Statistics in Medicine

Chi squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations

Journal:	*Statistics in Medicine*
Manuscript ID:	SIM-06-0349.R2
Wiley - Manuscript type:	Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Campbell, Ian; I C Statistical Services
Keywords:	two-by-two tables, chi squared test, Fisher-Irwin test, power, small sample recommendations

1. INTRODUCTION

One of the commonest problems in statistics is the analysis of a 2 × 2 contingency table, i.e. a table of the form of Table I(a). The results of observational and interventional studies are often summarised in this way, with one binary variable represented by the two rows and the other by the two columns. For example, Yates [1] discussed the data on malocclusion of teeth in infants shown in Table I(b).

Barnard [2] was the first to observe that such 2 × 2 tables can arise through at least three distinct research designs. In one, usually termed a comparative trial, there are two populations (denoted here by *A* and *not-A*), and we take a sample of size *m* from the first population, and a sample of size *n* from the second population. We observe the numbers of *B* and *not-B* in the two samples, and the research question is whether the proportions of *B* in the two populations are the same (the common proportion being denoted here by π). In the second research design, termed cross-sectional or naturalistic [3], or the double dichotomy [2], a single sample of total size *N* is drawn from one population, and each member of the sample is classified according to two binary variables, *A* and *B*. Like comparative trials, the results can be displayed in the form of Table I(a), but the row totals, *m* and *n* are not determined by the investigator. The research question is whether there is an association between the two binary variables. The proportions in the population of *A* and *B* will be denoted here by π_1 and π_2 respectively.

In the third research design, sometimes termed the 2 × 2 independence trial [2, 4], both sets of marginal totals are fixed by the investigator. Here there is no dispute that the Fisher-Irwin test (or Yates's approximation to it) should be used. This last research design is rarely used and will not be discussed in detail.

Statistical tests of 2 × 2 tables from comparative trials and cross-sectional studies

have been under discussion for a hundred years and dozens of research papers have been devoted to them. However, there is still a lack of consensus on the optimum method - most texts recommend the use of the chi squared test for large sample sizes and the Fisher-Irwin test for small sample sizes, but there is disagreement on the boundary between 'large' and 'small' sample sizes, and also on which versions of the chi squared and Fisher-Irwin tests should be used. An informal survey of fourteen medical and general statistics textbooks in print at the time of writing found only two agreeing in their recommendations. This makes life difficult for experienced statisticians. But most statistical calculations are carried out by non-statisticians, and for them the current lack of a consensus is confusing.

1.1. Versions of the chi squared test

In the original version of the chi squared test, due to K. Pearson [5] and Fisher [6], the value of the expression $(ad - bc)^2 N / mnrs$ (nomenclature as Table 1a) is compared with the chi squared distribution with one degree of freedom. Yates [1] recommended an adjustment to the original formula such that $(ad - bc)$ is replaced by $(|ad - bc| - \frac{1}{2}N)$. His basis was that the *P* value from the chi squared test then closely matches that from the Fisher-Irwin test. He termed the adjustment a 'continuity correction', but his theoretical justification for it is disputed (see Discussion), and so it will be termed here an 'adjustment' instead. Fleiss [3] recommends that Yates's adjustment is always used, whereas Armitage *et al.* [7] recommends that it is never used - a change from previous editions of the same textbook.

E. Pearson [8] recommended a third version of the chi squared test, where the expression $(ad - bc)^2 (N - 1) / mnrs$ is compared with the chi squared distribution with one degree of freedom, i.e. differing from the original by the factor $(N - 1) / N$. The

theoretical advantages have been discussed by K. Pearson [8], Barnard [9], Schouten *et al.* [10] and Richardson [11, 12]. A crucial point in its derivation is that, while an unbiased estimate of π is r / N, an unbiased estimate of $\pi(1 - \pi)$ is not $(r / N)(1 - r / N)$ (as has appeared in some books and research papers), but is instead $(r / N)(1 - r / N)\, N / (N - 1)$ [13]. The difference from the original version is small for large sample sizes, but becomes crucial in analyses with small sample sizes, which is the subject of this paper.

Criteria for when the chi squared tests become invalid at small sample sizes are generally based on the smallest expected cell number under the null hypothesis, which is equal to (the smaller of m and n) times (the smaller of r and s) / N. Most recommendations are that a chi squared test should not be used if the smallest expected number is less than 5. This rule is often attributed to Cochran [14, 15], but Yates [1] referred to it as customary practice, and Cochran [16] gave Fisher as the source. Cochran [14] noted that the number 5 appeared to have been arbitrarily chosen, and that the recommendations 'may require modification when new evidence becomes available.' A second recommendation by Cochran [14, 15] that the chi squared test should not be used where $N < 20$ is in fact redundant, as the smallest expected number will always be less than 5 whenever $N < 20$.

1.2. Versions of the Fisher-Irwin test

This test appears in the literature under various names including 'Fisher's exact test'. Because the test was developed independently by Fisher [1, 17] and Irwin [18], and because it is controversial whether the P values obtained are'exact' in all 2×2 tables, the test will be referred to here as the 'Fisher-Irwin test'. The procedure adopted in a one-sided test, as originally described by Fisher and Irwin, is to add the probabilities of the observed table and all tables with the same row and column totals that give a more

extreme difference than the table observed (calculated on the assumption that the marginal totals in the table are fixed). Fisher did not publish a procedure for a two-sided test, but in a private letter, favoured doubling the one-sided P value [19]. Irwin [18] described a different two-sided method: calculating the total probability of tables in either tail that are as likely as, or less likely than the one observed. This will always give a *P* value less than or equal to that of the first method. In this paper, this method will be referred to as 'Irwin's rule', following Cormack and Mantel [20]. More recently, a further version, the mid-*P* probability has gained some support [7, 22 - 25]. For a one-sided mid-*P* test, only half the probability of the observed table is included in the sum. This is based on the observation that the expectation of a one-sided *P* value under a null hypothesis is 0.5 for a continuous distribution (and a perfect test), but is greater than 0.5 when the distribution is discrete (as in the Fisher-Irwin test). However, if only half the probability of the observed data is included in the cumulative sum, the expectation for a one-sided test is then 0.5 [26]. Even though this theoretical justification does not hold for a two-sided test, the mid-*P* version of the Fisher-Irwin test can still be used as a two-sided test by doubling the one-sided mid-*P* value and this is the method recommended by Armitage *et al.* [7]. An alternative mid-*P* two-sided method is to take half the probability of the observed table plus the probabilities of tables in either tail that are less than that of the observed table [25]. This will be referred to here as the mid-*P* test by Irwin's rule.

