Social Studies 201
Notes for December 1, 2006

Chi-square Test for Independence

See text, Chapter 10, especially sections 10.1 (introduction) 10.2 (description of chi-square), and section 10.4 (test).

Appendix J contains a description of the chi-square distribution and the table on page 914 contains the critical values for the chi-square distribution.

These notes provide a short description of the chi-square test for independence in a cross-classification or crosstabulation table. The three tables referred to are on the last page of the notes in this file.

The purpose of the test is to examine whether there is any statistical relationship between the two variables of a cross-classification table. In the example, the question addressed is whether accumulated student debt is different for students from households of different income levels. Someone might claim, for example, that students from lower income backgrounds are more likely to finance their university studies through borrowing, as compared with students from households with higher income. The chi-square test provides a way to test this claim.

The chi-square test can be used whenever one wishes to test for the existence of a relationship between two variables in a cross-classification tables. The only limitation is that the sample size must be sufficiently large to have at least five expected cases in each cell of the table.

Example – test for a relationship between student debt and household income

The notes that follow use Tables 1 through 3 as an example to illustrate the test (see the last page of this file). The steps involved in the chi-square test are similar to those used in previous hypothesis tests, with a few variations. The test proceeds as follows.

1. **Hypotheses**. The chi-square test for independence of two variables begins with a null hypothesis of no relationship between the variables of a cross-classification table and an alternative hypothesis of some relationship between the two variables. In the case of the latter, the nature of the relationship is not specified – all the chi-square test allows us to determine is whether there is evidence of a relationship between

the two variables or not. For testing whether there is a relation between the two variables in Table 1, the null and alternative hypothesis are

Null hypothesis	H_0: No relationship between debt and income
Alternative hypothesis	H_1: Relationship between debt and income

As with other hypothesis tests, begin by assuming that the null hypothesis is true. It is only at the end of the test that the researcher decides either to reject this null hypothesis, H_0, and accept the alternative hypothesis, H_1. Alternatively, the evidence may be such that it is not possible reject the null hypothesis, H_0.

2. **Expected values**. This stage of the test is somewhat different than in earlier hypothesis tests. For a chi-square test, it is necessary to determine the expected number of cases in each cell of the cross-classification table, under the assumption that H_0 is correct. In this example, the expected values are the number of respondents that would be expected in each category, assuming no relation between debt and income. In Table 2, these are the numbers labelled "Expected Count." While the SPSS program calculates these, it is worth considering how these are obtained.

   Table 2 has nine cells – three categories of debt and three categories of income, so there are $3 \times 3 = 9$ cells. For the top left cell of zero debt (event A) and lowest income (event B), the probability that a randomly selected individual is in this category is

$$P(A \text{ and } B) = P(A)P(B).$$

   As discussed in Chapter 6, when the two events are independent of each other, the joint probability of the two events is the product of the individual probabilities (see text, p. 329). Since the null hypothesis of independence is assumed when working through the test, this probability is

$$P(A \text{ and } B) = P(A)P(B) = \frac{160}{549} \times \frac{371}{549}.$$

Note that this probability does not take into account the observed number of cases in this cell of the table, it uses only the row, column, and overall total number of cases.

The number of cases a researcher would expect in any cell of the table is the probability of finding an individual in this cell, times the total number of cases n. If the expected number of cases is given the symbol (E), the expected number of cases in any cell of the table can be obtained by using the formula:

$$E(A \text{ and } B) = P(A \text{ and } B) \times n$$

where n is the sample size. For this first cell, the expected number of cases would thus be

$$\begin{aligned} E(A \text{ and } B) &= \frac{160}{549} \times \frac{371}{549} \times 549 \\ &= \frac{160 \times 371}{549} \\ &= 108.1 \end{aligned}$$

This same procedure can be followed for each of the cells. Fortunately, SPSS is programmed to produce these expected values, and the observed number of cases ("Count") and the expected number of cases or expected count is shown for each cell of the table. In Table 1, these are in the second row of each cell of the table.

