Cut-off for ratio	Sensitivity	Specificity	Prevalence of HIV seropositivity 10% PPV	10% NPV	1% PPV	1% NPV
2.0	1.00	0.68	0.26	1.00	0.03	1.00
3.0	0.98	0.93	0.59	0.997	0.12	0.9997
4.0	0.90	0.98	0.81	0.99	0.28	0.999
5.0	0.82	0.99	0.87	0.98	0.38	0.998
6.0	0.65	0.99	0.91	0.96	0.49	0.996
12.0	0.24	1.00	1.00	0.92	1.00	0.992

predictive values. There is no reason to use the prevalence in the study data (23%) which has no meaning because the two samples of subjects were selected independently. The appropriate figure to use will depend upon the characteristics of the population being studied.

The choice of a cut-off is not a statistical decision. Assuming that it is felt that the values in Table 14.10 show that the test is *clinically* useful, then the 'best' cut-off must be chosen according to the relative costs (not necessarily financial) associated with a false positive and false negative test results. This in turn will be related to the clinical action that will follow a positive test, in particular whether the test is a *screening* test or a *diagnostic* test (see section 14.4.7). It is not always necessary, however, to impose a cut-off, as we will see below. The need to do so depends on whether the aim is to make a diagnosis or a prognosis. Again, this is not a statistical issue.

We can arrive at a similar situation with the results of a multiple regression analysis. As we saw in section 12.4.8 a regression model can be used to derive a continuous score or prognostic index. When the outcome variable is binary and logistic regression is used, that prognostic index can be converted into a probability of the presence (or absence) of that outcome. In section 12.5.2 I described the application of logistic regression to the problem of discrimination. It is a small jump to the use of the same model for diagnosis; indeed, the two concepts are arguably the same.

In the next section the calculations are examined more closely.

14.4.5 Calculations

Table 14.11 shows a general representation of any diagnostic test based on a binary indicator, such as the presence or absence of a particular symptom

Table 14.11 General representation of a diagnostic test

		Disease status Positive	Negative	Total
Test	Positive	a	b	$a + b$
	Negative	c	d	$c + d$
	Total	$a + c$	$b + d$	n

or test result. We can give names to the four cells:

Test	Disease status	Name
+	+	True positive (a)
+	−	False positive (b)
−	+	False negative (c)
−	−	True negative (d)

The quantities defined and discussed earlier are

$$\text{Sensitivity} = a/(a + c)$$
$$\text{Specificity} = d/(b + d)$$
$$\text{Positive predictive value} = a/(a + b)$$
$$\text{Negative predictive value} = d/(c + d)$$

The terms **false positive rate** and **false negative rate** are sometimes used, but these names are ambiguous. For example, the false negative rate might be $c/(c + d)$ or $c/(a + c)$, depending on your point of view.

The observed **prevalence** of disease in the study is $(a + c)/n$. If the study is carried out on a definable group of patients, such as those attending a particular clinic, then the prevalence may be useful, as may the calculation of positive and negative predictive values based on that prevalence. More generally, however, we may wish to consider the predictive ability of the test for groups with other prevalences of disease, such as different age groups or even the general population. These calculations depend upon **Bayes' theorem**, which is that

$$\text{Prob(disease|test positive)} = \frac{\text{Prob(test positive|disease)} \times \text{Prob(disease)}}{\text{Prob(test positive)}}$$

$$= \frac{\text{Prob(test positive|disease)} \times \text{Prob(disease)}}{\text{Prob(test positive|disease)} \times \text{Prob(disease)} + \text{Prob(test positive|no disease)} \times \text{Prob(no disease)}}$$

where Prob(disease|test positive) means the probability of disease when the

test is positive, and so on. From the earlier definitions it is clear that

Prob(disease) = prevalence of disease

Prob(disease|test positive) = positive predictive value (PPV)

Prob(test positive|disease) = sensitivity

Prob(test positive| no disease) = 1 − specificity

so that we can rewrite the above equation for the probability of disease when the test is positive as

$$\text{PPV} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}.$$

By a similar argument we can show that the negative predictive value (NPV) is

$$\text{NPV} = \frac{\text{specificity} \times (1 - \text{prevalence})}{(1 - \text{sensitivity}) \times \text{prevalence} + \text{specificity} \times (1 - \text{prevalence})}.$$

Two consequences of these formulae are clear. Firstly, it is simple to estimate the predictive values for any prevalence of disease. The effect of varying the prevalence can be marked, as is seen in Table 14.10. Secondly, if we have no idea of the prevalence we cannot estimate the predictive value of the test. Another way of interpreting the prevalence is as the probability before the test is carried out that the subject has the disease, known as the **prior probability** of disease. The values of PPV and 1 − NPV are the revised estimates of the same probability for those subjects who are positive and negative to the test, and are known as **posterior probabilities**. The difference between the prior and posterior probabilities is one way of assessing the usefulness of the test.

We can extend these ideas to diagnosis based on a continuous measurement, by considering each possible cut-off in turn. Table 14.10 illustrated the procedure for the association between assay results and HIV seropositivity.

The sensitivity and specificity are proportions, and so we can calculate confidence intervals for them using the methods of section 10.2.1. When two diagnostic tests are compared on the same sample of individuals, the sensitivities and specificities are paired and so the appropriate confidence interval (section 10.4.1) and the McNemar test (section 10.7.5) should be used.

14.4.6 Two further ways of looking at diagnostic tests

(This section can be omitted without loss of continuity.)

The apparent simplicity of diagnostic test data, particularly when presented as a 2 by 2 table, is belied by the many ways of expressing the results.

Here I consider two further approaches that are more informative than simply looking at sensitivity and specificity.

(a) The likelihood ratio

For any test result we can compare the probability of getting that result if the patient truly had the condition of interest with the corresponding probability if they were healthy. The ratio of these probabilities is called the **likelihood ratio** (LR), and it is calculated as

$$\text{LR} = \frac{\text{Prob(positive test|disease)}}{\text{Prob(positive test|no disease)}} = \frac{\text{sensitivity}}{1 - \text{specificity}}.$$

We can consider the likelihood ratio as indicating the value of the test for increasing certainty about a positive diagnosis. The prevalence is the probability of disease before the test is performed. The *odds* of having the disease are thus given as prevalence/(1 − prevalence). Thus if the prevalence is 10%, the odds are 0.11, or 9 to 1 against the disease being present. We can call this figure the **pre-test odds**, and the odds corresponding to the positive predictive value as the **post-test odds**. It is not difficult mathematically to show that

$$\text{Post-test odds} = \text{pre-test odds} \times \text{likelihood ratio}$$

demonstrating how the likelihood ratio measures the change in certainty of diagnosis.