1.3. Comparison of methods

As well as the above tests, other versions of the chi squared and Fisher-Irwin tests have been proposed, and there are also many alternative tests [4, 19, 21, 27]. Two-by-two tables can also be analysed via a comparison of two proportions together with a

confidence interval for the difference in proportions, but this is outside the scope of this paper.

The different tests will give similar *P* values for large sample sizes, but can give markedly different *P* values with even moderate sample sizes. A choice can be made between the competing tests by comparing, under the null hypothesis, the actual Type I error with the specified significance level α (also referred to as the nominal value). For an ideal test, under any null hypothesis, the actual Type I error (i.e. the total probability of sample tables significant by the test) at a specified significance level α will in fact be equal to α. A test that gives a Type I error appreciably lower than α is sub-optimal because of a loss of power in detecting alternatives to the null hypothesis. Such a test is often said to be *conservative*. A test that has a Type I error appreciably higher than α will be misleading in that it exaggerates the rarity of a table, and consequently is misleading as to the strength of evidence against the null hypothesis. Cochran [16] suggested that a 20% error be permitted in the actual Type I error, e.g. an error up to 1% at the 5% level, and up to 0.2% at the 1% level. Cochran noted that the criterion is arbitrary, but other authors, e.g. Upton [4], have generally followed this or a similar criterion.

Many published studies have evaluated a variety of statistical tests in this way, and the findings can be summarised as follows. For *comparative trials*:

1. Yates's chi squared test has Type I error rates less than the nominal, often less than half the nominal [4, 8, 11, 25, 28 - 33];

2. The Fisher-Irwin test has Type I error rates less than the nominal [4, 25, 28, 29, 32, 34];

3. K. Pearson's ('*N*') version of the chi squared test has Type I error rates closer to the nominal than Yates's chi squared test and the Fisher-Irwin test [4, 28 - 30, 32, 33, 35,

36], but in some situations gives Type I errors appreciably larger than the nominal value;

4. The '*N* - 1' chi squared test, behaves like K. Pearson's '*N* ' version, but the tendency for higher than nominal values is reduced [4, 8, 10, 25, 34, 37, 38];

5. The two-sided Fisher-Irwin test using Irwin's rule is less conservative than the method doubling the one-sided probability [25];

6. The mid-*P* Fisher-Irwin test by doubling the one-sided probability performs better than standard versions of the Fisher-Irwin test, and the mid-*P* method by Irwin's rule performs better still in having actual Type I errors closer to nominal levels [25, 39].

For *cross-sectional studies*, fewer investigations have been carried out, but the limited findings are similar to those for comparative trials [11, 30, 33, 35, 40].

Despite all these studies, no consensus on the optimum test has emerged. The main reason for this is the number of calculations needed for a comprehensive approach to the problem. For example, to compare the performance of statistical tests in analysing a cross-sectional study with sample size of 40, we need to consider the 12341 different sample tables, and calculate the probability of each table together with the *P* value for each table by each test. The number of calculations means that detailed study has not previously been possible - and previous investigations have studied just a subset of the possible combinations of sample size and population proportions. A major difficulty is that normally a null hypothesis is specified only in general terms; for example, in a comparative trial, the null hypothesis is specified as no difference between the two groups in the population proportions, without specifying the value of the common proportion, π. But the probability of a table in a comparative trial is given by [2, 8]

$$\pi^{r}(1 - \pi)^{s}\, m!\, n! / (a!\, b!\, c!\, d!) \qquad (1)$$

and it clearly depends on π. In assessing the performance of a statistical test, we therefore need to consider how it performs over the full range of values of π, and there

should be no region where the Type I error is so high as to render the test invalid. So for a comparative trial we need to consider the maximum Type I error over all values of π, and for a cross-sectional study, we need to find the maximum Type I error over the *two* population parameters π_1 and π_2. Previous studies of the K. Pearson chi squared test in analysing comparative trials have found the maximum Type I error to occur when *m* and *n* are very unequal, for values of π far away from 0.5; in these circumstances Type I errors may be more than twice the nominal.

Hence there is a need for comprehensive information on the performance of statistical tests in analysing comparative trials and cross-sectional studies, particularly when the tests are subject to restrictions such as Cochran's recommendations. These recommendations, which Cochran himself noted were arbitrary and provisional, date back over 50 years and have never been tested. This study addresses that need using the recent advances in computing power. The tests considered are the commonly recommended versions of the chi squared and Fisher-Irwin tests and their close variants. Only two-sided tests are considered in detail because in practice there is rarely justification for a one-sided test; conclusions for one-sided tests will be broadly similar.

1.4. Power and ordering of the sample space

As Storer and Kim [41] pointed out, the finding of one test being conservative compared to another would be of little interest unless it translates into a difference in power. Calculations [28, 41] and Monte Carlo simulations [11] in a limited number of situations have shown that both the Fisher-Irwin test and Yates's chi squared test are less powerful than K. Pearson's chi squared test. However, these kinds of investigations of particular situations are limited by the large number of different null and alternative hypotheses possible. But there is a general principle [4] that if the rejection region for one test (the

set of significant sample tables) is a subset of that for a second test, then the power of the second test will be greater than that of the first for all alternative hypotheses. This can enormously simplify the question of which test is the more powerful to detect real differences; if it so happens that all sample tables that are significant by one test are also significant by a second test, and there are also further sample tables significant by the second test, then the power of the second test will be greater than that of the first for all null and alternative hypotheses, i.e. whatever population proportions are specified, and there is no need to consider which of these are the most likely to occur in practice. This study includes a systematic comparison of the sets of sample tables significant by the tests under scrutiny, which seems not to have been previously performed.