Where the observed count and expected count are reasonably similar, this provides evidence that tends ot support the null hypothesis. For example, in the middle column of the table, the first cell has 169 cases observed and 170.3 expected cases, working on the assumption H_0 is true. In contrast, the observed and expected counts are further apart for each of the cells in the first column. For example, if there is no relation between income and debt, there would be only 17.8 students in the lower left cell, that is, only 17.8 of lowest income and debt of ten thousand dollars plus. But, from the sample, the observed number of cases is 23, so this provides evidence that there are more low income students with high debt than what would be expected if H_0 is true.

Assumption. As noted earlier, the only restriction on the use of the chi-square test is that there be five or more expected cases per cell of

the table. Note that it is not the observed number of cases that must exceed five but the expected number of cases that must be five or more – it may be that a cell has no cases in it, but so long as the expected number of cases is five or more, that satisfies the assumption for use of the test.

In Table 2, all the expected counts are much greater than five, so this assumption is met. For cross-classsification tables with one or two cells having less than five expected cases per cell, the chi-square test may be used, but if several of the cells have less than five expected cases, it may be advisable to use a different type of test. See the discussion on p. 741 of the text for guidelines.

3. **Chi-square statistic**. The next step is to calculate the chi-square statistic. This is given the symbol χ, a Greek letter, with the squared value being χ^2. There is a chi-square entry for each cell of the table and this entry is

$$\frac{(O-E)^2}{E}$$

where O is the observed count and E the expected count for the cell. The χ^2 value for the whole table is the sum of these entries. That is,

$$\chi^2 = \sum \frac{(O-E)^2}{E},$$

the sum, across all cells of the table, of the squares of the observed minus expected values, divided by the expected values.

For Table 2, the first entry is

$$\frac{(97-108.1)^2}{108.1} = \frac{-11.1^2}{108.1} = 1.140.$$

To obtain all the entries for the χ^2 statistic, a similar calculation for each of the cells of Table 2 is used. These are the entries in Table 4, with this table set up to correspond to the format of Table 2.

The sum of the chi-square values in Table 4 is $\chi^2 = 9.272$, consistent with the Pearson chi-square value listed in Table 3, from the SPSS printout. The question is whether this is a large chi-square value or a chi-square value that is not large enough to reject the null hypothesis.

Table 4: Chi-square calculations for Table 2

1.140	0.010	1.660
1.021	0.019	1.316
1.519	0.036	2.529

4. χ^2 **distribution**. Section 10.2 (p. 705) and Appendix J (p. 913) give a description of the chi-square distribution. This distribution is asymmetrical – it can be imagined as fixing the left part of a normal distribution and pulling the right end to the right. The χ^2 values are measured on the horizontal axis, starting at a value of 0 on the left and going to very large values on the right.

   If the null hypothesis is correct, so there is no relation between the two variables, then $O = E$ for each cell of the table and $\chi^2 = 0$. If the alternative hypothesis is correct, then $O \neq E$, each entry into the chi-square calculation is large and the overall chi-square value is large.

   The critical region for rejection of the null hypothesis is only at the right end of the distribution. As in earlier hypotheses tests, a significance level of α is selected. The set of χ^2 values that leave exactly α in the right tail is the critical region. If the χ^2 value calculated in item 3 is in this critical region, H_0 is rejected; in contrast, if the calculated χ^2 value is not in this critical region, then the null hypothesis cannot be rejected at the α significance level.

5. **Degrees of freedom (df)**. One additional consideration required for a χ^2 test is to obtain the degrees of freedom (df). For a test of independence in a cross-classification table, the degrees of freedom is the number of rows minus one, times the number of columns minus one. This represents the number of cells in a cross-classification table that can be freely assigned, with the counts in the remaining cells being constrained so the correct row and column totals result.

   Table 2 is considered a 3×3 table, with three rows and three columns of data. This table has $(3-1) \times (3-1) = 2 \times 2 = 4$ degrees of freedom. This is the degrees of freedom stated in the SPSS printout of Table 3.

6. **Significance level and critical region**. From the table of the χ^2 distribution in Appendix J, the critical χ^2 value can be determined once the significance level is selected and the number of degrees of freedom is known. The value in the body of the χ^2 table represents the point on the horizontal axis where the critical region begins, for a given significance level α and degrees of freedom df.