For the data in Table 14.6 the prevalence of abnormal pathology is 0.75, so the pre-test odds of disease are $0.75/(1 - 0.75) = 3.0$. The post-test odds of disease given a positive test are $0.878/(1 - 0.878) = 7.22$, and the likelihood ratio is $0.895/(1 - 0.628) = 2.406$, demonstrating the stated relation between these three quantities ($7.22 = 3.0 \times 2.406$). For the data in Table 14.7 the likelihood ratio is the same, but the pre-test odds of disease are $0.25/(1 - 0.25) = 0.33$. We can obtain the post-test odds as $2.406 \times 0.33 = 0.79$.

This approach may give further insight into the interpretation of diagnostic test data, but it does not add new information because the same quantities are used as before. As I have just shown, a high likelihood ratio may demonstrate that the test is useful but it does not necessarily indicate that a positive test is a good indicator of the presence of disease. For the data in Table 14.7, the low prevalence of 0.25 means that someone with a positive test is still more likely to be normal than abnormal – this is seen from both the post-test odds of 0.81 and the PPV of 0.45. Using odds rather than probabilities may be helpful, however, especially for seeing the usefulness of the test as assessed by the likelihood ratio (Ingelfinger *et al.*, 1987, p. 25).

(b) ROC curve

When a measurement is used to make a diagnosis the choice of the 'best'

cut-off is not simple. A graphical approach is to plot the sensitivity versus 1 – specificity for each possible cut-off, and to join the points. The curve thus obtained is known as a 'receiver operating characteristic' curve or **ROC curve**, because the method originated in studies of signal detection by radar operators. For the data in Table 14.10 the curve would thus be based on the second and third columns. However, the ROC curve is not very helpful for these data because the specificities are so high that the 'curve' follows the y axis. If the 'cost' of a false negative result is the same as that of a false positive result, the best cut-off is that which maximizes the sum of the sensitivity and specificity, which is the point nearest the top left-hand corner. With different costs it is hard to note the best point from the graph.

The ROC method is perhaps most useful when comparing two or more competing methods. For a single test it does not add anything to a table but it is preferable when there are many possible cut-off values. Of course, the ROC curve, being based only on sensitivity and specificity, takes no account of the prevalence of the disease being tested for.

14.4.7 What is the patient's true condition?

In section 14.4.3 I observed that the sensitivity and specificity calculated from a sample of subjects are unrelated to the prevalence of abnormality. This may not always be the case. We can consider three ways of categorizing a patient – their true condition, the diagnosis, and the test results. When we calculate the sensitivity and specificity of the test we do this in relation to the diagnosis, but we do not necessarily know that the diagnosis is always correct. Unless the diagnosis is perfect, so that it always gives the patient's true status (positive or negative), we are evaluating the test's ability to predict the diagnosis rather than the patient's true disease status. In this case, the sensitivity and specificity of the test in relation to the true state *are* related to the prevalence of abnormality (Begg, 1987). This suggests that unless it is known that the diagnosis is almost always correct, it is wise to evaluate a diagnostic test on patients with the same prevalence of disease as those for whom the test will be used in future.

14.4.8 Discussion

The analysis of data from diagnostic tests requires no complicated mathematics. The main difficulty is not statistical, but rather the need to decide how good the test should be to be clinically valuable. The answer to this question is related to the prevalence of the disease in the subjects being tested. Two extremes are when we are testing high risk individuals, perhaps in a tertiary referral centre, and when we are screening an ostensibly healthy population for early signs of rare serious disease, such as

cervical cancer. For screening tests it is very important to have high specificity and NPV. We do not want false negative results and are willing to accept a moderate number of false positive results. All those positive to the screening test will then be tested again, usually with a different test. Here the requirement will be a high sensitivity and PPV, because a positive result will probably lead to a diagnosis of disease and clinical intervention. A high specificity is also desirable, of course. The detection of HIV seropositivity is a good example of the case where the importance of a false positive diagnosis would have major consequences for the patient and so would a false negative diagnosis for someone receiving their blood in a transfusion. Another is the use of alpha-fetoprotein levels from amniocentesis to detect fetuses with Down's syndrome. These issues must be carefully weighed up when deciding where to put the cut-off between positive and negative diagnosis in the data in Table 14.9 or, indeed, whether it is wise to impose any cut-off.

One approach that could be adopted more frequently is to use the diagnostic test to divide subjects into three groups, with a central, 'uncertain' group who would be subjected to further testing. For the data shown in Table 14.9 Weiss *et al.*, (1985) considered assay results between 3.0 and 5.0 as 'borderline'.

Finally, a link with the earlier sections of this chapter is that it is a requirement of a good diagnostic test that the result is repeatable and is subject to minimal inter-observer variation.

Further discussion of the methodology and interpretation of diagnostic tests can be found in the paper by Sheps and Schechter (1984), the series of articles from the Department of Clinical Epidemiology and Biostatistics at McMaster University (1983) and in the books by Galen and Gambino (1975) and Ingelfinger *et al.* (1987). The logic of clinical diagnosis and computer applications are reviewed by Macartney (1987).

14.5 REFERENCE INTERVALS

Diagnostic tests use patient data to classify individuals as either normal or abnormal. A related statistical problem is the description of variability in normal individuals, to provide a basis for assessing test results for other individuals. The most common form of presenting such data is as a range of values, or interval, which encompasses the values obtained from the majority of a sample of normal subjects. The **reference interval** is often referred to as a **normal range** or **reference range**. 'Reference interval' is a better term, both because it avoids confusion with Normal in the statistical sense, and also because the word 'range' suggests that values excluded are by definition abnormal.

Reference intervals are used most often in clinical chemistry, for example to provide a standard reference against which to assess cholesterol

levels in blood samples from patients under investigation. As with diagnostic tests the calculations required are essentially simple and most of the problems are associated with interpretation. One point to note is that the procedure is equivalent to a diagnostic test where we know the specificity (usually 90% or 95%) but nothing else. Clearly such information should not be used *on its own* to make a diagnosis. Detailed discussion on the concepts of reference intervals are given in Solberg (1987) and the papers cited therein.

14.5.1 Selecting a sample

The concept of 'normality' is elusive, and any definition will be specific to the context. Reference intervals are often derived from samples taken in hospital from subjects subsequently found not to be seriously ill, but people in hospital are not normal in the sense of being representative of the healthy population. It is essential to describe how the reference subjects were selected and on what basis their health was determined.