2. METHODS

The study focussed on seven two-sided tests: three versions of the chi squared test and four versions of the Fisher-Irwin test. The tests were:

1. K. Pearson's chi squared test, comparing $(ad - bc)^2N / mnrs$ with the chi squared distribution with one degree of freedom;

2. Yates's chi squared test, comparing $(|ad - bc| - \frac{1}{2}N)^2N / mnrs$ with the chi squared distribution [1]. The adjustment by $\frac{1}{2}N$ was not taken beyond zero, as is standard.

3. The '$N - 1$' chi squared test, comparing $(ad - bc)^2(N - 1) / mnrs$ with the chi squared distribution with one degree of freedom.

4. The Fisher-Irwin test, by doubling the one-sided P value. In obtaining the one-sided value, Fisher [17] described adding the probabilities of tables with the same marginal totals that have 'a discrepancy from proportionality as great or greater than that observed'. Many textbooks give the procedure as progressively decrementing the

smallest cell of the table, but this can give misleading *P* values, for example from the table: $a = 2$, $b = 3$, $c = 4$, $d = 21$; and a better approach, which was adopted here, is to take the tail of the distribution of possible tables with the smaller total probability [19, 42]. If both directions give totals greater than 0.5, the table can be regarded as a central table, not belonging to either tail, with a two-sided probability of 1 [19].

5. The Fisher-Irwin test, taking tables from either tail as likely, or less, as that observed (Irwin's rule);

6. The mid-*P* Fisher-Irwin test, by doubling the one-sided mid-*P* level;

7. The mid-*P* Fisher-Irwin test, using Irwin's rule.

2.1. Calculation of maximum Type I errors in a comparative trial over all values of π

In outline, the method involved dividing the range of possible values of π into a number of intervals, and calculating the Type I error at the boundaries of the intervals together with upper bounds to the Type I error over each interval. By making the intervals sufficiently narrow the upper bounds could be made close to the actual values within any specified accuracy δ.

The Type I error at any value of π is just the sum of the probabilities of those sample tables that are significant at the chosen α. These individual probabilities are given by expression (1), and so the sum of them is a sum of polynomials in π and is therefore a smooth but possibly multimodal function of π. The maximum value of this sum of polynomials cannot be determined analytically, but can be obtained by the following numerical method to an accuracy of δ. Except for tables with zero marginal totals (see below), the probability (1) of any particular table increases from a value of zero, at $\pi = 0$, to a maximum at $\pi = r/N$, and then decreases again to zero at $\pi = 1$. We first consider an interval in π from π_L to π_U. Over this interval, the maximum value of

the probability (1) of a particular table will be at π_L if $r/N \leq \pi_L$, at π_U if $r/N \geq \pi_U$, and otherwise will be at r/N. By summing these maximum values of the probabilities of all the significant tables, we can obtain an upper bound to the Type I error over the interval. This upper bound will in general be larger than the actual maximum because the maximum probabilities for the significant tables will generally occur at different values of π. However, the narrower the interval, the smaller will be the difference between the upper bound and the actual maximum, and by making the interval sufficiently narrow, the difference (and therefore the inaccuracy in the estimate) can be made small. By repeating this process for the whole range of values of π divided into contiguous intervals, and taking the overall maximum, we can obtain a maximum value of the Type I error to an accuracy δ. Values of π can lie between 0 and 1, but for a two-sided test, there is symmetry around 0.5 and only the range 0 to 0.5 need be considered. This method has some similarities to that of Suissa and Shuster [43], although developed independently.

The method used in practice was an iterative technique. The range of 0 to 0.5 in π was divided into ten intervals, and the upper bound of the Type I error was calculated for each interval, together with the actual Type I error at the boundaries of the intervals. Any interval where the upper bound was smaller than the overall maximum plus δ was discarded from further study. Subsequent iterations divided the range in π formed by the residual intervals into smaller intervals, and the process continued until all interval upper bounds were less than the overall maximum plus δ. This overall maximum then gave the maximum Type I error to an accuracy of δ. Further details and the program itself are freely available online [44].

2.2. Calculation of maximum Type I errors in a cross-sectional study

For a cross-sectional study, the maximum Type I error must be found over every possible pair of values of π_1 and π_2. The method used was an extension of the method for comparative trials described above, where instead of one-dimensional intervals in π, the method studied two-dimensional areas of the combined space of π_1 and π_2 [44].

For both research designs, tables where one of the marginal totals was zero (where no statistical test would be applicable) were treated as non-significant, on the basis that if a scientist finds, for example, no cases at all of a particular side-effect in a comparison of two treatments, this is clear evidence against there being a large difference in the rates of that side-effect between the two groups; it is not just an inconclusive result. This treatment of invalid tables as non-significant is in line with previous studies [4, 35].

2.3. Power and sets of significant sample tables

Tests were compared in pairs in the analysis of cross-sectional studies of size equal to all values from 4 to 80, to determine the number of significant sample tables and whether the set of significant sample tables at $P < 5\%$ for one test is a subset of that for the other.

