The need for caution when analysing data from the two eyes of the same subject has become widely appreciated among ophthalmologists and eye researchers, but the statistical techniques available to overcome the resulting difficulties are much less well understood. In a recent survey of the statistical content of papers in the British Journal of Ophthalmology it was evident that most papers used only relatively simple statistical techniques such as the $\chi^2$ test or $t$ test; that data are frequently collected on pairs of eyes; and that few people know how to analyse such data in ways which are both valid and efficient.

The problems of correctly analysing correlated data are not trivial and this whole area is currently the subject of extensive theoretical research. It is thus not surprising that there is a lag between methods becoming available and their widespread use. A recent review published in the Archives of Ophthalmology\(^2\) has described many of the latest developments. However, without a knowledge of statistical theory it is not easy to put such information to practical use. In this paper we consider one simple technique, the $\chi^2$ test, and show how easy it is to adjust the test when dealing with correlated data.

Example
In a recent community survey\(^1\) we measured the vision of a random sample of Asian and Caucasian residents of Leicester. To illustrate the $\chi^2$ test we have formed a binary categorisation based on the subjects' Snellen visual acuities, with glasses if worn; vision of 6/9 or better is called good, and vision below 6/9 is called poor. We have measurements on both eyes of each of 369 subjects and wish to test whether the extent of poor vision is greater in one ethnic group when compared with the other. The data are presented in Table 1.

We could approach this problem in many ways. We might for example try to phrase our question in terms of subjects rather than eyes. A visually impaired person might be defined as anyone with poor vision in at least one eye. This strategy would remove the problem of correlated data on 738 eyes by using the 369 individuals. However, although a $\chi^2$ test at the person level would be valid, this approach loses some information; we do not distinguish between people with poor vision in one eye compared with both eyes.

To retain the data at the level of eyes it is common practice to analyse the right and left eyes separately. This is also valid but inefficient and can lead to a problem of interpretation. As Table 2 shows a $\chi^2$ test applied to the data on right eyes gives a test statistic of 6.43 and a significant $p$ value of 0.01 and the test applied to left eyes gives a test statistic of 0.42 and a non-significant $p$ value of 0.52. So what are we to conclude? Even when both tests are non-significant one cannot be sure that by combining the data appropriately there would not be sufficient information to produce a statistically significant result.

Of course, were we to combine left and right eyes in a single table, as in Table 2 (c), we would have 738 observations instead of 369 and apparently much greater power to detect a difference between the two ethnic groups. However this is misleading. People with poor vision in one eye frequently have poor vision in the other eye as well. Thus the situation is different from that in which we measure one eye from each of 738 individuals. Our data on two eyes from 369 people are equivalent in terms of power to single eye data on considerably fewer than 738 people.

A number of methods exist for testing correlated data of this type but perhaps the simplest is to make a direct adjustment to the invalid $\chi^2$ test based on combining all eyes.

The adjustment to the $\chi^2$ test depends on the extent to which the results on the two eyes of the same individual are associated. Such an association is measured by the intraclass correlation $\rho$, which we might in this case call an intraperson correlation. We may use an estimate of $\rho$ of the intraclass correlation to adjust the $X^2$ value obtained by pooling all eyes, so giving an adjusted $\chi^2$ statistic $X^2_A$ where,

$$X^2_A = X^2/(1-\rho)$$

This adjusted value can then be compared with the $\chi^2$ distribution in the usual way.

Intraclass correlation is the measure of association employed when there is no distinction between the source of two or more measurements. Thus in family studies, given a pair of siblings which person do we call X and which Y? The solution to the problem is to count each sibling pair twice, once as X and Y and once as Y and X. Calculating the usual Pearson (product moment) correlation with such 'doubled' data gives the intraclass correlation. Snedecor and Cochran describe this further. In our example we could score eyes

<table>
<thead>
<tr>
<th>Ethnic group</th>
<th>Left eye</th>
<th>Right eye</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good</td>
<td>Poor</td>
</tr>
<tr>
<td>Asian</td>
<td>106</td>
<td>4</td>
</tr>
<tr>
<td>Caucasian</td>
<td>141</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 3 Analysis of variance table and estimate of the intraclass correlation

<table>
<thead>
<tr>
<th></th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnic group</td>
<td>1</td>
<td>1006</td>
<td>10056</td>
</tr>
<tr>
<td>Subject within ethnic group</td>
<td>367</td>
<td>131-306</td>
<td>0.3578</td>
</tr>
<tr>
<td>Residual</td>
<td>369</td>
<td>131-306</td>
<td>0.0474</td>
</tr>
<tr>
<td>Total</td>
<td>737</td>
<td>1049-12</td>
<td>0.766</td>
</tr>
<tr>
<td>Intraclass correlation</td>
<td>(0.357-0.0474)</td>
<td>(0.3578-0.0474)</td>
<td>0.766</td>
</tr>
</tbody>
</table>
with good vision as 0 and eyes with poor vision as 1 and then calculate the intraclass correlation between pairs of eyes separately for Asians and Caucasians. The resulting estimated intraclass correlations are 0·848 and 0·686.

