CHI-SQUARE TEST - ANALYSIS OF CONTINGENCY TABLES

David C. Howell [1]

Professor Emeritus,University of Vermont, USA

The term “chi-square” refers both to a statistical distribution and to a hypothesis testing procedure that produces a statistic that is approximately distributed as the chi-square distribution. In this entry the term is used in its second sense.

PEARSON’S CHI-SQUARE

The original chi-square test, often known as Pearson’s chi-square, dates from papers by Karl Pearson in the earlier 1900s. The test serves both as a “goodness-of-fit” test, where the data are categorized along one dimension, and as a test for the more common “contingency table”, in which categorization is across two or more dimensions. Voinov and Nikulin, this volume, discuss the controversy over the correct form for the goodness of fit test. This entry will focus on the lack of agreement about tests on contingency tables.

In 2000 the Vermont State legislature approved a bill authorizing civil unions. The vote can be broken down by gender to produce the following table, with the expected frequencies given in parentheses. The expected frequencies are computed as $R_i \times C_j/N$, where R_i and C_j represent row and column marginal totals and N is the grand total.

	Vote		
	Yes	No	Total
Women	35 (28.83)	9 (15.17)	44
Men	60 (66.17)	41 (34.83)	101
Total	95	50	145

The standard Pearson chi-square statistic is defined as

$$\chi^2 = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(35 - 28.83)^2}{28.83} + \cdots + \frac{(41 - 34.83)^2}{34.83} = 5.50$$

where i and j index the rows and columns of the table. (For the goodness-of-fit test we simply drop the subscript j.) The resulting test statistic from the formula

[1] David Howell is a Professor Emeritus (since 2002), and former chair of the Psychology department at the University of Vermont (1987–1992) and (2000–2002). Professor Howell’s primary area of research is in statistics and experimental methods. He has authored well known texts: *Statistical Methods for Psychology* (Wadsworth Publishing, 7th ed., 2009), *Fundamental Statistics for Behavioral Sciences* (Wadsworth Publishing, 7th ed., 2010), and is a co-author (with Brian Everitt) of a four volume *Encyclopedia of Statistics in Behavior Science* (Wiley & Sons, 2005).

on the left is approximately distributed as χ^2 on $(r-1)(c-1)$ degrees of freedom. The probability of $\chi^2 \geq 5.50$ on 1 *df* = .019, so we can reject the null hypothesis that voting behavior is independent of gender. (Pearson originally misidentified the degrees of freedom, Fisher corrected him, though Pearson long refused to recognize the error, and Pearson and Fisher were enemies for the rest of their lives.)

LIKELIHOOD RATIO CHI-SQUARE

Pearson's chi-square statistic is not the only chi-square test that we have. The likelihood ratio chi-square builds on the likelihood of the data under the null hypothesis relative to the maximum likelihood. It is defined as

$$\begin{aligned} G^2 = 2\sum O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right) &= 2\left[35\ln\left(\frac{35}{28.83}\right) + 9\ln\left(\frac{9}{15.17}\right) + 60\ln\left(\frac{60}{66.17}\right)\right. \\ &\left. +41\ln\left(\frac{41}{34.83}\right)\right] \\ &= 5.81 \end{aligned}$$

This result is slightly larger than the Pearson chi-square of 5.50. One advantage of the likelihood ratio chi-square is that G^2 for a large dimensional table can be neatly decomposed into smaller components. This can not be done exactly with Pearson's chi-square, and G^2 is the usual statistic for log-linear analyses. As sample sizes increase the two chi-square statistics converge.

SMALL EXPECTED FREQUENCIES

Probably no one would object to the use of the Pearson or likelihood ratio chi-square tests for our example. However, the chi-square statistic is only approximated by the chi-square distribution, and that approximation worsens with small expected frequencies. When we have very small expected frequencies, the possible values of the chi-square statistic are quite discrete. For example, for a table with only 4 observations in each row and column, the only possible values of chi-square are 8, 2, and 0. It should be clear that a continuous chi-square distribution is not a good match for a discrete distribution having only 3 values. The general rule is that the smallest expected frequency should be at least five. However Cochran (1952), who is generally considered the source of this rule, acknowledged that the number "5" seems to be chosen arbitrarily.

