Testing association with Fisher's Exact test

Pamela Warner

Reader in Medical Statistics, Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK

Correspondence to
Dr Pamela Warner, Centre for Population Health Sciences, University of Edinburgh, Medical School, Teviot Place, Edinburgh EH8 9AG, UK;
p.warner@ed.ac.uk

Received 3 August 2013
Accepted 3 August 2013

To cite: Warner P. *J Fam Plann Reprod Health Care* 2013;**39**:281–284.

BACKGROUND

Fisher's Exact tests have been used to test association in a paper in this issue of the Journal, namely that by Akintomide *et al.*[1] These notes are intended to provide some supplementary explanation of this method (see Box 1 for a glossary of terms used in this article).

WHAT IS FISHER'S EXACT TEST?

Undoubtedly the most widely known test of association between two binary variables is the 2×2 Chi-square (χ^2) test.[2–5] However, many readers will also have learned about Fisher's Exact test at some point – most likely in a basic statistics course – that Fisher's Exact test is the advised, or in fact the obligatory, alternative to the 2×2 χ^2 test in the situation that 'the sample size is small'.[2–5] It might seem surprising then that Fisher Exact tests have been used for all analyses of association in the article by Akintomide *et al.*, even though the n available for analysis is >100 in all analyses reported, and despite the fact that the cross-tabulations are not 2×2, but 3×3 or, in one case, 3×4.[1]

The fact is, Fisher's Exact test of association between two categorical (classification) variables is much more widely applicable than basic statistics courses have led learners to believe. There is an historical reason why it has been so 'overlooked', and that is because of the torturous arithmetic calculations that are required to achieve the Fisher Exact test for a cross-tabulation with large overall n, even more so to complete tests analogous to 2×2 Fisher Exact test, for tables of larger dimension (R×C rather than 2×2). The calculations necessary would be pretty much impossible using a calculator, and have not even been much available in statistical software for personal computers. It is only with recent improvements in desktop computing power that the necessary procedures have come to be added into statistical software packages.[6]

Fisher's Exact test (or an analogous test for tables larger than 2×2) enables, for *any* cross-classified R×C table, calculation of the exact probability of obtaining a set of cell frequencies at least as extreme as the observed data. Reflection on the size of this calculated probability then allows evaluation of the null hypothesis of no association (or equivalently, of independence) between the two classification variables.

WHEN/WHY IS IT USEFUL?

The well-known χ^2 test is an asymptotic test (i.e. it depends on large-sample approximation) and so the larger the sample the better it will perform. Of course the reverse is also true, which has consequences for the circumstances in which χ^2 is valid (dependable). We have referred above to the well-known caveat that the 2×2 χ^2 test is not valid if there is a small sample. [Small sample size is variously defined by textbooks, along the lines of 'in all cases where total $n<20$, or when $n<40$ and any expected cell count is<5'.][2–4] It is also the case that the χ^2 test is not valid in any R×C tables where more than 20% of the *expected* cell counts for the table are less than 5, or if *any* expected cell count is less than 1.[2–4] In either of these circumstances (small n, or 'lop-sided' classification, i.e. reasonable n but too many small expected counts), Fisher's Exact test is invaluable in enabling a (valid) test of association to be performed. However, it can also be used in tables where these validity concerns do not apply, and in such circumstances has the advantage that it provides an exact probability for the significance test, rather than an approximation.

It should be pointed out also that although Fisher Exact test for R×C tables is now included in many statistical software packages, it remains very demanding on computing power/time, and for some particular table arrangements the calculation would not be