Sample size is also an important consideration, and is discussed in section 14.5.3. Also there may be variation in the distribution of the measurement of interest between different groups of subjects. In particular it is frequently necessary to calculate separate intervals for males and females. There is often also variation by age, especially among children; this topic is considered in section 14.5.4.

14.5.2 Calculating the reference interval

The reference interval is simply the estimated range of values that includes a certain percentage of the values among the relevant population. As with other intervals discussed in earlier chapters, reference intervals usually encompass 90%, 95% or 99% of the values, with 95% the most frequently used. The same method is used whether both low and high values are considered suspicious or only those at one extreme.

There are two basic approaches to the calculation. We can either take the appropriate (per)centiles from the empirical distribution of the observations, or we can use the Normal distribution, perhaps after transforming the data. Many serum constituents, for example, have Lognormal distributions. The options are thus the same as for the general methods introduced in section 3.4 for summarizing the distribution of a set of observations. In that section the serum IgM values from 298 healthy children aged 0 to 6 years were analysed. In section 3.4.2 the $2\frac{1}{2}$th and $97\frac{1}{2}$th centiles were calculated as 0.2 and 2.0 g/l. The range of values from 0.2 and 2.0 thus defines a 95% reference interval using the percentile method. The distribution of IgM was skewed (Figure 3.3) but $\log_{10}$ IgM had a symmetrical distribution (Figure 3.13), with mean −0.158 and standard deviation 0.238.

If we can consider the distribution of $\log_{10}$ IgM as close to Normal we can use the standard Normal distribution to estimate the required centiles (see section 4.5.2). The 95% reference interval for $\log_{10}$ IgM is calculated as mean ±1.96SD, and the values are antilogged to give the 95% reference interval for IgM. We thus calculate first

$$-0.158 - (1.96 \times 0.238) \text{ and } -0.518 + (1.96 \times 0.238)$$

that is, −0.624 and 0.308, and back-transform these values (using 10^x as in section 3.4) to get a 95% reference interval for IgM as 0.24 to 2.03. The two approaches give very similar answers for these data.

As always there are advantages and disadvantages of the alternative approaches and each has strong advocates. The parametric approach depends on the data having a closely Normal distribution, perhaps after transformation. We can use a formal test of non-Normality, as described in section 7.5.3. The Normal plot for the log IgM data in Figure 14.5 shows that the data are indeed close to a Normal distribution. The alternative percentile approach makes no assumptions about the data, but is less reliable when the data are Normal.

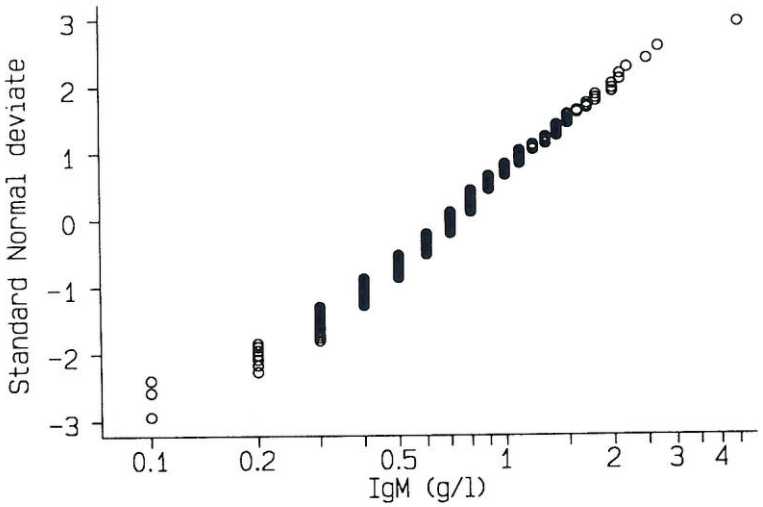

Figure 14.5 Normal plot of log serum IgM data in children (Isaacs *et al.*, 1983).

14.5.3 Uncertainty and sample size

Whereas the parametric approach is based on estimates of the mean and standard deviation, the percentile approach is based on observations in the tails of the distribution. For both methods the reference interval is

obtained as two values which are subject to sampling variability. Several samples from the same population of healthy individuals will give different reference intervals, with the variability depending on sample size. Samples from different populations would be even more variable, and the use of different types of machine to measure the quantity of interest would increase variability further. Table 14.12 shows mean fetal scalp blood pH and reference intervals from 14 different samples of women in 12 centres. Five different types of pH meter were used. There is marked variation in the reference intervals with two (numbers 3 and 14) hardly overlapping. Most noticeable, however, is the fact that most of the studies are very small, all but one being based on fewer than 50 subjects.

Table 14.12 Reference intervals from 14 studies of fetal scalp blood pH (Lumley *et al.*, 1971)

Study	Mean pH	95% reference interval*	Sample size
1	7.29	7.15 to 7.43	43
2	7.29	7.21 to 7.37	24
3	7.29	7.25 to 7.33	10
4	7.30	7.20 to 7.40	12
5	7.30	7.22 to 7.38	18
6	7.30	7.22 to 7.38	129
7	7.32	7.20 to 7.44	16
8	7.32	7.22 to 7.42	49
9	7.35	7.23 to 7.47	45
10	7.35	7.25 to 7.45	26
11	7.35	7.25 to 7.45	29
12	7.35	7.25 to 7.45	21
13	7.37	7.27 to 7.47	45
14	7.38	7.30 to 7.45	22

*mean ± 2SD

The standard error may be obtained for any estimated centile of the Normal distribution. For example, the values describing a 95% reference interval have a standard error of

$$\sqrt{\frac{s^2}{N} + \frac{1.96^2 s^2}{2N}}$$

where s is the standard deviation of the observations. This is approximately equal to $s\sqrt{3/N}$. The widths of confidence intervals for the limits of 95% reference intervals for different sample sizes are shown in Figure 14.6. For sample sizes smaller than about 50 the values defining the reference interval themselves have a confidence interval wider than the standard

