

Crosstabulation & Chi Square

Robert S Michael

Chi-square as an Index of Association

After examining the distribution of each of the variables, the researcher's next task is to look for relationships among two or more of the variables. Some of the tools that may be used include correlation and regression, or derivatives such as the *t*-test, analysis of variance, and contingency table (crosstabulation) analysis. The type of analysis chosen depends on the research design, characteristics of the variables, shape of the distributions, level of measurement, and whether the assumptions required for a particular statistical test are met.

A *crosstabulation* is a joint frequency distribution of cases based on two or more categorical variables. Displaying a distribution of cases by their values on two or more variables is known as contingency table analysis and is one of the more commonly used analytic methods in the social sciences. The joint frequency distribution can be analyzed with the *chi-square* statistic (χ^2) to determine whether the variables are statistically independent or if they are associated. If a dependency between variables does exist, then other indicators of association, such as Cramer's *V*, gamma, Sommer's *d*, and so forth, can be used to describe the degree which the values of one variable predict or vary with those of the other variable. More advanced techniques such as log-linear models and multinomial regression can be used to clarify the relationships contained in contingency tables.

Considerations:

Type of variables. Are the variables of interest continuous or discrete (e.g., categorical)? Categorical variables contain integer values that indicate membership in one of several possible categories. The range of possible values for such variables is limited, and whenever the range of possible values is relatively circumscribed, the distribution is unlikely to approach that of the Gaussian distribution. Continuous variables, in contrast, have a much wider range, no limiting categories, and have the potential to approximate the Gaussian distribution, provided their range is not artificially truncated. Whenever you encounter a categorical or a nominal, discrete variable, be aware that the assumption of normality is likely violated.

Shape of the distribution. Categorical variables often have such a small number of possible values that one cannot even pretend that the assumption of normality is approximated. Consider for example, the possible values for sex, grade levels, and so forth. Statistical tests that require the assumption of normality cannot be used to analyze such data. (Of course, a statistical program such as SPSS will process the numbers without complaint and yield results that may appear to be interpretable — but only to those who ignore the necessity of examining the distributions of each variable first, and who fail to check whether the assumptions were met). Because the assumption of normality is a requirement for the *t*-test, analysis of variance, correlation and regression, these procedures cannot be used to analyze count data.

Level of measurement. The count data for categorical variables that appear in contingency tables is nominal level. Nominal level measurement is encountered with enumerative or classificatory data. Examples include classification of individuals into five income brackets; the classification of schools into eight categories by the Census Bureau; classification of performance into four letter grade categories; and an art historian who classifies paintings into one of k categories.

The chi-square test of statistical significance, first developed by Karl Pearson, assumes that both variables are measured at the nominal level. To be sure, chi-square may also be used with tables containing variables measured at a higher level; however, the statistic is calculated *as if* the variables were measured only at the nominal level. This means that any information regarding the order of, or distances between, categories is ignored.

Assumptions:

The assumptions for chi-square include:

1. Random sampling is not required, *provided the sample is not biased*. However, the best way to insure the sample is not biased is random selection.
2. Independent observations. A critical assumption for chi-square is independence of observations. One person's response should tell us nothing about another person's response. Observations are independent if the sampling of one observation does not affect the choice of the second observation. (In contrast, consider an example in which the observations are not independent. A researcher wishes to estimate to what extent students in a school engage in cheating on tests and homework. The researcher randomly chooses one student to interview. At the completion of the interview the researcher asks the student for the name of a friend so that the friend can be interviewed, too).
3. Mutually exclusive row and column variable categories that include all observations. The chi-square test of association cannot be conducted when categories overlap or do not include all of the observations.
4. Large expected frequencies. The chi-square test is based on an approximation that works best when the expected frequencies are fairly large. No expected frequency should be less than 1 and no more than 20% of the expected frequencies should be less than 5.

Hypotheses

The null hypothesis is the k classifications are independent (i.e., no relationship between classifications). The alternative hypothesis is that the k classifications are dependent (i.e., that a relationship or dependency exists).

