Analysis of 2 × 2 tables of frequencies: matching test to experimental design

John Ludbrook

Accepted 15 July 2008

Background

Biomedical investigators often use unsuitable statistical techniques for analysing the 2 × 2 tables that result from their experimental observations. This is because they are confused by the conflicting, and sometimes inaccurate, advice they receive from statistical texts or statistical consultants.

Methods

These consist of a review of published work, and the use of five different statistical procedures to analyse a 2 × 2 table, executed by StatXact 8.0, Testimate 6.0, Stata 10.0, SAS 9.1 and SPSS 16.0.

Discussion and Conclusions

It is essential to classify a 2 × 2 table before embarking on its analysis. A useful classification is into (i) Independence trials (doubly conditioned). These almost never occur in biomedical research because they involve predetermining the column and row totals in a 2 × 2 table. The Fisher exact test is the best method for analysing these trials. (ii) Comparative trials (singly conditioned). These correspond to the usual experimental design in biomedical work, in which a sample of convenience is randomized into two treatment groups, so that the group (column) totals are fixed in advance. The proper tests of significance are exact tests on the odds ratio, on the ratio of proportions (relative risk and risk ratio) or on the difference between proportions. (iii) Double dichotomy trials (unconditional). In these, a genuine random sample is taken from a defined population. Thus, neither column nor row totals are fixed in advance. The only practicable test is Pearson’s χ²-test. In analysing any of the above trials, exact tests are to be much preferred to asymptotic (approximate) tests. The different commercial software packages use different algorithms for exact tests, and can give different outcomes in terms of P-values and confidence intervals. The most useful are StatXact and Testimate.

Keywords

Comparative trial, conditioning, double dichotomy trial, exact tests, independence trial, statistics

Most readers of the International Journal of Epidemiology will recognize, in connection with 2 × 2 tables of frequencies, phrases such as Pearson’s χ²-test, Yates’ correction and Fisher’s exact test. A superficial survey of articles published from 1998 through 2007 in the journal, using Google Scholar, suggests that these tests have been used, respectively 5 times, 3 times and 25 times. I shall argue that in most cases these tests were used inappropriately.
Biomedical investigators may be surprised to learn that the analysis of such a simple data set as a $2 \times 2$ table of frequencies (a contingency table) has generated fierce controversies, not merely in the past but also in the present. Yates gave a good account of the history up to 1984. I give a brief summary and update below.

Karl Pearson described the $\chi^2$ goodness-of-fit test in 1900. Unfortunately, he made a mistake in the degrees of freedom with which to evaluate the $\chi^2$-statistic—he proposed $df = 3$ for a $2 \times 2$ table. It was not until 1922 that R. A. Fisher corrected this misapprehension.3 This seems to have been the origin of their life-long feud.4 In 1934, Frank Yates gave a good account of the history up to 1984.1 I give a brief summary and update below.

For $2 \times 2$ tables, his rules were that the Fisher exact test should be used whenever $N < 20$; or, when $N = 20–39$, if the smallest expected frequency in any cell is <5. For tables in which $N \geq 40$, he recommended that Yates' correction to Pearson's $\chi^2$-test should be preferred, because of the enormous computational task of executing the Fisher test in the early 1950s.

In 1945, Barnard proposed an exact test on $2 \times 2$ tables in which it was assumed that only the sample sizes were fixed in advance, and which he suggested was more powerful than Fisher's test.9 Fisher immediately challenged Barnard's argument,10 chiefly on the basis that Barnard was referring his inferences to hypothetical populations. Yet, Fisher accepted that Barnard was, in effect, using a permutation technique to analyse $2 \times 2$ tables—the very technique that Fisher himself had pioneered and which does not presuppose that a population has been randomly sampled.11 Barnard later elaborated on his test, proposing an exact test of significance on the equality of proportions.12 But he retracted his claim of greater power 2 years later,13 under the pressure of Fisher's earlier criticism.10 I shall argue that it is time to rehabilitate Barnard's exact test on proportions.12

Barnard's most lasting contribution was to identify and define the three sampling arrangements that can lead to a $2 \times 2$ table.12 Before describing these, I set out a stereotype $2 \times 2$ table (Table 1). Barnard distinguished (i) the double dichotomy trial, in which only the total number of observations is fixed (that is, $N$ in Table 1); (ii) the $2 \times 2$ comparative trial, in which only the size of the two samples (groups) is fixed in advance [for instance, the column totals ($a+c$) and ($b+d$) in Table 1]; and (c) the $2 \times 2$ independence trial, in which both the column and row totals are fixed in advance: ($a+c$, ($b+d$, ($a+b$) and ($c+d$) in Table 1.
I have never come across an example of an independence trial in biomedical research.