In fact with binary data there is a shorter method of calculation that produces identical results. We illustrate it using the data on Asians. First estimate the chance that both eyes of an individual are affected - that is, have poor vision, it is \( P_+ = 48/165 \). Now find the chance that any eye is affected, it is \( P_* = 107/330 \). Given these values the intraclass correlation is,

\[
\frac{(P_+ - P_*)^2}{(P_*(1 - P_*))}
\]

Providing that the two intraclass correlations are in approximate agreement we can combine them to give a single value with which to adjust our \( \chi^2 \) statistic. A weighted average based on the sample sizes gives good results. In our example the weighted average is \( \bar{P} = 0·759 \) and the adjusted \( \chi^2 \) value is \( 4·954/1·759 = 2·82 \). Comparing this with a \( \chi^2 \) distribution with one degree of freedom gives a corrected \( p \) value of 0·09. There is thus insufficient evidence for us to claim that the two ethnic groups have different extents of visual impairment.

In general we recommend obtaining the pooled estimate of the intraclass correlation from the associated analysis of variance table. The analysis of variance can be calculated by once again scoring each eye as 0 for good and 1 for poor and running the data through a statistical package with analysis of variance capabilities. The result of such a calculation is set out in Table 3 and produces a combined estimate of the intraclass correlation of 0·766. The conclusion for our example is not affected.

Rao and Scott have cautioned against the use of the adjusted \( \chi^2 \) statistic when the two groups have different intraclass correlations and they present an alternative approach that remains valid under those circumstances. For the data given in Table 1 the Rao and Scott test gives a \( \chi^2 \) value of 2·78 with a \( p \) value of 0·095. Once again the conclusion is unaffected.

My own simulations suggest that for paired binary data collected on eyes, use of an average intraclass correlation as suggested by Donner would only be misleading in extreme situations where both the two group sizes and the intraclass correlations are widely different. For example, if one group is four times as large as the other and the intraclass correlations are 0·6 in the small group and 0·0 in the large group, then a result based on Donner’s adjusted \( \chi^2 \) statistic will underestimate the impact of the correlation and tests with a nominal \( p \) value of 0·05 will have an actual \( p \) value of about 0·08. The practical impact is very small, none the less it is advisable to calculate the intraclass correlations for both groups in order to check that they are in rough agreement, if only because any difference may itself be of interest.

One advantage claimed for Donner’s method of adjusting the \( \chi^2 \) statistic is that it can make full use of both two eye and single eye data. Thus had we only measured the vision in one eye of some of the subjects, that information could be incorporated into the analysis so optimising the power of the study. However, this possibility should be treated with caution. We need to ask whether the fact that one eye was not measured is linked to the attribute that we were measuring. In many cases they will be associated, possibly quite strongly, and then whatever method we employ we need to interpret the results with greater care.

We have shown how the \( 2 \times 2 \times \chi^2 \) test may be adjusted for use with correlated eye data. It is straightforward to extend this method to a \( k \times 2 \) table, as we would have had if we had divided our sample into more than two ethnic groups. The adjustment is still based on the intraclass correlation which could be obtained either by averaging over \( k \) groups or from the appropriate analysis of variance table.

In our example we have assumed that we wish to test a single eye specific measurement made on subgroups of people. A slightly different situation arises if we make two eye specific binary measurements and wish to test for an association between them. For example, we might have wished to test for an association between our categorisation of vision and the presence or absence of lens opacity. There are now two intraclass correlations that can each be measured over the whole sample, \( \rho_1 \) between visions and \( \rho_2 \) between lens opacities. Rosner and Milton considered this situation and showed that the appropriate form of the adjusted test statistic now becomes

\[
\frac{X^2}{(1+\rho_1 \rho_2)}
\]

Sample size

The factor \( 1+\rho_1 \rho_2 \) (or \( 1+\rho_1 \rho_2 \)) occurs in many contexts when adjusting for correlation between eyes and one of its uses is to obtain an approximate sample size for a study. Rather than \( N \) subjects or \( 2N \) eyes, the effective size of our sample is \( 2N/(1+\rho) \) and \( 1+\rho \) is sometimes called the design effect. Thus in the example of 738 eyes from 369 subjects where the intraclass correlation is 0·76, the design effect is 1·76 and the effective sample size is 419. As far as the question of ethnic origin and vision is concerned, our survey has approximately the same power as would have been obtained from a single eye selected from each of 419 people. Measuring both eyes only gives the equivalent of 50 extra subjects.