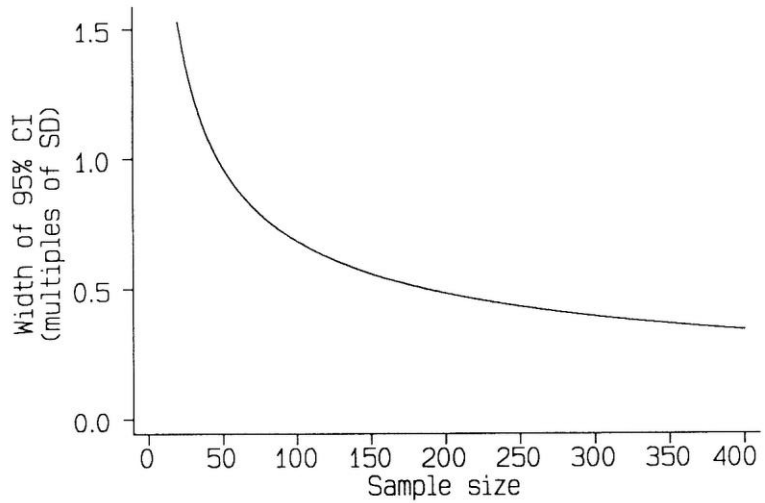

Figure 14.6 Width of parametric 95% confidence interval for limits of reference interval as a multiple of the standard deviation if the data have a Normal distribution.

deviation of the observations. In order to reduce the uncertainty we need much larger samples, preferably of at least 200 observations. Reference intervals derived by the non-parametric percentile method have confidence intervals that are much wider than those shown in Figure 14.6 (Linnet, 1987). The parametric approach is therefore much better if we can make the data conform closely to a Normal distribution, unless we have a very large sample.

14.5.4 Relation to age

Many clinical and biochemical variables vary with age in healthy individuals. For example, as people get older their blood pressure tends to rise and they tend to put on weight. During childhood we are especially likely to find changes with age, and the same applies to both mother and fetus during pregnancy. It is important to investigate possible relations with age, especially for measurements on children or during pregnancy. Failure to do so may lead to the finding of a spurious change in prevalence of abnormality with age.

Not only the mean but also the standard deviation may vary with age. Further, the assessment of Normality needs to be made for small age groups. Regression can be used to fit a curve to the means and, if necessary, a separate curve to the standard deviations. The residuals from these analyses should show no relation to age. Careful analysis of the IgM

data from children aged 6 months to 6 years showed that both the mean and standard deviation of log IgM increased slightly and then decreased in the $5\frac{1}{2}$ year period. Quadratic regression lines were fitted separately to the mean and SD of log IgM for 6 month age groups. These two curves were then combined to give mean ± 1.96SD at each age, and everything was antilogged to give the age-related reference interval shown in Figure 14.7.

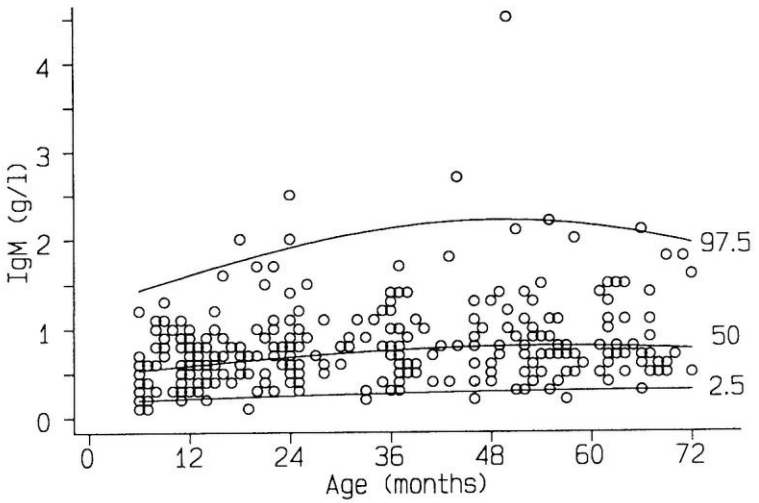

Figure 14.7 95% age-related reference interval for IgM (Isaacs *et al.*, 1983).

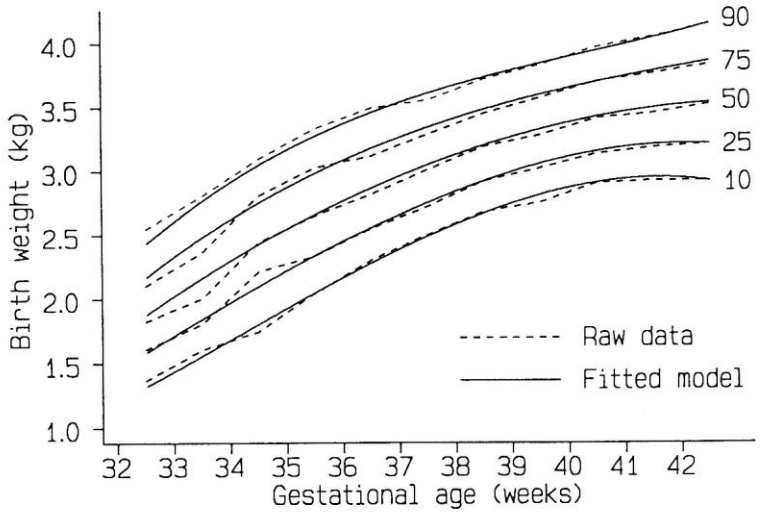

Figure 14.8 Centiles for birthweight of first-born male babies (Altman and Coles, 1980), showing empirical (raw) centiles and curves derived from regression models.

Further details of the method are given in the original paper (Isaacs *et al.*, 1983).

Exactly the same statistical problem arises in constructing 'standards' of fetal or child growth. For example, as well as fitting quadratic curves to mean birthweight as shown in Figure 11.16, cubic curves were fitted to the standard deviations and several age-related centiles obtained, as shown in Figure 14.8 for first born male babies.

14.5.5 Discussion

It is common in clinical practice to classify subjects as normal or abnormal with regard to some clinical or biochemical measurement as an aid to decision-making and thus treatment. When data are available for normal (healthy) and abnormal (ill) subjects we have the type of data that form the basis of a diagnostic test, as discussed in section 14.4. If we wish to use the measurement itself to be a measure of abnormality, then we need to describe the variation among some defined group, usually of healthy subjects. The creation of a reference interval will, however, inevitably lead to the inference that subjects whose values fall outside the interval are abnormal. While this may be true, such an inference is not valid both because the interval by definition excludes a fixed small percentage of healthy subjects, and also because the values of the variable in ill subjects are not known. Where the measurement itself defines the condition, such as blood pressure above a certain level being termed 'hypertension', the logic becomes even more diffuse (Pickering, 1978).

From a statistical point of view, the most interesting question is whether to use the parametric method or the percentile method. While the percentile approach is attractive both in its simplicity and validity for all data sets, there are two important advantages of using the parametric method based on Normal distribution theory. Firstly, the confidence intervals for the values defining the reference interval are much narrower than for the equivalent percentile reference interval. Secondly, the use of the Normal distribution allows any subject's measurement to be expressed as a standard deviation score, and hence located at a particular percentile, which is much more informative than knowing whether they are inside or outside the reference interval. In other words, we can see *how* unusual a value is. (There is a strong analogy here to P values.) Where it is possible, therefore, to treat the data or some transformation of the data as Normal the parametric approach should be used.