The rather vague hypothesis being tested is that the columns and rows are independent. The alternative is that there is an interaction between columns and rows. The statistical inference refers only to the unique experiment that was conducted. The inference is made under the randomization model of inference.\textsuperscript{15,16}

Double dichotomy trials

In trials such as these, a random sample of predetermined size is taken from a defined population, and each member is classified according to two categories. Let us suppose that investigators take a random sample of size \( N \) from the population of preschool children in the Australian State of Victoria. Each child is classified according to sex (M/F), and also according to obesity (obese/not obese). In the resultant \( 2 \times 2 \) table, the sample size (\( N \) in Table 1) was fixed in advance. However, neither of the two sets of marginal totals was fixed, but depended on the findings of the trial. The table can be described as unconditional.

This trial design lends itself to statistical inferences under the Neyman–Pearson population model of inference.\textsuperscript{16} The statistical inferences refer to defined populations that have been randomly sampled, in this case all Victorian pre-school children. This sort of design is common enough in fields such as epidemiology and demography, but it is rare in biomedical research. Pearson's \( \chi^2 \)-test is often used to analyse the outcomes of such trials. That procedure tests goodness-of-fit; that is, how closely the observed entries in the table coincide with those to be expected under a null hypothesis.

Comparative trials

In this sort of trial, the group sizes (column totals) are determined in advance, by randomizing the members of a sample of convenience to receive one or another treatment. Samples of convenience might be taken from, for instance, patients attending a hospital; or by selecting rabbits, rats or mice from breeding colonies. As an example, imagine that 42 patients with advanced cancer are recruited from St Peregrine's Hospital. They are randomly allocated to one of two groups. Group 1 will be given a placebo. Group 2 will receive a course of treatment with an anti-cancer drug. At the end of 5 years, the patients are classified as alive or dead (Table 3). Only the column marginal totals (group sizes) are fixed in advance, so the \( 2 \times 2 \) table is singly conditioned.

The statistical inference must be made under the randomization model of inference.\textsuperscript{15,16} It is not referable to a population, but only to the patients that were recruited. Any wider inference can be made only by verbal, not statistical, argument. Moreover, the statistical inference can be quite specific. It can refer to

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dead</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>Alive</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>19</td>
</tr>
</tbody>
</table>

\( OR = (17/2)/(14/9) = 5.464 \). \( RR = Risk\ ratio = Ratio\ of\ proportions \( (p_2/p_1) = (17/19)/(14/23) = 1.470 \). Difference in proportions \( (p_2 – p_1) = (17/19) – (14/23) = 0.286 \).

the odds ratio (OR), the difference in proportions or the ratio of proportions (sometimes called the relative risk, risk ratio or RR) (see footnotes to Tables 1 and 3).

Matching tests of significance to trial design

Independence trials

This is the type of trial for which Fisher invented his exact test some 70 years ago. In the unlikely event that a biomedical independence trial was conducted, Fisher's exact test would be the proper one for analysing the result. Yates' correction to the \( \chi^2 \)-test provides only an approximation to the exact test and is primarily of historical interest.

Double dichotomy trials

The choice of best test to analyse the outcome of these trials is not easy. The difficulty is that the sample space for unconditional trials is a multinomial one,\textsuperscript{17} and this is not susceptible to exact analysis even by very powerful computers. The usual solution is to use the \( \chi^2 \)-distribution as an approximation. Because these trials are usually on a very large scale, it is of little consequence whether or not Yates' correction is applied. Other solutions, for instance the Fisher exact test or those listed below under comparative trials, give outcomes that are far too conservative (that is, give too great a value of \( P \)).

Comparative trials

As I suggested earlier, there is a choice of test statistic with which to summarize the information in \( 2 \times 2 \) tables that result from comparative trials. These are the OR, the relative risk, sometimes called the risk ratio (RR) or the difference between proportions. Definitions of these are given in the footnote to Table 1. Comparative trials should be analysed by exact tests on one or other of these statistics. Exact tests on categorical variables are members of the family of permutation (randomization) tests,\textsuperscript{15,16} though I am not aware of this having been pointed out before. The formula for executing a two-sided randomization test, adapted to
a 2 × 2 table with the constraint that the column totals are fixed (single conditioning), is:

\[
P = \begin{cases} 
\text{All tables for which the summary statistic is} & \text{at least as extreme as that observed,} \\
\text{in either direction} & 
\text{All possible tables with the same column totals}
\end{cases}
\]

Examples of summary statistics are the OR, RR and \( p_2 - p_1 \). Because the number of possible tables is limited by prior fixation of the column (group) totals, computation of the two-sided \( P \)-values is well within the capacity of a desktop computer. The commercial software available for these exact tests is listed and discussed in Table 4 and Appendix 1.