3. RESULTS

3.1. Comparative trials

Figure 1 gives the maximum Type I error across all values of π and all possible m, n pairs at an α of 0.05 for the seven two-sided tests, as a function of N. The findings replicate those of previous studies, with rates considerably higher than the nominal for the K. Pearson and the 'N - 1' chi squared tests. The mid-P Fisher-Irwin test by Irwin's rule also has rates that are too high, but to a lesser extent. Of the remaining four tests

studied, the mid-*P* Fisher-Irwin test by doubling the one-sided value performs closest to the nominal, with a good match at *N* over 30. As in previous studies, Yates's chi squared and the standard Fisher-Irwin tests have rates much lower than the nominal. Several of the lines show a sawtooth effect, especially those for the K. Pearson and '*N* - 1' chi squared tests. This is due to firstly the small number of sample tables that form the bulk of the Type I error under the conditions that make it a maximum, and secondly the arbitrary nature of the 5% cutoff for α. Further details are given online [44].

3.1.1. Policies based on a minimum expected number

This section and the following section show the effect on maximum Type I errors of limiting tests according to expected cell numbers. In these two sections, tables that do not meet the criteria are counted with the non-significant tables. Section 3.4 describes the effect on the Type I error of analysing these excluded tables by the Fisher-Irwin test. Figure 2 (upper) compares Type I errors for four policies relating to K. Pearson's chi squared test. The top line shows no restriction and so repeats the corresponding line in Figure 1. The bottom line shows the effect of restricting use of the test to tables where all expected numbers are at least 5, i.e. to Cochran's recommendations. This is effective in abolishing the excessively high rate of Type I errors, so that the maximum rate lies between 0.05 and 0.06 for most values of *N* greater than 25. The middle two lines (showing restriction to tables with expected numbers of at least 1 and at least 3, respectively) show that there is limited scope for relaxing the limit of 5, since the maximum Type I error has peaks above 0.07, which would seem unacceptable.

Figure 2 (lower) repeats this analysis for the '*N* - 1' chi squared test. The maximum Type I error is generally slightly lower than that for K. Pearson's chi squared test, and even with a restriction of a minimum expected number of 1, the peak rates are

generally in the range of 0.04 to 0.06, even with very small values of *N*.

A similar analysis for the mid-*P* Fisher-Irwin test by Irwin's rule (chart not shown) found that no restriction according to minimum expected numbers could remove the peak Type I error rates well above 0.05. These findings at an α of 0.05 also apply at an α of 0.02, 0.01 and 0.1. In particular: (1) maximum Type I error rates are generally no more than 20% above the nominal (but in a few cases can be up to 40% above the nominal) for K. Pearson's chi squared test when restricted to tables with expected numbers of at least 5; (2) the same applies to the '*N* - 1' chi squared test when restricted to tables with expected numbers of at least 1; and (3) the latter test has an acceptable match to the nominal significance level for lower values of *N* than the former. This can be confirmed by downloading and running the software used [44]. It is concluded that, in the case of comparative trials, the '*N* - 1' chi squared test used at a minimum expected number of 1 is preferable to the K. Pearson chi squared test used with a minimum expected number of 5, in having Type I errors that are a good match to the nominal over a wider range of values of *N*.

3.2. Cross-sectional studies

In general, the findings for cross-sectional studies were a repeat of those for comparative trials. This is not surprising since the frequency of any sample table in a cross-sectional study is the weighted mean of the frequencies of that table under all the possible *m*, *n* pairs that might occur with the total sample size *N*, with the weightings being governed by the value of π for the binary row variable. Three tests (the same three as in comparative trials) had maximum Type I error rates that can be considerably above the nominal 0.05 (Figure 3), while four tests (the same four as in comparative trials) had Type I error rates considerably below the nominal until $N = 30$, or more.

The effect of policies of limiting the tests to sample tables with particular minimum expected numbers was similar to that for comparative trials: K. Pearson's chi squared test restricted to a minimum expected number of 5 and the '*N* - 1' chi squared test restricted to a minimum expected number of 1 both give a good match to the nominal Type I error when α is 0.01, 0.02, 0.05 and 0.1, but the latter test gives a good match to the nominal down to a smaller value of *N*. The mid-*P* Fisher-Irwin test by Irwin's rule did not give a good match to the nominal under any limitation of the minimum expected number. The results for the K. Pearson and the '*N* - 1' chi squared test at an α of 0.05 are given in Figure 4.

This good performance of the '*N* - 1' chi squared test, when restricted to tables with expected numbers of at least 1 is not limited to the *maximum* Type I error. For example, in cross-sectional studies when α is 0.05, the Type I error is generally at least 0.04 for 'central' pairs of values of π_1 and π_2 (e.g. $\pi_1 > 0.3$ and $\pi_2 > 0.3$) for *N* of 14 or more.

3.3. The power of the tests

Of the seven tests whose Type I errors were studied (above), six were compared in terms of the number of different sample tables in a cross-sectional study that are significant at 5%. The mid-*P* Fisher-Irwin test by Irwin's rule was excluded because of the high Type I error rates that cannot be prevented by restriction to sample tables with a minimum expected number.

For each value of *N* investigated (all values from 4 to 80), the number of significant sample tables was highest for the K. Pearson chi squared test, followed by (in order) the '*N* - 1' chi squared test, the Mid-P Fisher-Irwin test by doubling the one-sided value, the Fisher-Irwin test by Irwin's rule, and finally Yates's chi squared test and the

Fisher-Irwin test by doubling the one-sided value had approximately equal numbers. Furthermore, for all N up to 43, the sets of tables significant by these last four tests were each a subset of the set of tables significant by the 'N - 1' chi squared test (which was a subset of the set of tables significant by the K. Pearson chi squared test). From this, it can be concluded that the higher Type I errors of the K. Pearson and 'N - 1' chi squared tests in cross-sectional studies *do* translate into higher power for all alternative hypotheses for all N up to 43. Since the sample tables that occur in a comparative trial are a subset of those that occur in a cross-sectional study, this result also applies to comparative trials.

3.4. Summary of results and consideration of test policies

The results presented here give strong support to the use of the 'N - 1' chi squared test, provided it is restricted to tables where all expected numbers are at least 1. In those relatively few cases where the smallest expected number is less than 1, it seems reasonable to perform an analysis by the Fisher-Irwin test by Irwin's rule, as this has the advantages of being well known, and of having Type I errors close to the nominal. Further calculations in this study found that such a test policy results in small increases in the maximum Type I error rates for both comparative trials and cross-sectional studies, but not to an unacceptable level.