The sample size should be large enough to restrict uncertainty about the limits of the reference interval, preferably with a bare minimum of 100 subjects for a parametric analysis and 200 for the percentile method. For age-related intervals it is important to smooth the data across ages. Apart from the fact that smoothly changing values are more plausible, there is

much better statistical use of the data. In all cases, reports of new reference intervals should specify the criteria for inclusion of subjects and the statistical methods used.

14.6 SERIAL MEASUREMENTS

14.6.1 Introduction

Two types of study may yield a series of observations, or **serial measurements**, on each subject. Firstly, there are designed studies where repeated measurements are taken on each individual at specific times chosen in advance. Even when there are complete data for each individual, the appropriate analysis and interpretation of such data are not obvious. Secondly, data can arise from observational studies where multiple measurements are taken at unspecified times. With such data there may be doubts about the reason for the observations. For example, women with many measurements of blood pressure during pregnancy are likely to be a high risk group.

There are several approaches to analysing serial data, each with advantages and disadvantages. In particular some methods are complex both to perform and interpret, and some can be applied only to data at fixed time points. Here I shall consider a simple approach which gives useful results in most situations. It can be applied to experimental or observational data, and can thus be used for structured data sets with missing observations, which is a common phenomenon. A fuller discussion is given by Matthews *et al.* (1990). The method will be illustrated using the data in Table 14.13 and Figure 14.9, which show serum progesterone levels at several times up to two hours after nasal administration of progesterone for four groups of women.

14.6.2 The usual approach to analysis

The most common method of analysing data like these is to perform independent analyses at each time point, such as two-sample t tests or one way analysis of variance. Frequently the data are displayed graphically by a plot joining the mean values at each time point, often with 'error bars' of ± 1 standard error (or perhaps ± 1 standard deviation). There are several important criticisms of this approach:

1. It ignores the design of the study, as no account is taken of the fact that the values at each time point are from the same individuals;
2. The curve joining the means may not be a good indicator of the typical curve for an individual, and will hide any variation in the shape of the curves for different individuals;

Table 14.13 Serum levels of progesterone (nmol/l) after nasal administration in women (Dalton *et al.*, 1987)

	Time after administration (min)										Peak value (nmol/l)	Time to peak (min)
	0	1	3	5	10	15	30	45	60	120		
Group 1 (0.2 ml of 100 mg/ml progesterone in one nostril)												
1	1.0	–	10.0	16.0	22.0	20.0	16.0	–	18.0	14.0	22.0	10
2	6.5	5.7	9.5	11.6	17.5	27.3	28.5	22.4	19.3	10.0	28.5	30
3	3.0	4.0	4.0	13.0	15.8	19.5	21.2	17.9	10.7	13.4	21.2	30
4	1.0	2.1	9.7	–	21.8	–	27.5	–	15.5	6.2	27.5	30
5	1.0	1.0	1.0	4.2	22.6	23.9	45.5	42.6	35.0	10.6	45.4	30
6	1.0	1.0	1.0	1.0	3.9	14.7	17.6	16.1	8.8	10.8	17.6	30
Mean	2.3	2.8	5.9	9.2	17.3	21.1	26.0	24.8	17.9	10.8	27.0	26.7
(SE)	(0.9)	(0.9)	(1.8)	(2.8)	(2.9)	(3.5)	(4.4)	(6.1)	(3.8)	(1.1)	(4.0)	(3.3)
Group 2 (0.3 ml of 100 mg/ml progesterone in one nostril)												
7	1.0	1.5	5.0	11.0	16.0	23.0	15.0	9.0	6.0	5.0	23.0	15
8	1.0	1.0	6.5	20.0	22.5	27.8	19.0	9.0	8.2	8.0	27.8	15
9	1.0	1.0	7.3	7.5	18.0	20.0	18.9	12.8	6.3	4.8	20.0	15
10	3.0	2.5	2.0	2.7	3.4	3.6	14.0	7.3	7.7	4.7	14.0	30
11	8.3	7.5	9.6	11.0	11.5	15.7	15.2	15.8	14.0	11.5	15.8	45
12	6.2	5.9	6.8	7.7	9.0	9.3	12.1	12.2	11.0	9.0	12.2	45
Mean	3.2	3.2	6.2	10.0	13.4	16.6	15.7	11.0	8.1	7.1	18.8	27.5
(SE)	(1.3)	(1.1)	(1.0)	(2.4)	(2.8)	(3.7)	(1.1)	(1.3)	(1.3)	(1.1)	(2.4)	(6.0)

Table 14.13 (*cont'd*) Serum levels of progesterone (nmol/l) after nasal administration in women

	Time after administration (min)								Peak value (nmol/l)	Time to peak (min)
	0	1	3	5	10	15	30	45	60	120		
Group 3 (0.2 ml of 200 mg/ml progesterone in one nostril)												
13	8.4	10.8	8.1	7.8	8.5	12.0	19.8	22.2	25.2	40.5	120	
14	3.5	3.2	3.4	3.3	8.5	9.4	14.5	12.7	11.5	10.2	14.5	30
15	3.5	4.0	4.8	3.5	3.7	13.0	12.5	15.0	22.0	10.5	22.0	60
16	3.7	3.2	4.3	4.5	5.5	8.5	10.3	11.1	8.0	6.0	11.1	45
Mean (SE)	4.8 (1.2)	5.3 (1.8)	5.2 (1.0)	4.8 (1.0)	6.7 (1.2)	10.7 (1.1)	14.3 (2.0)	15.3 (2.5)	16.7 (4.1)	16.8 (8.0)	22.0 (6.7)	63.8 (19.7)
Group 4 (0.2 ml of 100 mg/ml progesterone in each nostril)												
17	5.0	5.6	6.1	7.2	13.8	26.0	26.1	25.7	20.5	11.0	26.1	30
18	4.5	5.1	13.2	21.0	26.8	28.0	22.0	17.8	15.7	14.0	28.0	15
19	8.4	6.2	8.0	18.5	33.8	35.0	26.2	23.0	19.0	12.6	35.0	15
20	4.2	3.2	4.2	4.8	10.3	13.7	17.1	18.3	17.4	15.8	18.3	45
Mean (SE)	5.5 (1.0)	5.0 (0.7)	7.9 (1.9)	12.9 (4.0)	21.2 (5.5)	25.7 (4.4)	22.8 (2.2)	21.2 (1.9)	18.2 (1.0)	13.4 (1.0)	26.9 (3.4)	26.3 (7.2)

Figure 14.9 Serum progesterone levels after nasal administration of progesterone in four groups of women. Data from Table 14.13.