Box 1 Glossary of statistical terms used in this article

Association	Relationship between two variables. For categorical variables, the data can be envisaged as a cross-tabulation of counts of respondents with each combination of values for the two variables. 'An association' means the occurrence for an individual of a particular value of one variable, is associated with (more likely to be in conjunction with) a particular value of the other variable. 'No association' means the distribution of values across rows will be approximately the same in each column, or vice versa. See also *Independence, Ordinal association.*
Asymptotic test (large sample approximation)	A test where the *p* value is obtained by approximation, but it is known that as sample size becomes very large, the calculated *p* value approaches to the true value very closely.
Binary variable	Has only two possible values (e.g. oral contraceptive user or not, female or not).
Categorical variable	Such a variable has a limited set of distinct values (categories), and these values can be nominal (i.e. simply descriptive, such as blood group) or ordered (such as degree of impact, duration of professional experience).
Cell frequency (count)	See *Cross-tabulation.*
Chi-square (χ^2) test	Test applied to cross-tabulated data for two categorical variables, to assess association between them. It is designed for nominal data (no inherent ordering) so for any table of counts the *same* χ^2 would be obtained *whatever* the ordering of rows and/or columns. It compares observed cell counts against what would be expected under the null hypothesis of no association, and the greater the discrepancy, the stronger the suggestion of association.
Confidence interval (95%)	This defines a range of values within which we are 95% confident the true population effect (in this article, rho) lies.
Conservative	Tending to give a *p* value that is larger than it might truly be (i.e. less likely to lead to rejection of the null hypothesis).
Cross-classified	See *Cross-tabulation.*
Cross-tabulation	A way of setting out data for individuals cross-classified by two categorical variables. The size of the table is specified R×C, where R=number of distinct categories in row classification, and C=number of column categories. Each cell of the table gives the frequency count of the number of individuals with that combination of row and column values [e.g. a 3×4 table has 3 rows and 4 columns, comprising 12 distinct cells (combinations)].
Effect size	Used loosely here to mean the *strength* of the association. In 2×2 tables the 'effect' could be summarised in a number of ways including, say, the difference between two study groups in percentage with some characteristic of interest (e.g. subgroup percentages of 34% vs 46% having some condition X, give an effect size 'a difference of 12 %-points in prevalence of X').
Expected (cell counts)	Calculated from the total *n*, and the marginal totals for each of the two classification variables.
Fixed marginal totals	Preset by the design. This is sometimes the case for one variable, say group sizes for a trial, but not often for both variables in a cross-tabulation.
Frequency count	See *Cross-tabulation.*
Hypothesis test	See *Significance test.*
Independence (between two variables)	There is independence between two random variables R and C, if the probability that a participant has any specified value of R is *unchanged* by knowledge of that person's value for variable C. Or, independence is the same as 'no association' between values for R and values for C. See also *Association.*
Marginal totals	The totals for each row and column in a cross-tabulation.
Monte Carlo method	While exact results are preferred because they are reliable, the calculations required are sometimes too unwieldy. The Monte Carlo method is a general iterative method of obtaining an unbiased *estimate* of the exact value it is wished to calculate, by repeatedly sampling subsets of the entire 'problem', obtaining a calculated value for each subset, and then 'averaging' these (subset) values across all repeated iterations. For the Fisher Exact test, the difficulty is usually too many possible tables for which probabilities need to be calculated. In this case, Monte Carlo calculations of Fisher *p* values for a large enough number of subsets of tables, provides an unbiased estimate of the exact *p* value sought.
Nominal variable	Has a set of distinct 'naming' values, such as type of contraception (sterilisation, barrier, hormonal), recruitment method (mailshot, general practitioner, Internet)
Null hypothesis (NH)	A statement, prior to testing, of no effect (e.g. 'no association' between row and column classifications). See also *Significance probability.*
Ordinal association	This is association between two ordinal variables such that, when a person has an 'ordinally higher' response (relative to group) on one of the two variables, he/she tends generally to give a response on the other variable that is also ordinally higher (*direct or positive association*) or tends generally to give a response that is ordinally 'lower' (*inverse or negative association*).
Ordinal variable	Specifically this is a special subset of categorical variables where the values are conceptually ordered (e.g. degree of pain: none, mild, moderate, severe, etc.). In terms of statistical analysis, count variables (e.g. parity: 0, 1, 2, etc.) and continuous variables (e.g. weight: 62, 74, 91, etc.) are also conceptually ordinal, but might have too many possible values to be amenable to categorical methods of analysis. See *Ordinal association.*
p value	See *Significance probability.*
Power	This term is used here, loosely, as the probability of rejecting the stated NH on the basis of the study data, when that is in fact the *correct* decision.

Box 1 Glossary of statistical terms used in this article (continued)

Significance probability (*p* value) — The probability, if the NH is true, of obtaining the observed data (combinations of responses on the two variables) or something more 'extreme' (i.e. further from the NH). The smaller *p* is, then the less likely this data would be under the NH, and so the greater our doubts that NH is indeed true.

Significance test (or hypothesis test) — The process of testing aims to enable a binary *decision* to be made about the NH: reject NH or not. This decision is based on the significance probability (*p* value) obtained via the test. If the *p* value is low enough we will decide the data are so inconsistent with NH, that NH of 'no association' should be *rejected* as untenable – hence, in this application, concluding there must be some association. However, the *p* value reflects the size (power) of the study as well the strength of the association, so a more extreme *p* value does not necessarily mean a 'stronger' association.

Spearman rho (non-parametric correlation coefficient) — Spearman rho is an index of the strength of association between values for two ordinal variables measured on the same individuals. Rho takes values between −1 and 1, with zero indicating no correlation, a positive value indicating a direct or positive correlation, and a negative value an inverse or negative correlation. Values 1 and −1 indicate perfect (direct or inverse) correlation. See also *Ordinal association.*

Valid test — Used here loosely to mean a test that is suited to the research question and data variable(s) to be analysed, in the sense that the data to be analysed satisfy any data assumptions required for the test to perform adequately.

feasible computationally even by personal computer. Therefore it is often the case that an alternative calculation method is provided, a 'Monte Carlo' *estimate* of the exact probability. This calculation method is adopted by the software if the true 'exact' calculation would not be possible. For our purposes we will not distinguish between the two, but for further explanation see Mehta.[6]

WHAT PRECAUTIONS ARE NEEDED?