An alternative approach to 2 × 2 tables is to test the value of Z from the ratio of the difference in two proportions to the standard error of the difference (which is exactly equivalent to the chi squared test). To adjust this technique in line with the 'N - 1' chi squared test, the current standard formula for Z [7] can be modified by the factor $\{(N-1)/N\}^{½}$ prior to the comparison with the $N(0,1)$ distribution.

4. DISCUSSION

The results of this study have confirmed and extended the findings of previous studies of the advantages of the '*N* - 1' chi squared test over the alternatives studied. But calculation of Type I error is not the only consideration in the choice of test - some theoretical arguments have been central to the controversies concerning the analysis of 2 × 2 tables that have continued over many decades. The more important theoretical arguments are summarised here; more detailed material is available online [44].

4.1. Whether the row and column totals carry useful information

The Fisher-Irwin test and the '*N* - 1' chi squared test can give very different *P* values for the same set of data, even when the total sample size is quite large e.g. 50. Why should this be, if both are based on valid statistical principles? In fact it can be argued that there is a flaw in the basis of the Fisher-Irwin test, as follows. When putting forward the Fisher-Irwin test, Fisher [17] argued as if the marginal totals (the row and column totals) in a 2 × 2 table carry no useful information:

> 'Let us blot out the contents of the table, leaving only the marginal frequencies. If it be admitted that these marginal frequencies by themselves supply no information on the point at issue, namely, as to the proportionality of the frequencies in the body of the table, we may recognize the information they supply as wholly ancillary; and therefore recognize that we are concerned only with the relative probabilities of occurrence of the different ways in which the table can be filled in, subject to these marginal frequencies'.

This point is crucial to the theory of the Fisher-Irwin test and one might expect Fisher to discuss whether or not it is correct. But Fisher says nothing at all on this question, but

proceeds as if it has been proved. However, it can be shown that the marginal totals *do* carry information on the elements of the table. For example, Berkson [45] pointed out that in the case of a comparative trial with $m = n = 5$ ($N = 10$), information on the observed proportions *is* contained within the column totals. For example, when $r = 0$ (and $s = 10$), the cells of the table, denoted by (a, b, c, d) must be (0, 5, 0, 5), so there is no difference between the two groups in the sample proportion with value A; and when $r = 1$, the cells of the table must be either (1, 4, 0, 5) or (0, 5, 1, 4), and the difference between the sample proportions must be +20% or -20%; when $r = 2$, the cells must be (2, 3, 0, 5) or (1, 4, 1, 4) or (0, 5, 2, 3), and the difference between the sample proportions is +40%, 0%, or -40%; and so on. So once we know the column totals, we *do* have some information on the sample proportions, which is presumably what Fisher meant by 'the proportionality of the frequencies in the body of the table'.

4.2. Whether Yates's continuity adjustment should be applied

The question of a continuity correction to a chi squared test arises because we are trying to determine the probability of *discrete* values of the test statistic by reference to a *continuous* distribution. Generally when this situation arises, and we have, for example, three successive discrete values of a test statistic at x_1, x_2 and x_3, then the probability of the outcome x_2 is best approximated by the probability of the continuous distribution over the interval $\{½(x_1 + x_2), ½(x_2 + x_3)\}$, and in determining a P value, the cumulative probability of outcomes up to and including x_2 is best approximated at the point $½(x_2 + x_3)$ of the continuous distribution (see e.g. Plackett, [46]).

Following Yates's proposal of his continuity adjustment [1], subsequent authors (e.g. [11, 12, 46]) have confirmed that when both pairs of marginal totals are fixed (which rarely occurs in practice), successive possible values of the cross product (ad -

bc) differ by *N* and so Yates's adjustment for continuity of *N*/2 is appropriate. However, in comparative trials and cross-sectional studies, one or both sets of marginal totals are free to vary and successive possible values of (*ad* - *bc*) differ by considerably less than *N*. So Yates's continuity adjustment of *N*/2 is in fact a large *over*correction, and is inappropriate.

4.3 Should randomised trials be analysed differently to other comparative trials

Several authors [e.g. 8, 19] have advocated that randomised trials are analysed on the basis of the hypergeometric distribution and the Fisher-Irwin test, but there are counter-arguments to this, and the present author believes that no distinction should be made. This point is debated in the additional material online [44].

5. CONCLUSIONS AND RECOMMENDATIONS

In 1979, Kempthorne [47] wrote about the analysis of 2 × 2 tables: 'The importance of the topic cannot be stressed too heavily ... 2 × 2 contingency tables are the most elemental structures leading to ideas of association. ... The comparison of two binomial parameters runs through all sciences. ... It is remarkable that a consensus has not been reached.' Over two decades later, these remarks are still applicable, perhaps more so with the increasing use of statistical software by non-statisticians.

The current recommendations on the restriction of the chi squared test to tables with a minimum expected number of at least 5 date back to Cochran [14, 15] and before, but Cochran [14] noted that the number 5 appeared to have been arbitrarily chosen, and could require modification once new evidence became available. This paper provides such new evidence and allows Cochran's guidelines to be updated. The data and

arguments presented here provide a compelling body of evidence that the best policy in the analysis of 2 × 2 tables from either comparative trials or cross-sectional studies is:

(1) Where all expected numbers are at least 1, analyse by the '*N* - 1' chi squared test (the K. Pearson chi squared test but with *N* replaced by *N* - 1),

(2) Otherwise, analyse by the Fisher-Irwin test, with two-sided tests carried out by Irwin's rule (taking tables from either tail as likely, or less, as that observed).

This policy extends the use of the chi squared test to smaller samples (where the current practice is to use the Fisher-Irwin test), with a resultant increase in the power to detect real differences.