Yates proposed a correction to the formula for chi-square to bring it more in line with the true probability. However, given modern computing alternatives, Yates' correction is much less necessary and should be replaced by more exact methods.

For situations in which we do not satisfy Cochran's rule about small expected frequencies, the obvious question concerns what we should do instead. This is an issue over which there has been considerable debate. One of the most common alternatives is Fisher's Exact Test (see below), but even that is controversial for many designs.

ALTERNATIVE RESEARCH DESIGNS

There are at least four different research designs that will lead to data forming a contingency table. One design assumes that all marginal totals are fixed. Fisher's famous "tea-tasting" study had four cups of tea with milk added first and four with milk added second (row totals are fixed). The taster had to assign four cups to each guessed order of pouring, fixing the column totals. The underlying probability model is hypergeometric, and Fisher's exact test (1934) is ideally suited to this design and gives an exact probability. This test is reported by most software for 2×2 tables, though it is not restricted to the 2×2 case.

Alternatively we could fix only one set of marginals, as in our earlier example. Every replication of that experiment would include 44 women and 101 men, although the vote totals could vary. This design is exactly equivalent to comparing the proportion of "yes" votes for men and women, and it is based on the binomial distribution. The square of a z-test on proportions would be exactly equal to the resulting chi-square statistic. One alternative analysis for this design would be to generate all possible tables with those row marginals and compute the percentage of obtained chi-square statistics that are as extreme as the statistic obtained from the actual data. Alternatively, some authorities recommend the use of a mid-p value, which sums the probability of all tables less likely than the one we obtained, plus half of the probability of the table we actually obtained.

For a different design, suppose that we had asked 145 Vermont citizens to record their opinion on civil unions. In this case neither the Gender nor Vote totals would be fixed, only the total sample size. The underlying probability model would be multinomial. Pearson's chi-square test would be appropriate, but a more exact test would be obtained by taking all possible tables (or, more likely, a very large number of randomly generated tables) with 145 observations and calculating chi-square for each. Again the probability value would be the proportion of tables with more extreme outcomes than the actual table. And, again, we could compute a mid-p probability instead.

Finally, suppose that we went into college classrooms and asked the students to vote. In this case not even the total sample size is fixed. The underlying probability model here is Poisson.

Computer scripts written in R are available for each model with a fixed total sample size at
`http://www.uvm.edu/~dhowell/StatPages/chi-square-alternatives.html`

SUMMARY

Based on a large number of studies of the analysis of contingency tables, the current recommendation would be to continue to use the standard Pearson chi-square test whenever the expected cell frequencies are sufficiently large. There seems to be no problem defining large as “at least 5.” With small expected frequencies Fisher’s Exact Test seems to perform well regardless of the sampling plan, but randomization tests adapted for the actual research design, as described above, will give a somewhat more exact solution. Recently Campbell (2007) carried out a very large sampling study on 2×2 tables comparing different chi-square statistics under different sample sizes and different underlying designs. He found that across all sampling designs, a statistic suggested by Karl Pearson’s son Egon Pearson worked best in most situations. The statistic is defined as $\chi^2 \frac{N}{N-1}$. (For the justification for that adjustment see Campbell’s paper.) Campbell found that as long as the smallest expected frequency was at least one, the adjusted chi-square held the Type I error rate at very nearly α. When the smallest expected frequency fell below 1, Fisher’s Exact Test did best.

References

[1] Campbell, I. Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine*, 26, 3661–3675, (2007).

[2] Cochran, W. G. The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, 25, 315–345, (1952).

[3] Fisher, R. A. The logic of inductive inference. *Journal of the Royal Statistical Society.*, 98, 39–54, (1934).