

# THE UNIVERSITY OF TEXAS AT AUSTIN SCHOOL OF INFORMATION

## MATHEMATICAL NOTES FOR LIS 397.1 INTRODUCTION TO RESEARCH IN LIBRARY AND INFORMATION SCIENCE

Ronald E. Wyllys  
Last revised: 2003 Jan 15

### THE CHI-SQUARE TESTS AND THEIR USES

#### Two Uses of the Chi-Square Statistic

The chi-square distribution has two main uses in basic statistical inference. One of these, the *chi-square goodness-of-fit test*, is to determine whether a variable appears to follow some specified distribution. An example of this use might involve gathering a sample of observed values of a certain variable and seeing whether they appear to be Gaussianly distributed. Although this use can be important, it is infrequent.

The other main use of chi-square is encountered much more often. This use, the *chi-square test of association*, is to determine whether two categorical variables appear to be associated, in the sense that they exhibit a tendency for their respective values to co-occur in some pattern.

#### The Chi-Square Test of Association

The chi-square test of association involves the cross-tabulation of the frequencies with which the variables co-occur in a sample. For example, suppose you have a sample of book-circulation transactions that occurred in a university library system during a certain week. Suppose further that each of the circulation transactions can be characterized by the branch from which the book was borrowed and by the class (undergraduate student, graduate student, faculty member, university staff member, or other) of the borrower. Then you can construct a table such as the following one, which displays the numbers of book-circulation transactions in your sample falling into each of the possible combinations of categories.

	Undergrad	Graduate	Faculty	Staff	Other	
Main Library		61		87	99	80 53
Undergraduate Library		113		63	43	101 47
Special Collections Lib.		35		91	88	52 39

From the data in this table, you can determine whether it is reasonable to believe that there is an association between the category of borrower and the branch of the library system used by the borrower. That is, you can find out whether it appears that different types of borrowers prefer different library branches.

For either of these uses of the chi-square test, the calculation of the chi-square statistic is the same. To each observed frequency,  $O_i$ , you match an expected frequency,  $E_i$ . What  $E_i$  is depends on which use of the chi-square you are employing. In the case of the chi-square goodness-of-fit test, you have some idea (theory, model) concerning the distribution of the variable, and you compare the actual observed frequencies with the expected frequencies that your model specifies. In the case of the chi-square test of association the expected frequency in each cell in the cross-tabulation table is calculated as follows:

$$\text{expected frequency in cell} = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

The terms "row total" and "column total" refer to the specific row and specific column in which is located the cell whose expected frequency you are calculating.) These expected frequencies are those that are derived from, and hence correspond to, the null hypothesis (or model) that there is no association between the two variables by which the observations have been categorized. For example, here again is the table above, now including the expected frequencies.

	Undergrad	Graduate	Faculty	Staff	Other	Row Totals
Main	61	87	99	80	53	380
Library	75.4943	87.0532	83.0798	84.1635	50.2091	
Undergraduate	113	63	43	101	47	367
Library	72.9116	84.0751	80.2376	81.2842	48.4914	
Special Col- lections Lib.	35	91	88	52	39	305
	60.5941	69.8717	66.6825	67.5523	40.2994	
Column Totals	209	241	230	233	139	Grand Total 1052

Once these expected frequencies have been calculated, you are ready to calculate the chi-square test statistic itself. It is given by

Formula 1

$$\chi^2_{obs} = \sum \frac{O_i - E_i}{E_i}^2 = \sum \frac{O_i^2}{E_i} - n$$

In Formula 1,  $n$  represents the total number of observations. Formula 1 shows two ways of doing the arithmetic. Either way may be used. The left-hand form shows clearly that what matters is the total (squared) discrepancy between the expected frequencies and the actually observed frequencies; the right-hand form is well suited to calculations on a simple calculator. Using this formula on the data in the table above, we find that

$$\chi^2_{obs} = 1135.27 - 1052 = 83.27$$

When you have obtained the observed value of  $\chi^2$ , you compare it with the tabled threshold value for the level of significance at which you have chosen to work. The number of degrees of freedom,  $df$ , for the *goodness-of-fit test* is given by  $df = c - 1$ , where  $c$  is the number of categories observed. For the *association test*,

$$df = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

where the phrase "number of rows" means the number of categories of the row variable, and "number of columns" means the number of categories of the column variable.

For the table above,  $df = (3 - 1)(5 - 1) = 8$ . From a table of  $\chi^2$  thresholds, we find that the threshold is 15.507 for a level of significance  $\alpha = .05$  with  $df = 8$ . Since the observed value of  $\chi^2$  exceeds this threshold, our decision is to reject the null hypothesis of no association between the two categorical variables of class of borrower and branch of library system; i.e., we conclude that different classes of borrowers prefer different branches.

### The Minimum Acceptable Value for an Expected Frequency

In using the chi-square test of association, you need to observe caution with respect to the expected frequencies. For technical reasons, small expected frequencies cause difficulties in the interpretation of the chi-square coefficient. (Note that what you must be cautious about is the size of the *expected* frequencies. No special cautions need apply to the *observed* frequencies. The observed frequencies are the reality, against which you are checking your hypothesis that there is no association between the two categorical variables.)

Different statisticians advocate different minimum acceptable values for the expected frequencies, but almost all would accept a minimum of 10; many would accept a minimum of 5; and some would, in certain circumstances, accept a minimum of 1. In LIS 397.1 we will use 5 as the minimum acceptable value for the expected frequencies in the chi-square test of association.

What should you do if your calculations of the expected frequencies show one or more of them to be less than 5? The ideal solution is to gather enough additional observations so that the minimum expected frequency increases to at least 5. If that is impossible, then the process of "collapsing categories" may help. To collapse categories, you combine two rows (or columns); the entries in the new row (or column) are the sums of the entries in the original two rows (or columns).

In choosing the rows (or columns) to be merged, you should try to choose values of the variable concerned that make sense as a merged category. For example, if the book-circulation situation above had different data and happened to have an expected frequency of 3 in the "undergraduate" column, it would probably make the most sense to combine the "undergraduate" and "graduate" columns into a single merged category, "student".

### Note on the Chi-Square Test of Association for $2 \times 2$ Tables

When the chi-square test of association is applied to data in a  $2 \times 2$  table (i.e., to a situation in which each of the two categorical variables has only two values), you will sometimes see a special formula used for the calculation of the chi-square coefficient. I recommend that you *not* use this formula; but since some authors use it, you need to know what it looks like.

In a  $2 \times 2$  table there are four cells. We can use the labels A, B, C, and D for the observed frequencies in the four cells, as follows:

A	B
C	D

If we let  $n$  represent the total number of observations, then the chi-square test-statistic can be calculated as

Formula 2

$$\chi_{obs}^2 = \frac{n(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}$$

Though Formula 2 may seem simpler to use than Formula 1, it has a major defect. It does not let you see the expected frequencies, and therefore you will not be warned if there is an expected frequency less than 5. It is true that to use Formula 1 you will have to calculate the expected frequencies, and to do this, you will have to compute the row and column totals. But the various row and column totals are the contents of the parentheses in the denominator of Formula 2, and thus you will, in fact, have to compute them for either of the formulas.

It requires very little extra effort to do what I recommend: viz., to compute also the expected frequencies, and then check whether any of them is less than 5. If all the expected frequencies are 5 or more, you can go ahead and use Formula 1; if not, you might as well stop (unless there is some way to make further observations).