ACKNOWLEDGEMENT

Thanks are due to M. H. Campbell for many comments on early versions of the manuscript.

REFERENCES

1. Yates F. Contingency tables involving small numbers and the Π^2 test. *Journal of the Royal Statistical Society Supplement* 1934; **1**:217-235.
2. Barnard GA. Significance tests for 2 × 2 tables. *Biometrika* 1947; **34**:123-138.
3. Fleiss JL. *Statistical Methods for Rates and Proportions*, 2nd edition. Wiley: Chichester, 1981.
4. Upton GJG. A comparison of alternative tests for the 2 × 2 comparative trial. *Journal of the Royal Statistical Society Series A* 1982; **145**: 86–105.
5. Pearson K. On the criterion that a given system of deviations from the probable in

the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5* 1900; **50**: 157-172.

6. Fisher RA. On the interpretation of Π^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 1922; **85**: 87-94.
7. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research*, 4th edition. Blackwell: Oxford, 2002.
8. Pearson ES. The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table. *Biometrika* 1947; **34**: 139-167.
9. Barnard GA. 2×2 tables: A note on ES Pearson's paper. *Biometrika* 1947; **34**: 168-169.
10. Schouten HJA, Molenaar IW, van Strik R, Boomsma A. Comparing two independent binomial proportions by a modified chi square test. *Biometrical Journal* 1980; **22**: 241-248.
11. Richardson JTE. Variants of chi-square for 2×2 contingency tables. *British Journal of Mathematical and Statistical Psychology* 1990; **43**: 309-326.
12. Richardson JTE. The analysis of 2×1 and 2×2 contingency tables: a historical review *Statistical Methods in Medical Research* 1994; **3**: 107-134.
13. Stuart A, Ord JK, Arnold S. *Kendall's Advanced Theory of Statistics*, Vol. 2A, 6th edition. Arnold: London, 1999, p17.
14. Cochran WG. The Π^2 test of goodness of fit. *Annals of Mathematical Statistics* 1952; **25**: 315-345.
15. Cochran WG. Some methods for strengthening the common Π^2 tests. *Biometrics* 1954; **10**: 417-451.
16. Cochran WG. The Π2 correction for continuity. *Iowa State College Journal of*

Science 1942; **16**: 421-436.

17. Fisher RA. The logic of inductive inference. *Journal of the Royal Statistical Society* 1935; **98**: 39-54.
18. Irwin JO. Tests of significance for differences between percentages based on small numbers. *Metron* 1935; **12**: 83-94.
19. Yates F. Tests of significance for 2 × 2 contingency tables (with discussion) *Journal of the Royal Statistical Society Series A* 1984; **147**: 426-463.
20. Cormack RS, Mantel N. Fisher's exact test: the marginal totals as seen from two different angles. *Statistician* 1991; **40**: 27-34.
21. Hill ID, Pike MC. Algorithm 4: TWOBYTWO *Computer Bulletin* 1965; **9**: 56-63.
22. Barnard GA. Discussion on Tests of significance for 2 × 2 contingency tables (by F Yates). *Journal of the Royal Statistical Society Series A* 1984; **147**: 449-450.
23. Barnard GA. On alleged gains in power from lower P values. *Statistics in Medicine* 1989; **8**: 1469-1477.
24. Plackett RL. Discussion on Tests of significance for 2 × 2 contingency tables (by F Yates). *Journal of the Royal Statistical Society Series A* 1984; **147**: 458.
25. Hirji KF, Tan S, Elashoff RM. A quasi-exact test for comparing two binomial proportions. *Statistics in Medicine* 1991; **10**: 1137–1153.
26. Berry G, Armitage P. Mid-P confidence intervals: a brief review. *Statistician* 1995; **44**: 417–423.
27. Sahai H, Khurshid A. On analysis of epidemiological data involving a 2 × 2 contingency table: an overview of Fisher's exact test and Yates' correction for continuity. *Journal of Biopharmaceutical Statistics* 1995; **5**:43-70.
28. Berkson J. In dispraise of the exact test. *Journal of Statistical Planning and Inference* 1978; **2**: 27-42.

29. D'Agostino RB, Chase B, Belanger A. The appropriateness of some common procedures for testing the equality of two independent binomial populations. *The American Statistician* 1988; **42**: 198–202.

30. Grizzle JE. Continuity correction in the Π^2 -test for 2 × 2 tables. *The American Statistician* 1967; **21**: 28-32.

31. Kurtz TE. A role of time-sharing computing in statistical research. *The American Statistician* 1968; **22**: 19-21.

32. Garside GR, Mack C. Actual Type I error probabilities for various tests in the homogeneity case of the 2 × 2 contingency table. *The American Statistician* 1976; **30**: 18-21.

33. Camilli G, Hopkins JD. Applicability of chi-square to 2 × 2 contingency tables with small expected cell frequencies. *Psychological Bulletin* 1978; **85**: 163-167.

34. Overall JE, Rhoades M, Starbuck RR. Small-sample tests for homogeneity of response probabilities in 2 × 2 contingency tables. *Psychological Bulletin* 1987; **102**: 307-314.

35. Camilli G, Hopkins JD. Testing for association in 2 × 2 contingency tables with very small sample sizes. *Psychological Bulletin* 1979; **86**: 1011-1014.

36. Roscoe JT, Byars JA. An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association* 1971; **66**: 755-759.

37. Haber M. An exact unconditional test for the 2 × 2 comparative trial. *Psychological Bulletin* 1986; **99**: 129-132.

38. Rhoades HM, Overall JE. A sample size correction for Pearson chi-square in 2 × 2 contingency tables. *Psychological Bulletin* 1982; **91**: 418-423.

39. Hwang JTG, Yang M.-C. An optimality theory for mid *p*–values in 2 ×2

contingency tables. *Statistica Sinica* 2001; **11**: 807-826.