There is an apparent difficulty if one of the cells in a 2 × 2 table contains zero. For instance, if 0 is substituted for 2 in cell d of Table 3, then OR becomes infinity. One solution is to swap the rows in this table, when OR becomes 0. In either case, two-sided \( P \) is the same, so the difficulty is more apparent than real.

### How to execute tests on 2 × 2 tables: asymptotic vs exact tests

#### Independence trials

Fisher designed his exact test for just this sort of trial. The test can be performed by hand only on very small tables, but almost all statistical software can execute it for even very large tables. For this reason, I believe that there is no longer a place for Yates’ correction to the Pearson \( \chi^2 \)-test as an approximation to the Fisher exact test.

#### Double dichotomy trials

I can calculate Pearson’s \( \chi^2 \)-statistic on my handheld calculator, using the formula for \( \chi^2 \) that can be found in any elementary statistical text (for Table 3, \( \chi^2 = 4.404 \)). Those who no longer possess a calculator can use almost any general purpose statistical software package.

#### Comparative trials

These require exact (permutation) tests on the OR, RR or difference in proportions (Table 1). These tests cannot be performed by hand, but require desktop computers with fast processors and 256–512 Mb of RAM. Two pieces of commercial software specialize in exact tests (StatXact and Testimate). Some general purpose statistical packages have modules with which some of these exact tests can be performed, for instance, Stata, SAS and SPSS (Table 4 and Appendix 1). However, there are important differences in the algorithms that different software packages use to execute the tests. This is reflected in differences in the resultant \( P \)-values and confidence intervals (CIs) (Table 5).

---

**Table 4** Exact tests on 2 × 2 tables provided by five statistical software packages operating under Windows

<table>
<thead>
<tr>
<th>Exact test</th>
<th>StatXact 8.0</th>
<th>Testimate 6.0</th>
<th>Stata 10.0</th>
<th>SAS 9.1</th>
<th>SPSS 16.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>H₀: OR = 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H₀: RR = 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H₀: ( p_2 - p_1 = 0 )</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

H₀: null hypothesis. Details of vendors in Appendix 1. Refer to Table 5 for a description of how the packages perform the various exact tests.

**Table 5** Outcome of analyses of Table 3: two-sided \( P \)-values and (95% CIs)

<table>
<thead>
<tr>
<th>Test</th>
<th>StatXact 8.0</th>
<th>Testimate 6.0</th>
<th>Stata 10.0</th>
<th>SAS 9.1</th>
<th>SPSS 16.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson ( \chi^2 ) (^{a})</td>
<td>0.0359 (NA)</td>
<td>0.0359 (NA)</td>
<td>0.0359 (NA)</td>
<td>0.0359 (NA)</td>
<td>0.0359 (NA)</td>
</tr>
</tbody>
</table>

**Exact**

<table>
<thead>
<tr>
<th>Test</th>
<th>StatXact 8.0</th>
<th>Testimate 6.0</th>
<th>Stata 10.0</th>
<th>SAS 9.1</th>
<th>SPSS 16.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher exact</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H₀: Exact on OR = 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Actual OR = 5.464)</td>
<td>0.0761 (0.88–57.8)</td>
<td>0.0440 (1.03–41.62)</td>
<td>0.0753 0.0753 (0.88–57.8)</td>
<td>0.0753 (0.88–57.8)</td>
<td>0.0753 (0.88–57.8)</td>
</tr>
<tr>
<td>H₀: Exact on RR = 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Actual RR = 1.470)</td>
<td>0.0391 (1.01–2.31)</td>
<td>0.0440 (1.003–2.35)</td>
<td>0.0753 (1.02–2.11)</td>
<td>0.0753 (0.88–57.8)</td>
<td>0.0753 (0.88–57.8)</td>
</tr>
<tr>
<td>H₀: Exact on ( p_2 - p_1 = 0 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Actual ( p_2 - p_1 = 0.286 ))</td>
<td>0.0391 (0.02–0.53)</td>
<td>0.0440 (0.008–0.54)</td>
<td>0.0753 (0.04–0.53)</td>
<td>0.0753 (0.04–0.53)</td>
<td>0.0753 (0.04–0.53)</td>
</tr>
</tbody>
</table>

H₀, null hypothesis. Actual values of summary statistics from the data of Table 3.

\(^{a}\)Note that the expected frequency for cell d in Table 3 is <5. According to Cochran’s rules, Yates’ correction should be applied. This would result in \( P = 0.0808 \).

\(^{b}\)Performed with double conditioning. NA, CI not applicable.

\(^{c}\)Performed with single conditioning.
Presenting the outcomes of statistical tests on $2 \times 2$ tables

$P$-values or confidence intervals?