3. It is difficult, if not impossible, to interpret the multiple non-independent P values that are obtained when different groups of subjects are compared;
4. No allowance can be made for any missing observations, so the data at different times may not relate to exactly the same group.

The first point above is the critical one, from which the others follow. It is not at all clear how we would interpret the analysis of the data in Figure 14.9 if, for example, the first two groups were significantly different only at 15 minutes. Further, should we take account of any differences in baseline (time zero) values in the two groups and, if so, how? The purpose of this type of study is usually to assess the response over time, so it is far better to tailor the analysis to the clinical objective.

14.6.3 Analysis using summary measures

Probably the most useful general approach to the analysis of serial measurements is to simplify the analysis by reducing each subject's data to certain features of particular interest. Either a statistical model may be fitted to each individual's data or the necessary quantities can be derived directly from the observed data. These **summary measures** are then

analysed in the same way as if they were the original observations. Clearly this approach relies on the ability to choose summary measures of clinical relevance.

For clinical measurements the only commonly used model is to fit a linear regression of each subject's data on time. The slope of the line represents the rate of change of the measurement per unit of time (e.g. per hour). Clearly, linear regression is appropriate only for data which tend either to rise or fall systematically over time. Many data sets, such as that in Figure 14.9, have a general tendency to rise and then fall (or vice versa). It is unlikely that any simple statistical model would fit such data at all well.

A simpler and more common approach is to take summary statistics directly from the observed data, perhaps after some simple mathematical calculation. Some of the more frequent derived statistics are:

- mean of all the measurements (i.e. ignore the time response)
- height of peak
- time to reach peak
- time to reach a given level
- time to change by a given amount
- time above a given level
- time to achieve maximum change from original level (baseline)
- time to return (near) to baseline level
- change from first to last measurement
- final level (perhaps the average of the last few measurements)
- area under the curve (AUC)

Several of these suggestions incorporate some arbitrary definitions which should be chosen in advance of the analysis rather than after inspection of the data. Several are specifically aimed at data with peaks. Where initial values vary considerably the change from baseline may be used.

The AUC may be interpreted in some circumstances as the cumulative response to the intervention. The calculation is described in section 14.6.5. Note that for equally spaced observations the AUC, which is the hardest of these summary statistics to calculate, is virtually the same as the mean of all the measurements.

Dalton *et al.* (1987) used three measures to summarize the data in Figure 14.9: the time of the peak, the maximum increase from time zero and the AUC. In general it is reasonable to consider two or three derived statistics, but as in any study it is highly desirable to identify a single measure of primary interest. The choice of appropriate measures should relate to the study objectives. For example, if the study is one of treatment efficacy we may reasonably be most interested in the values at the end of the study, perhaps in relation to starting values. If the study is to evaluate the effectiveness of analgesics, then we would probably be interested in the

rapid effectiveness of the drug, perhaps by looking at the timing c peak and the level achieved, and perhaps also the time above some c level.

Although the analysis of summary statistics is usually simple, ther some difficulties with this approach too:

1. it may be difficult to specify the feature(s) of major importance, be the study objective is too vague;
2. the choice of statistics to use may be influenced by inspecting the d
3. it is difficult to study any possible variation between groups in the of the curves (but this is always difficult).

Against these disadvantages we must set some important further a tages; the ability to cope with missing observations (see Table 14.13 variable timing of observations; the ability to handle the comparis serial measurements for the same subjects under different conditions the ease of understanding and explaining the results (a notable pro with several alternative approaches). It may seem that when we ar summary measures we discard a lot of data. In fact the large numb observations is more apparent than real, as consecutive readings i patient will be very similar. The *patient* is the unit of investigation, s easier and more meaningful to handle such data when we have onl value per patient.

14.6.4 Graphical display

Because of the potentially misleading effect of plotting mean values at time point it is important to examine graphs of individuals' data, a possible to include these in the published paper. A graph will show quickly if the curves are similar or dissimilar. Unfortunately, gra display is effective only for small samples. Figure 14.9 showed th serum progesterone data in one form; an alternative is shown in F 14.10.

The summary measures can also be plotted. One interesting form 'peaked' data is to plot the height of the peak against its time. Figure shows such a plot for the progesterone data. This type of plot may patterns that are not evident in other graphs. More generally, w produce a scatter diagram of any two summary measures. The data i example were collected at the same times for all subjects, but gra display may be even more useful for data collected at varying times.

14.6.5 The area under the curve

The area under the curve (AUC) is a useful way of summarizin information from a series of measurements on one individual.

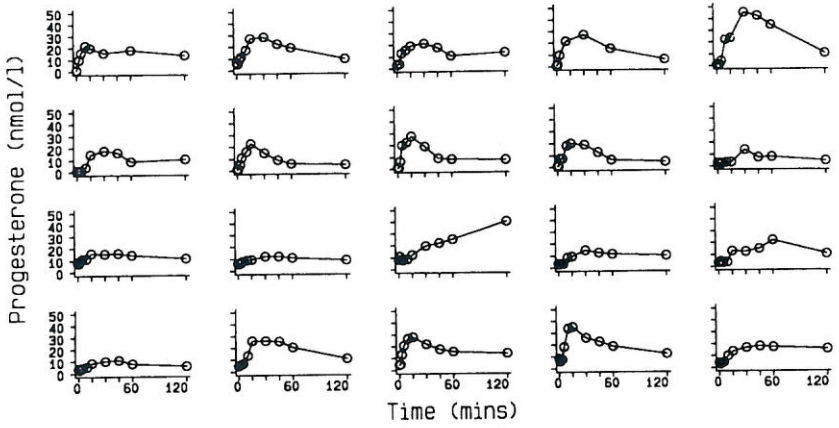

Figure 14.10 Alternative display of serum progesterone data in Figure 14.9.