   For the example of income and debt, select a significance level of $\alpha = 0.10$. Since Table 2 is a 3×3 cross-classification table, there are 4 degrees of freedom. For $\alpha = 0.10$ and 4 df, the critical χ^2 value is 7.779, that is, 0.10 of the area under this χ^2 distribution is to the right of $\chi^2 = 8.496$ and the other 0.90 of the area lies to the left of this. If the sample yields a value greater than χ^2 of 7.779, reject H_0, but if the sample has a value less than 7.779, do not reject H_0.

7. **Conclusion**. From the SPSS printout, $\chi^2 = 9.250 > 7.779$, and the chi-square statistic is in the critical region. At the 0.10 level of significance, the null hypothesis is rejected and the alternative hypothesis accepted. This conclusion is that there is some relationship between debt and income. As a result, the data from Tables 1 through 3 demonstrates that there is some relationship between household income of students and the level of student debt.

   By itself, the chi-square test for independence does not indicate thee nature of this relationship – an analysis of the way that income and debt are related requires examining the table in more detail. One way of doing this is to consider the difference between the counts and expected counts in the table.

   In Table 2, for the middle income group, with income of \$40-80,000, the counts (observed values or O) and expected counts (expected values or E) are similar. This means that for the distribution of debt, there is little difference between the sample as a whole and students of middle income. In contrast, for students from lower income backgrounds (less that \$40,000), there are fewer with no debt (97) than what would be expected (108) assuming no relationship between income and debt. And there are more with debt, and high debt levels, than expected. For students from the highest income category, there are more students with no debt (105) than expected (93), and fewer with debt than expected.

What these results point to is that students from lower income households tend to have student debt while students from higher income households tend to have limited debt. The connection between debt and household income is not strong but the data in Tables 1 through 3 provide sufficient evidence to demonstrate that lower income students may be more burdened with debt than are those from higher income households.

Summary of chi-square method

The following summarizes the method of performing the chi-square test.

H_0: No relationship between variables	$O = E$	small χ^2 value	Do not reject H_0
H_1: Relationship between variables	$O \neq E$	large χ^2 value	Reject H_0, accept H_1

The test begins by assuming the null hypothesis of no relationship between the two variables of the cross-classification table. Under this assumption, the observed and expected values are similar so the χ^2 value will be small. If the χ^2 value reported or calculated is small, then the conclusion is that the null hypothesis of no relationship between the variables cannot be rejected.

By contrast, if the observed and expected values are quite different from each other, then this produces a large χ^2 value. If this value is large enough to be in the critical region, then the null hypothesis is rejected. In this case, there is evidence that the alternative hypothesis is correct and the researcher concludes there is some relationship between the two variables.

Last edited December 1, 2006.

Social Studies 201 – December 1, 2006
Chi-square test for a cross-classification table

Table 1

D1 amount of debt * I household income Crosstabulation

			I household income			
			1 under $40,000	2 $40-80,000	3 $80,000 plus	Total
D1 amount of debt	1 none	Count	97	169	105	371
		% within I household income	60.6%	67.1%	76.6%	67.6%
	2 under $10,000	Count	40	54	23	117
		% within I household income	25.0%	21.4%	16.8%	21.3%
	3 $10,000 plu	Count	23	29	9	61
		% within I household income	14.4%	11.5%	6.6%	11.1%
Total		Count	160	252	137	549
		% within I household income	100.0%	100.0%	100.0%	100.0%

Table 2

D1 amount of debt * I household income Crosstabulation

			I household income			
			1 under $40,000	2 $40-80,000	3 $80,000 plus	Total
D1 amount of debt	1 none	Count	97	169	105	371
		Expected Count	108.1	170.3	92.6	371.0
	2 under $10,000	Count	40	54	23	117
		Expected Count	34.1	53.7	29.2	117.0
	3 $10,000 plu	Count	23	29	9	61
		Expected Count	17.8	28.0	15.2	61.0
Total		Count	160	252	137	549
		Expected Count	160.0	252.0	137.0	549.0

Table 3

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
Pearson Chi-Square	9.272[a]	4	.055	
Likelihood Ratio	9.603	4	.048	
Linear-by-Linear Association	9.056	1	.003	
McNemar Test				.[b]
N of Valid Cases	549			

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 15.22.

b. Both variables must have identical values of categories.