There is growing pressure on biomedical investigators to use CIs instead of, or as well as, $P$-values, though this has been resisted. CIs were designed by Neyman to be used in association with the population model of inference, and refer to randomly sampled populations. This fits the case of double dichotomy trials.

There are two special difficulties in using CIs in the analysis of $2 \times 2$ tables. The first is that a CI presupposes that there is a summary statistic around which the CI is located. There is no meaningful summary statistic in the cases of the $\chi^2$ and Fisher exact tests. However, summary statistics can be extracted from $2 \times 2$ tables. These are OR, RR and the difference between proportions ($p_2 - p_1$). It is possible to invert the $P$-values that result from tests on these statistics so as to arrive at CIs. However, the discrete nature of the data in $2 \times 2$ tables, especially when the latter are small, can result in CIs that are incompatible with the $P$-values. For example, in Table 3, Stata gives a 95% CI for RR that does not include 1, whereas $P > 0.05$; and gives a 95% CI for $p_2 - p_1$ that does not include 0, whereas $P > 0.05$ (Table 5).

One- or two-sided $P$-values?

This issue is not statistical, but one of the ethics of biomedical research. Surely no animal experiment or clinical trial should be carried out unless the investigators are genuinely uncertain which of two treatments (in a broad sense of the word ‘treatment’) is the better? On that premise, only two-sided values are safe to use. Investigators are sometimes tempted to report one-sided $P$-values because they are smaller than two-sided values. Peer reviewers of their ethics and grant applications, or of the manuscripts they have submitted for publication, should always insist on two-sided $P$-values.

Which statistics program?

It is safe to use any of the statistics programs listed in Table 4 for the asymptotic form of Pearson’s $\chi^2$-test and for Fisher’s exact test. But for an exact test on OR $= 1$, it is safe only to use the Testimate single-conditioned option. For the exact test on RR $= 1$, and for the exact test on the difference in proportions ($p_2 - p_1 = 0$), it is safe to use StatXact and Testimate, but not Stata.

Acknowledgements

Amanda Thrift, of the Baker IDI Heart & Diabetes Research Institute, helped me with Stata, and Ewa Karafilowska, of The University of Melbourne, with SPSS and SAS. Cyrus Mehta and Volker Rahlfs provided me with valuable information about StatXact and Testimate respectively. Thanks, also, to an anonymous reviewer for his/her constructive comments.

Conflict of interest: None declared.

References

2. Pearson K. On a 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. Philos. Mag. Series 5 1900;50:157–75.
Appendix 1

Commercial software packages for exact tests

I list below five statistics software packages that are written for the Microsoft Windows operating system and with which one can perform some or all of the exact tests listed in Table 4. There is an excellent recent and comparative review by Oster24,25 of the exact analysis of categorical variables in all these packages except SPSS. However, Oster does not indicate the conditioning under which the exact tests were done. I have remedied this omission in Table 5. The algorithms that the software packages use for the various exact tests are not usually made explicit, so I have had to infer them from the outcome of the tests. In particular, the software packages tend to describe as 'conditional' doubly conditioned procedures such as Fisher's exact test; and as 'unconditional' singly conditioned procedures for the OR, RR and \( p_2 - p_1 \).

StatXact 8.0 with Cytel Studio

StatXact 8.0 (Cytel Software Corporation, Cambridge MA, USA) is menu-driven and offers the most comprehensive menu of exact tests.

Testimate 6.0

Testimate 6.0 (Institute for Data Analysis and Study Planning, Gauting/Munich, Germany) also caters for all the exact tests mentioned in this article. Note that it offers a choice between two possible formulations of the exact test on \( RR = 1 \), and it analyses \( OR = 1 \) with single or double conditioning. The slight difference in \( P \)-values from those of StatXact (Table 5) suggests that the two packages use slightly different algorithms.

Stata 10.0

Stata 10.0 (Stata Corporation, College Station, TX, USA) provides 'exact' 95% CIs for the test statistics OR, RR and the difference in proportions. However, the CI for OR is a good deal wider than that given by StatXact, which suggests that this is the result of (inappropriate) double conditioning. The only \( P \)-values that Stata gives are those that result from Fisher's exact test (that is, with double conditioning).

SAS 9.1

SAS 9.1 (SAS Institute Inc, Cary, NC, USA) has introduced modules for exact tests, developed by the Cytel Software Corporation. These include PROC FREQ, which provides two exact tests on \( 2 \times 2 \) tables of frequencies: Fisher’s and an exact test on \( OR = 1 \).

SPSS 16.0

SPSS 16.0 (SPSS Inc, Chicago, IL, USA) very popular statistics package has an Exact Tests add-on (leased from StatXact), with an exact routine for the Fisher test.