40. Bradley DR, Cutcomb S. Monte Carlo simulations and the chi-square test of independence. *Behaviour Research Methods and Instrumentation* 1977; **9**: 193-201.

41. Storer BE, Kim C. Exact properties of some exact test statistics for comparing two binomial proportions. *Journal of the American Statistical Association* 1990; **85**: 146-155.

42. Jagger G. Discussion on tests of significance for 2 × 2 contingency tables (by F Yates). *Journal of the Royal Statistical Society Series A* 1984; **147**: 455.

43. Suissa S, Shuster JJ. Exact unconditional sample sizes for the 2 × 2 binomial trial. *Journal of the Royal Statistical Society Series A* 1985; **148**: 317-327.

44. http://www.iancampbell.co.uk/twobytwo 2006.

45. Berkson J. Do the marginal totals of the 2 × 2 table contain relevant information respecting the table proportions? *Journal of Statistical Planning and Inference* 1978; **2**: 43-44.

46. Plackett RL. The continuity correction in 2 × 2 tables. *Biometrika* 1964; **51**: 327-337.

47. Kempthorne O. In dispraise of the exact test: reactions *Journal of Statistical Planning and Inference* 1979; **3**: 199-213.

Table I. Two-by-two tables.

(a) *Nomenclature*

	B	not-*B*	Total
A	*a*	*b*	*m*
not-*A*	*c*	*d*	*n*
Total	*r*	*s*	*N*

(b) *Example data*

	Normal teeth	Malocclusion	Total
Breast-fed	4	16	20
Bottle-fed	1	21	22
Total	5	37	42

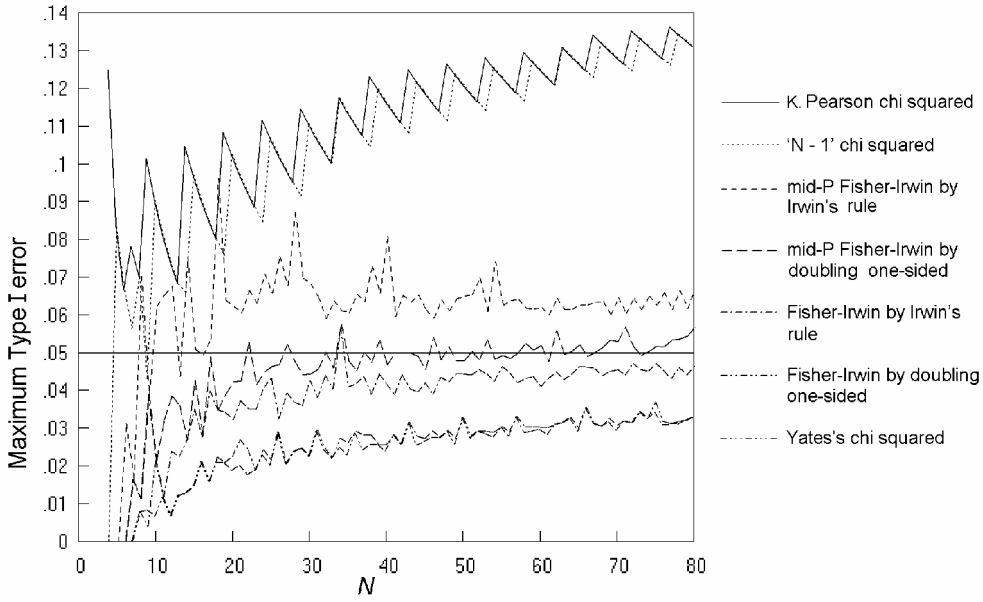

Figure 1. Comparative trials: the maximum Type I error over all values of π and all possible *m*, *n* pairs at a nominal α of 0.05 for seven tests, with an inaccuracy of less than 0.001.

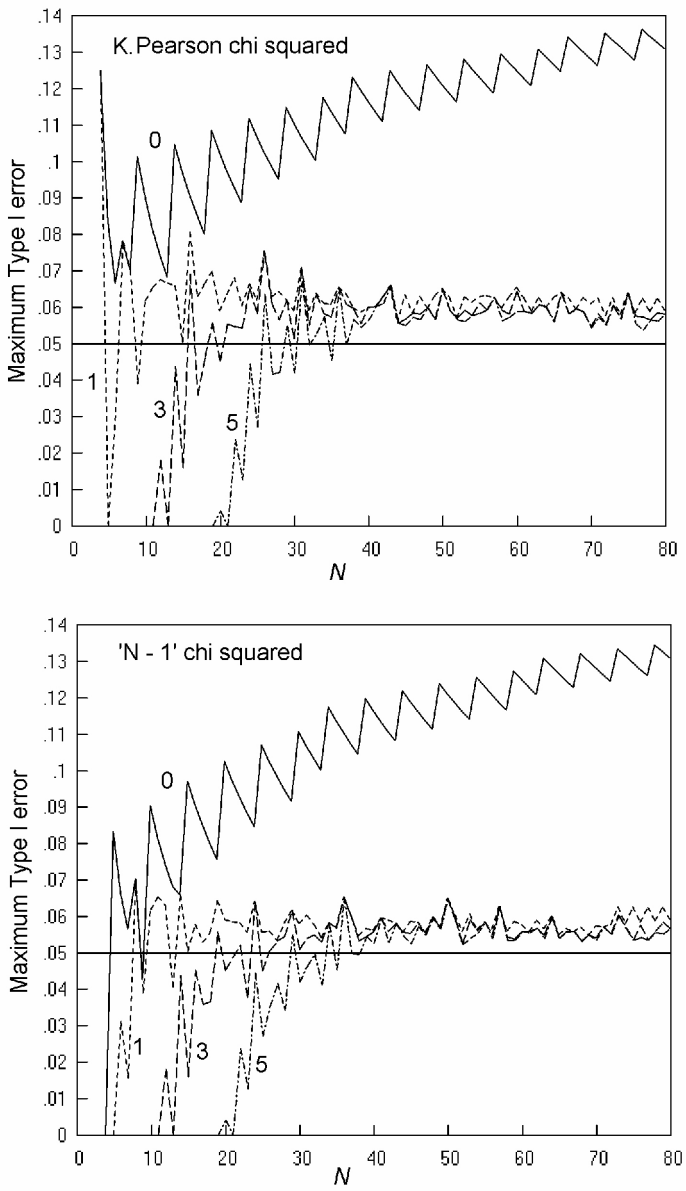

Figure 2. Comparative trials analysed by K. Pearson's and the '*N* - 1' chi squared tests: The maximum Type I error at a nominal α of 0.05 is shown when there is no restriction (labelled '0'), and when the test is restricted to sample tables with expected numbers of at least 1, 3 or 5, with an inaccuracy of less than 0.001.