There are three reservations applying to Fisher Exact test, but note that these also apply to the χ^2 test. First, the test is designed for nominal level data (i.e. categorical but with no inherent ordering). This means that the test will be under-powered if the data variables (and the association) are in fact ordinal, as has been pointed out previously regarding χ^2.[5 7 8] Second, both Fisher and χ^2 are solely significance tests, and as such provide no quantification of the *size* of effect (i.e. the degree/strength of association), which is these days the preferred approach to statistical analysis.[2 5 7] The third reservation is too complex to explain here, but hinges on the fact that the tests are theoretically designed for cross-tabulations where the marginal totals are fixed (i.e. set/specified prior to data collection), not whatever count happens to turn out randomly. Yet tables with both sets of marginal totals fixed are seldom found in health research. There has been considerable debate among statisticians about this issue, and the consequences for analysis findings in a table where marginal totals are *not* fixed in advance. The pragmatic view is that although Fisher's Exact test might tend to be on the conservative side in such circumstances, its use for small samples that are unsuited to χ^2 is acceptable.[4]

EXAMPLE

To illustrate with an example, Figure 1 shows the data reported in the third section of Table 1 of Akintomide *et al.*[1] for the association between health professional tendency to use local anaesthetic (LA) for intrauterine device (IUD) insertions, and the number of insertions performed in the past year. It can be seen that those always/sometimes using LA are more likely to be those who have undertaken more than 50 IUD insertions in the past year. If a standard χ^2 had been performed, despite the fact that the data fails the requirements for χ^2 (in that 33% of the expected cell frequencies are less than 5), the *p* value found would be 0.011. The Fisher Exact probability, as reported, was 0.010, so in this case there is very little disparity (and the Fisher *p* value is not more conservative than χ^2). However, depending on the precise table pattern, disparities can be in the other direction, and/or greater, particularly for smaller *n*.

With respect to the points made above: (1) As is usual in health research, this cross-tabulation did not have fixed marginal totals; the marginal numbers are as occurred randomly in the sample surveyed (e.g. column marginal totals=36, 59 and 32). Nevertheless, Fisher Exact test is regarded as an acceptable test to use. (2) It is the case here that both cross-tabulation variables are ordinal: degree of use of LA, and number of insertions performed. An alternative analysis approach could have

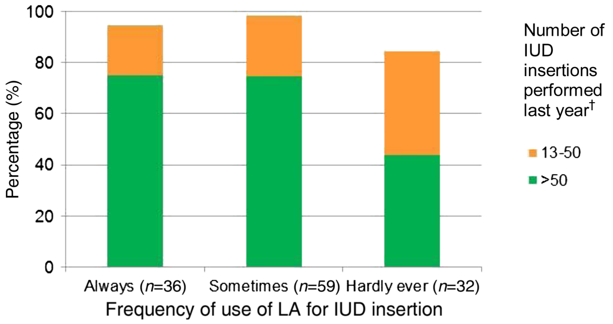

Figure 1 Percentage distribution of respondents by annual number of intrauterine device (IUD) insertions, separately for subgroups based on reported frequency of use of local anaesthetic (LA) *Graph has been created from the third panel of data reported in Table 1 of Akintomide *et al.*[1] †Those performing <13 insertions in past year are not plotted (6%, 2% and 16% across the three columns), but if they had been this would have brought each column up to 100%.

been non-parametric correlation,[7] which would have given a Spearman rank order correlation rho of 0.24 (95% confidence interval 0.06–0.42).

OVERVIEW

Fisher Exact tests are preferable to χ^2 for hypothesis-testing in small or sparse cross-tabulations, whether 2×2 or R×C tables. They can also be used for larger samples to obtain an exact p value.

Competing interests None.

Provenance and peer review Commissioned; internally peer reviewed.

REFERENCES

1 Akintomide H, Sewell RDE, Stephenson JM. The use of local anaesthesia for intrauterine device insertion by health professionals in the UK. *J Fam Plann Reprod Health Care* 2013;39:276–80.

2 Swinscow TDV. *Statistics at Square One* (9th edn) (revised by M J Campbell, University of Southampton). London, UK: BMJ Publishing Group, 1997. http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one [accessed 2 August 2013].

3 Kirkwood BR, Sterne JAC. *Essential Medical Statistics*. Oxford, UK: Blackwell Science, 2003.

4 Bland M. *An Introduction to Medical Statistics* (3rd edn). Oxford, UK: Oxford University Press, 2000.

5 Warner P. Testing and quantifying association in binary data. *J Fam Plann Reprod Health Care* 2009;35:26–27.

6 Mehta CR, Patel NR. *IBM SPSS Exact Tests*. IBM Corporation, 2011.

7 Warner P. Quantifying association in ordinal data. *J Fam Plann Reprod Health Care* 2010;36:83–85.

8 Warner P. Ordinal logistic regression. *J Fam Plann Reprod Health Care* 2008;34:169–170.