
NOTEWORTHY STATISTICS (NS)

Testing and quantifying association in binary data

Pamela Warner

Background

Both Chi-square test and logistic regression have been used to explore association in an article in this issue of the Journal.1 These notes are intended to provide readers with some supplementary explanation and comparison of these statistical methods. [See Box 1 for glossary of terms.]

What is it?
The Chi-square test can be used to test the null hypothesis (NH) of ‘no association’ between two categorical variables. When both outcome variable and explanatory variable are binary, there are only four possible combinations of values (or percentages) of individuals with the outcome.

When and why is it useful?
The Chi-square test is extremely useful in most 2 × 2 tables for testing association. It examines the observed cell counts (the data from the study sample), and compares these to cell counts that would be expected if there were truly no association between the two variables. The Chi-square value calculated from the sample data is referred to tables of the Chi-square distribution, to ascertain the significance probability (under the NH). If this probability is sufficiently low (conventionally <5% or <1%) then the data are judged too unlikely for the NH to be true, so we conclude, by reverse logic, that the NH must be false, that there is therefore an association.

What precautions are needed?
The validity of the Chi-square test for 2 × 2 tables is good if the total n is greater than 40. If the total n is between 20 and 40 then validity remains good provided none of the four expected cell counts is less than 5. Otherwise, if total n is less than 20, Fisher Exact test should be used instead of Chi-square.2 The Chi-square test does not provide a measure of the degree of association. The significance probability cannot serve this purpose, since it reflects the overall n as well as the degree of association. Therefore in reporting Chi-square results it is recommended to present, in addition to the numbers and percentages with the outcome in both explanatory variable subgroups, a summary statistic estimating the association, preferably with a confidence interval for the estimate. Possibilities for the summary are the difference in proportions/percentages or, alternatively, a ratio summary statistic [e.g. odds ratio (OR)].

Example of techniques

Table 1 shows the results when applying (to the data reported by Lorimer et al. in Table 1) the Chi-square test of association of screening uptake and sex, both separately by setting and overall (ignoring setting). Across settings (which are fairly similar in size, n = 104 to 127), the more extreme the sex difference in uptake percentage, the larger is the Chi-square value, and smaller (more significant) the p value. The difference overall is similar to that for the education setting (about 13 percentage points), but the much larger overall n


Public Health Sciences, University of Edinburgh Medical School, Edinburgh, UK
Pamela Warner, BSc, PhD, Senior Lecturer in Medical Statistics and Associate Editor, Journal of Family Planning and Reproductive Health Care

Correspondence to: Dr Pamela Warner, Public Health Sciences, University of Edinburgh Medical School, Teviot Place, Edinburgh EH8 9AG, UK. E-mail: p.warnner@ed.ac.uk

©FSRH J Fam Plann Reprod Health Care 2009: 35(1)
Table 1 Association of screening ‘uptake’ with gender: tests and summaries – overall, separately by setting, and adjusted for setting

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Study n</th>
<th>Difference in percentage ‘uptake’ (female-male pp)</th>
<th>Chi-square ($\chi^2$) test</th>
<th>Logistic regression</th>
<th>p value if study n doubled</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\chi^2$ value</td>
<td>Significance probability (P)</td>
<td>Odds ratio (OR)</td>
</tr>
<tr>
<td>Separately by setting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workplace</td>
<td>104</td>
<td>–7.5 pp</td>
<td>0.7</td>
<td>0.407</td>
<td>0.69</td>
</tr>
<tr>
<td>Education</td>
<td>115</td>
<td>–13.2 pp</td>
<td>3.1</td>
<td>0.078</td>
<td>0.42</td>
</tr>
<tr>
<td>Health and fitness</td>
<td>127</td>
<td>–23.1 pp</td>
<td>6.6</td>
<td>0.010</td>
<td>0.39</td>
</tr>
<tr>
<td>Overall (ignoring setting)</td>
<td>346</td>
<td>–13.3 pp</td>
<td>6.9</td>
<td>0.009</td>
<td>0.546</td>
</tr>
<tr>
<td>‘Adjusted for’ setting</td>
<td>346</td>
<td>–46 pp</td>
<td>–</td>
<td>~</td>
<td>~</td>
</tr>
</tbody>
</table>

(i.e. 346 vs 115) results in a larger Chi-square value, and hence a much smaller p value (i.e. 0.009 vs 0.078).

Table 1 also presents the ORs and their p values, obtained by means of univariate logistic regression analysis of association between uptake and sex (as reported for ‘overall’ in the first column of Lorimer et al. Table 2). The p values are very similar to those obtained for Chi-square and a similar pattern across settings is seen for ORs, in that the more extreme the difference in percentages, the more extreme the OR (smaller/further from 1).

Multivariate logistic regression allows testing/estimation of the association adjusted for setting. The n is unchanged from the overall analysis (n = 346) but the p value is considerably more significant (0.0002 vs 0.009), partly because the OR is now more extreme (0.472 vs 0.546), but mainly because more precise estimation is possible by combining ‘within setting’ estimates. As reported by Lorimer et al. in Table 2, this association is estimated to be even more extreme if adjusted for both setting and age group: OR = 0.42, a 58% lower odds of uptake of screening for females compared to males.

Finally, the analyses have been rerun for the hypothetical situation that the study size is doubled, but the associations within settings and overall are unchanged (i.e. unchanged percentage point differences and ORs). The p values for these are given in the final column of Table 1 above, and it can be seen how much lower they are, despite identical associations, simply because they are derived from a larger study. This highlights the fact that the p value is not a dependable indicator of degree of association.

Overview

The Chi-square test is easy to apply and suitable for testing association in most $2 \times 2$ tables. Since p values reflect n, reporting of results is more informative if the test is supplemented with a summary statistic of the degree of the association (e.g. OR). Univariate logistic regression analysis of a $2 \times 2$ association will give very similar statistical significance to Chi-square. Logistic regression has the advantage that multivariate analyses are possible, allowing adjustment for other variables that might be affecting the estimate of the association between outcome and the explanatory variable of interest.

Statements on funding and competing interests

Funding None identified.

Competing interests None identified.

References


Testing and quantifying association in binary data

Pamela Warner

J Fam Plann Reprod Health Care 2009 35: 26
doi: 10.1783/147118909787072414

Updated information and services can be found at:
http://jfprhc.bmj.com/content/35/1/26.citation

Email alerting service

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/