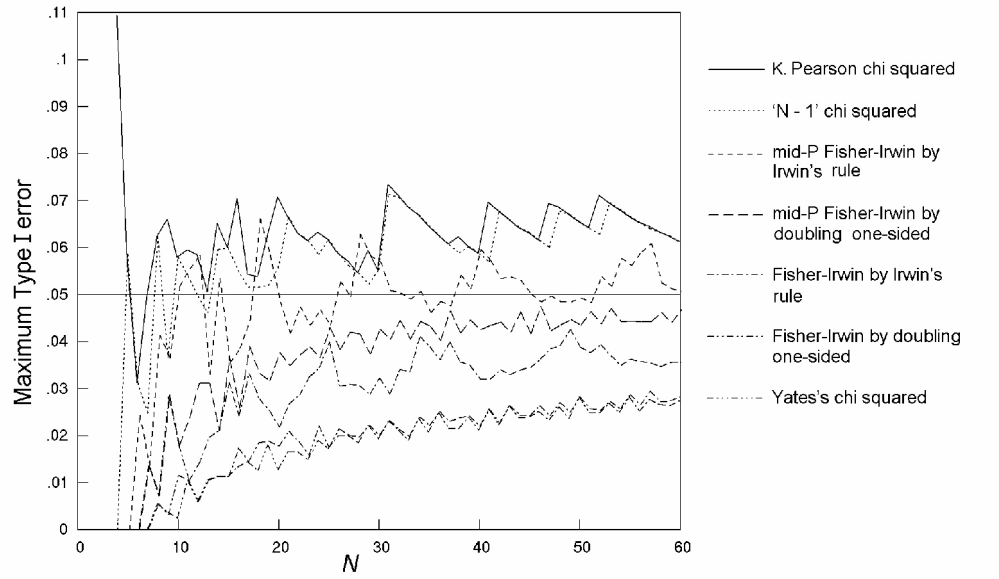

Figure 3. Cross-sectional studies: The maximum Type I error over all values of π_1 and π_2 at a nominal α of 0.05 for seven tests, with an inaccuracy of less than 0.001.

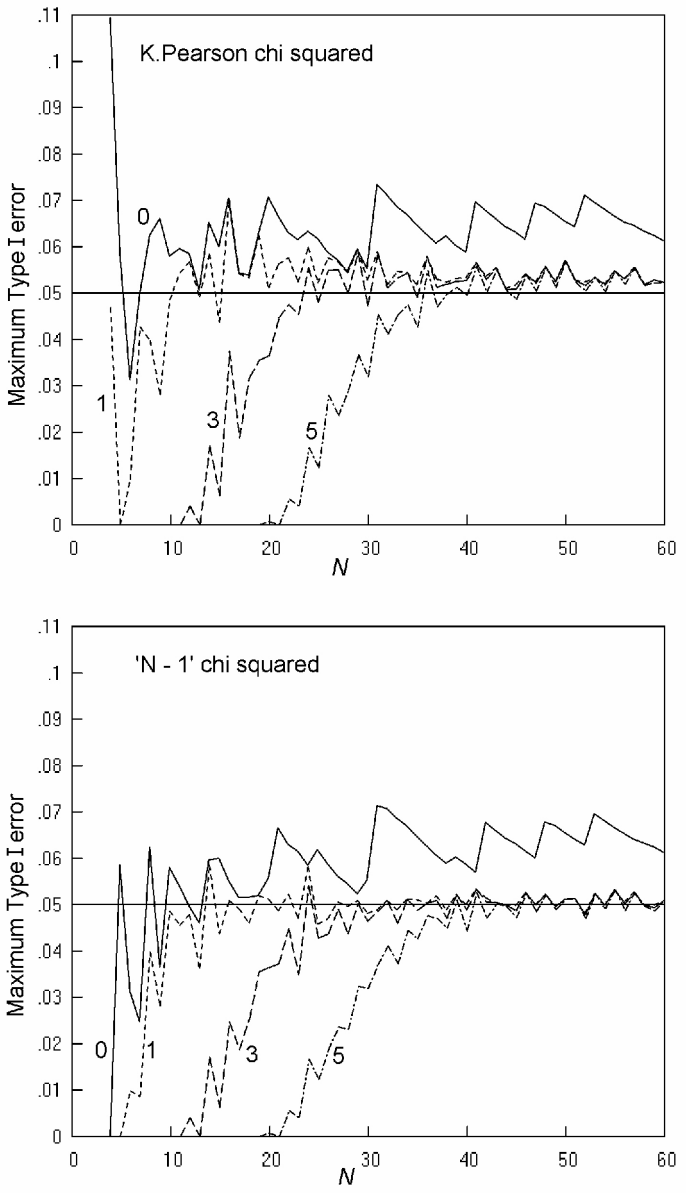

Figure 4. Cross-sectional studies analysed by K. Pearson's and the '*N* - 1' chi squared tests: The maximum Type I error at a nominal α of 0.05 is shown when there is no restriction (labelled '0'), and when the test is restricted to sample tables with expected numbers of at least 1, 3 or 5, with an inaccuracy of less than 0.001.