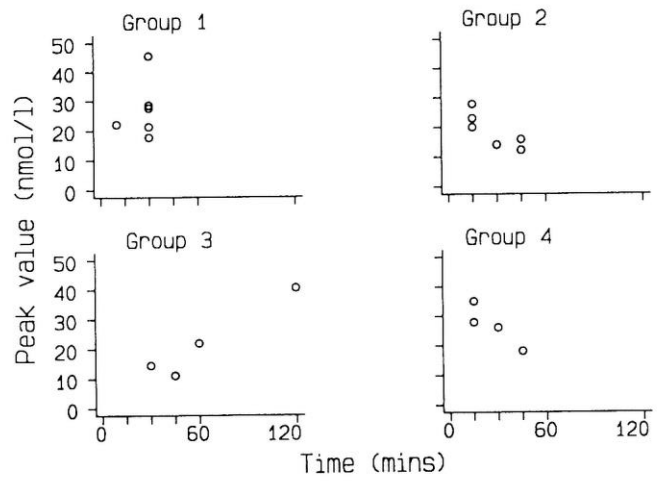

Figure 14.11 Plot of peak values of progesterone by time.

frequently used in clinical pharmacology, where the AUC from serum levels can be interpreted as the total uptake or bioavailability of whatever had been administered.

The data are joined by straight lines to get a 'curve'. The AUC is usually calculated by adding the areas under the curve between each pair of consecutive observations. If we have measurements y_1 and y_2 at times t_1 and t_2, then the AUC between those two times is the product of the time difference and the average of the two measurements. Thus we get

$(t_2 - t_1)(y_1 + y_2)/2$. This is known as the **trapezium rule** because of the shape of each segment of the area under the curve.

If we have $n + 1$ measurements y_i at times t_i $(t = 0, \ldots, n)$ then the AUC is calculated as

$$\frac{1}{2}\sum_{i=0}^{n-1}(t_{i+1} - t_i)(y_i + y_{i+1}).$$

The units of the AUC are the product of the units used for y_i and t_i, for example nmol.min/l, and are not easy to understand. It may be useful to divide the AUC by the total time to get a sort of weighted average level over the time period.

The calculation for the first subject in Table 14.13 goes as follows. There were eight observations for this subject, so seven areas to calculate. We have

$$\text{AUC} = 3 \times \left(\frac{1 + 10}{2}\right) + 2 \times \left(\frac{10 + 16}{2}\right) + 5 \times \left(\frac{16 + 22}{2}\right) + \ldots + 60 \times \left(\frac{18 + 14}{2}\right)$$

$$= 1930 \text{ nmol.min/l}.$$

This value can also be expressed as an average level of 1930/120 = 16.1 nmol/l.

We can calculate the AUC even when there are missing data, except when the final observation is missing.

14.6.6 Interpretation

Performing an analysis that does not relate to the questions of clinical interest often leads to incorrect inferences. When data are analysed separately at each of several time points it is common to see inferences based upon the time when groups become significantly different. Clearly the answer to this question will depend strongly on sample size, and has little if any scientific credibility. Presentation of all the raw data either in a table or figure is valuable, but neither may be feasible in a large study.

The use of summary statistics as the basis of statistical analysis avoids many difficulties by relating the analysis directly to one or more questions of specific interest. Interpretation is usually simplified by having one 'observation' per subject. Simple methods of estimation and hypothesis testing can be used.

14.7 CYCLIC VARIATION

Many measurements vary according to time of day. For example, most people's blood pressure is lowest at night and highest during the morning.

Circadian variation is also seen in many hormone levels and even our height tends to be slightly lower in the evening than in the morning.

Similarly, individual measurements and also population data may vary by month of the year. Table 14.14 shows the number of births with normal and abnormal cord blood IgE levels by month of birth in a study of over 5000 Belgian newborns. A high level of IgE is used to detect those predisposed to become allergic, and the study was carried out to confirm the results of a previous study that had found an association with month of birth.

Table 14.14 Cord blood IgE by month of birth (Kimpen *et al.*, 1987)

	Number of babies			
Month	Total	Normal IgE (≤ 1.0 IU/ml)	Abnormal IgE (> 1.0 IU/ml)	%Abnormal
January	331	319	12	3.6
February	416	401	15	3.6
March	528	503	25	4.7
April	503	481	22	4.4
May	496	468	28	5.6
June	462	447	15	3.2
July	518	504	14	2.7
August	411	396	15	3.6
September	456	449	7	1.5
October	446	437	9	2.0
November	374	368	6	1.6
December	412	398	14	3.4

When data come from ordered groups we should examine directly the possibility of a *linear* trend. With data like the IgE values, which relate to months, the groups are ordered but are also cyclic. Clearly it makes no sense to look for a linear trend; rather, we should explore the possibility of a systematic *cyclic* trend. Data like these may arise from repeated measurement of the same individuals, or where the data at different times are from independent groups of subjects. When data at different times come from the same individuals this analysis is thus a special form of the analysis of serial measurements. Examples are the measurement of hormone levels throughout the menstrual cycle or blood pressure over 24 hours.

Several methods exist for analysing such data. Frequencies can be analysed using a non-parametric method given by Freedman (1979), for example to see if the incidence of new cases of disease varies seasonally. Continuous variables or proportions can be examined by fitting a **sinusoidal**

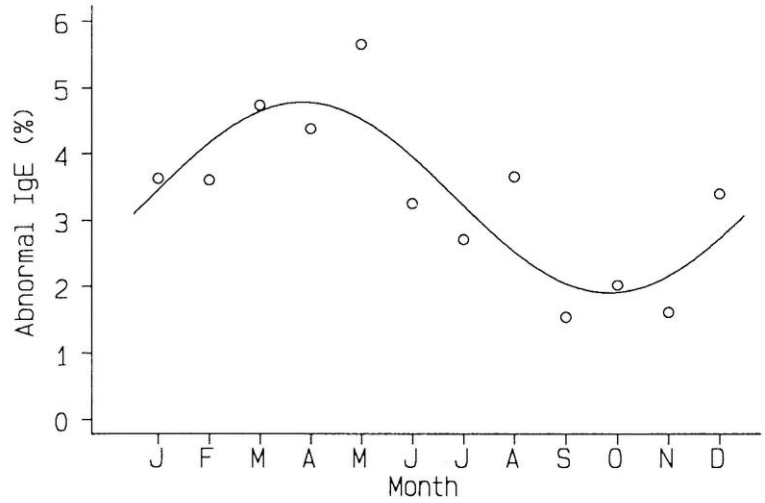

Figure 14.12 Observed percentages of IgE values above 1.0 IU/ml and fitted sine curve.

(or **sine**) curve to the data. This analysis can be regarded as a complex form of regression. Figure 14.12 shows the observed proportions of abnormal IgE values together with the fitted curve. The analysis, which is not described here, shows a highly significant seasonal pattern.

Cyclic variation may require complicated statistical analysis. The purpose of introducing the topic here is to show again how the nature of the data needs to be considered explicitly when selecting the most appropriate analysis. I recommend expert statistical advice for data of this type.