At the design stage it is usual to perform power calculations to decide how large a study needs to be. Values for \( N \) assuming independent observations have been widely tabulated. If the usual calculations assuming independence suggest \( N \) individuals are needed, then an equivalent sample would be to take both eyes of \( \frac{N(1+\rho)}{2} \) people, where \( \rho \) is the guess at the intraclass correlation. Unfortunately \( \rho \) is often close to 1·0 for eye data, in which case measuring both eyes will add comparatively little to the information available.

Further methods

The analysis of correlated observations, sometimes called clustered data, has a large literature mainly because of its importance to studies of members of households and families. Cohen and Altman were the first to consider the adjustment of the \( \chi^2 \) test. Their work was extended by Brier and applied to ophthalmic data by Donner who suggested the analysis of variance estimate of the intraclass correlation as well as a method for combining single and paired eye data. Other approaches to testing association in a \( 2 \times 2 \) table of correlated data have been suggested by Rosner and Dallal. Their tests have similar properties to the adjusted \( \chi^2 \) test.

The method of Rao and Scott mentioned above involves estimating the design effect separately for each group and then dividing both the numbers of successes and failures by that design effect. This effectively reduces the size of the data set to its equivalent sample size. They show that if such reduced data are analysed by an unadjusted \( \chi^2 \) test, the analysis will produce valid significance levels. They give an alternative way of estimating the design effect and also apply their method to the problems of testing for a trend in proportions and for a common odds ratio in stratified data.

Where the structure of the data is more complex, an approach based on model fitting is likely to be appropriate. For example, we know that Asians living in Leicester have a lower average age than the Caucasians, so we might have wanted to adjust our survey data for age. Glynn and Rosner have recently reviewed the problems of model fitting with correlated data and give several references. Unfortunately there is little statistical software that can be used to fit these
models without a good understanding of both the package and the theory.

The best approach available at present is that based on generalised estimating equations as described by Liang and Zeger.14 This method was used by Christen et al11 in their recent study of smoking and the risk of cataract in order to adjust the usual logistic model of binary data for the correlation between eyes. Liang and Zeger's method is applicable to a wide variety of problems involving correlated data, for it only requires the researcher to specify a model for the separate sets of observations plus an approximation to the correlation structure. The resulting generalised estimating equations produce valid estimates whose efficiency depends on the quality of the approximated correlation structure. The computer implementation of this method has been discussed by Lipsitz and Harrington10 who include an SAS10 program.

In a series of papers Rosner has developed an alternative conditional method for modelling clustered binary data.18 He has written a FORTRAN program that carries out the rather extensive calculations needed to fit this model which he will supply at a small cost.19 Although this model would probably fit most data sets adequately, the conditioning in its derivation requires that its coefficients be interpreted with great care. Anyone proposing to use this model would be advised to read the comments of Newhaus and Jewell.19

Conclusions
The $\chi^2$ test is widely used and generally well understood by ophthalmologists. However it assumes that the observations are independent, an assumption that usually breaks down when data are collected on pairs of eyes. By calculating the intraclass correlation one can obtain a simple adjustment to the $\chi^2$ test that adequately corrects the significance levels. Donner's adjusted $\chi^2$ test assumes equal intraclass correlation in all of the subgroups, but the test seems very robust for paired data. The method of Rao and Scott20 is only slightly more difficult to implement and might be considered when the correlations differ greatly between the groups.

One can still find examples in the ophthalmic literature where unadjusted $\chi^2$ tests are calculated from data collected on eyes. A useful guide to interpreting such studies is to assume a very high intraclass correlation and to divide the quoted $\chi^2$ statistic by $2.0$. If the result is still significant then it would remain so whatever the actual degree of correlation between eyes.

More detailed analysis of binary data collected on eyes will probably involve some form of regression modelling. This is a fast developing and complex area and anyone contemplating such an analysis would be advised to seek the assistance of a statistician.

JOHN R THOMPSON
Department of Ophthalmology, University of Leicester, Clinical Sciences Building, Leicester Royal Infirmary, PO Box 65, Leicester LE2 7LX

This paper was prepared while the author was visiting the Dana Center, Johns Hopkins Hospital, Baltimore supported by the David Cole Travel Fellowship. Support for the survey used to illustrate the adjusted $\chi^2$ test was given by Help the Aged, the British Council for the Prevention of Blindness, the Royal National Institute for the Blind and the Anne Allerton Fund.