EXERCISES

14.1 The following table shows red cell volume measured simultaneously in 19 patients using radioactive (^{51}Cr) and non radioactive (biotin) cell labels (Cavill *et al.*, 1988):

Patient	^{51}Cr volume (ml)	Biotin volume (ml)
1	1267	1954
2	1710	1651
3	1882	1887
4	1914	2043
5	1940	2054
6	1976	2075
7	2033	1976
8	2039	2120

Patient	^{51}Cr volume (ml)	Biotin volume (ml)
9	2077	2061
10	2087	2152
11	2102	1894
12	2139	1982
13	2184	2153
14	2192	2288
15	2393	2628
16	2425	2495
17	2554	2463
18	2600	3186
19	3420	3488

The authors compared the two sets of data by the Wilcoxon matched pairs rank sum test, for which they got $P > 0.05$. They concluded that the comparison of methods 'showed no consistent clinically significant difference between the two'.

(a) Comment on their analysis and interpretation.
(b) Carry out a better analysis.
(c) What is the relevance of the fact that the patients had all been referred for the measurement of red cell volume.
(d) The largest differences between the methods are those for subjects 1 and 18. The biotin method is affected by prior consumption of eggs, and the authors note that 'at least one of these patients had had an egg for breakfast'. Should the analysis take account of this information?

14.2 Furst and Paulus (1975) reported a study to compare the metabolism of clonixin in 12 patients with rheumatoid arthritis and 12 normal controls. The drug was under investigation as an anti-inflammatory analgesic for treatment of rheumatoid arthritis. Serum clonixin levels were measured at 0, $\frac{1}{2}$, 1, 2, 4, 6 and 8 hours after administration of a single dose of three 250 mg tablets of clonixin. The authors did not report the initial (0 hour) values; the remaining data are shown below:

Patients with rheumatoid arthritis:

	Clonixin levels (μg/ml) Time (hours)					
Patient	0.5	1	2	4	6	8
1	12.70	32.20	42.00	19.80	7.09	2.10
2	18.48	40.24	45.87	15.61	5.58	3.25
3	6.70	20.60	27.70	11.49	2.48	0.56

	Clonixin levels (μg/ml) Time (hours)					
Patient	0.5	1	2	4	6	8
4	24.20	16.20	7.84	5.30	0.38	0.00
5	14.70	28.30	31.90	16.08	9.20	3.60
6	6.55	29.17	33.30	15.17	3.17	0.00
7	41.70	29.40	16.90	7.04	3.48	2.56
8	1.49	47.26	32.78	15.89	4.72	2.61
9	13.04	19.08	39.47	12.42	4.91	2.86
10	29.28	44.94	45.72	12.71	4.43	1.67
11	8.61	20.34	44.33	6.74	2.15	1.11
12	28.10	56.10	36.68	19.10	5.62	1.82

Control subjects:

	Clonixin levels (μg/ml) Time (hours)					
Patient	0.5	1	2	4	6	8
13	58.10	65.90	46.89	17.50	5.40	1.67
14	19.20	22.20	36.50	10.70	2.74	0.94
15	14.21	22.35	32.50	16.49	5.44	2.42
16	5.25	11.13	29.13	7.84	2.21	1.19
17	4.44	43.74	38.22	12.10	2.78	0.02
18	21.20	41.20	46.30	21.70	9.46	4.31
19	31.60	32.80	53.00	39.50	17.88	6.90
20	0.58	2.68	4.01	51.90	21.80	7.64
21	40.90	49.00	38.24	11.09	4.57	0.80
22	31.70	44.20	58.10	24.10	12.60	5.30
23	36.35	47.12	30.96	7.45	2.42	1.75
24	17.57	34.01	40.20	23.80	10.80	7.80

(a) Plot the mean levels in each group.
(b) Compare the peak levels and the area under the curve in the two groups using a suitable analysis assuming that the clonixin level is 0.0 at time zero. (The AUC is easy to calculate in a computer program, but is rather tedious to do by hand.)
(c) Are the plots from (a) a good representation of the data?

14.3 A search of the literature for studies concerning the polygraph (lie-detector) led to the assessment of the sensitivity and specificity of the machine as 0.76 and 0.63 respectively (Brett *et al.*, 1986). It is proposed that the polygraph be used in association with questioning potential blood donors about whether they are drug users. (Assuming that all non-drug users tell the truth.)

(a) If 5% of potential donors use drugs and a third of them lie about it, what proportion of blood donations will be from drug users?
(b) What proportion of people failing the polygraph test will be drug users?

14.4 Acute lower respiratory tract infection is one of the commonest causes of death among infants and under-5s in developing countries. A simple test is needed to identify those infants with acute respiratory infection who have lower respiratory tract infection (LRI) and should receive antibiotics from those with upper respiratory tract infection (URI). The following data come from a study of the usefulness of the respiratory rate for this purpose in infants (Cherian *et al.*, 1988):

Respiratory rate (breaths/min)	Number of children (%) LRI	URI
0–30	1 (1%)	16 (11%)
31–40	4 (3%)	77 (51%)
41–50	10 (7%)	46 (30%)
51–60	41 (29%)	9 (6%)
61+	86 (61%)	3 (2%)
Total	142 (100%)	151 (100%)

(a) Construct 2×2 tables for each of the four cut-offs 30, 40, 50 and 60 breaths/min relating low and high respiratory rate to the correct classification (LRI or URI). Which cut-off gives the best balance of sensitivity and specificity? (This is where their sum is a maximum.)
(b) The authors of the report estimated that the prevalence of LRI among all infants with acute respiratory infection in a developing country is 3%. Which cut-off gives the best balance of positive and negative predictive values when the prevalence is 3%?
(c) If a respiratory rate of >50 breaths/min is taken as an indication of LRI and all such children are treated with antibiotics, what proportion of treated infants will have been treated unnecessarily? What proportion of LRI infants would not get antibiotics?
(d) At present general practitioners cannot tell which infants have LRI and about 80% of infants with respiratory tract infection (LRI or URI) receive antibiotics. What would be the effect on the amount of antibiotics used of the policy suggested in (c)?

14.5 A study of observer variation was performed using radiographic diagnosis of caries on 3869 molars and premolars (Espeland and Handelman, 1989). The following table shows the results for three dentists. Teeth were diagnosed as sound (S) or carious (C).

Dentist 1	Dentist 2	Dentist 3	Frequency
S	S	S	2128
S	S	C	1122
S	C	S	54
S	C	C	226
C	S	S	36
C	S	C	87
C	C	S	7
C	C	C	209

(a) Which pair of dentists agreed best?
(b) Is this a good level of agreement?