Examples:

Example 1. Generations of students at Washington School have taken field trips at both the elementary and secondary levels. The principal wonders if parents still support field trips for children at either level. Five hundred letters were mailed to parents, asking them to indicate either approval or disapproval; 100 parents returned the response postcard. Each postcard indicated whether the parents' children were currently enrolled in elementary or high school, and the parents' approval or disapproval of field trips. Table 1 contains the data.

Table 1: Parents' Opinion about Field Trips (Observed Frequencies)

	Approve	Disapprove	No Opinion	Row Totals
Elementary	28	14	5	47
High School	19	28	6	53
Column Totals	47	42	11	100

Example 2. A teacher wondered if students and teachers have differing views on grade inflation. At a faculty meeting the teacher asked whether grade inflation is a problem in this high school and the same question was posed to students who were waiting in a bus line. The responses appear in Table 2:

Table 2: Is Grade Inflation a Serious Problem? (Observed Frequencies)

	Yes	No	Row Totals
Teachers	24	6	30
Students	8	12	20
Column Totals	32	18	50

Example 3. One of the questions in the 1991 General Social Survey attempted to determine whether exposure to television weakens or strengthens confidence in the (presumably) television press. The results are present in Table 3:

Table 3: Confidence in Television Press

<i>As far as the people running the press, would you have . . .</i>	Average Hours of Daily TV Watching			Row Total
	0-1 hours	2-4 hours	5 or more	
A good deal of confidence	276	41	17	334
Only some confidence	196	174	47	417
Hardly any confidence	130	97	15	242
Column Total	602	312	79	993

Example 4. Researchers conducted a study of 58 children admitted to the child psychiatry ward of the University of Iowa for aggressive behavior.¹ The purpose of the study was to determine the influence of social class, sex, and age on the clinical characteristics of the children. A small portion of their data is shown in Table 4, which lists the numbers of children exhibiting antisocial behavior for two categories of social class. Classes I to III represent children from middle class families. Those in Classes IV and V were from poor families.

Table 4: Is Social Class and Antisocial Behavior? (Observed Frequencies)

	Yes	No	Row Totals
Classes I-III	24	4	28
Classes IV-V	17	13	30
Column Totals	41	17	58

Worked Example

The field trip opinion data is used to illustrate the computation of chi-square.

If the assumptions required for chi-square are met, the first step in computing the chi-square test of association (also referred to as the “chi-square test of independence”), is to compute the *expected* frequency for each cell, under the assumption that the null hypothesis is true. To calculate the expected frequency for the first cell (elementary, approve) of the field trip opinion poll, first calculate the proportion of parents that approve without considering whether

1. Behar, D., & Stewart, M. A. (1984). Aggressive conduct disorder: The influence of social class, sex, and age on the clinical picture. *Journal of Child Psychology and Psychiatry*, 25.

their children are elementary or high school. The table shows that of the 100 postcards returned, 47 approved. Therefore, 47/100, or 47% approve.

If the null hypothesis is true, the expected frequency for the first cell equals the product of the number of people in the elementary condition (47) and the proportion of parents approving (47/100). This is equal to $47 \times 47 / 100 = 22.09$. Thus, the expected frequency for this cell is 22.09, compared with the observed frequency of 28.

The general formula for each cell's expected frequency is: $E_{ij} = \frac{T_i \times T_j}{N}$, where:

- E_{ij} is the expected frequency for the cell in the i th row and the j th column.
- T_i is the total number of counts in the i th row.
- T_j is the total number of counts in the j th column.
- N is the total number of counts in the table.

The calculations are shown in Table 5:

Table 5: Parents' Opinion about Field Trips (Formula for *Expected* Frequencies)

	Approve	Disapprove	No Opinion	Row Totals
Elementary	$= \frac{47 \times 47}{100}$	$= \frac{47 \times 42}{100}$	$= \frac{47 \times 11}{100}$	47
High School	$= \frac{53 \times 47}{100}$	$= \frac{53 \times 42}{100}$	$= \frac{53 \times 11}{100}$	53
Column Totals	47	42	11	100

The second step, for each cell, after the expected cell frequencies are computed, is to subtract the observed frequencies from the expected frequencies, square the difference and then divide by the expected frequency. Then sum the values from each cell. The formula for chi-square is: $\chi^2 = \sum \frac{(E - O)^2}{E}$. The corresponding values for each cell are shown in Tble 6:

Table 6: Chi-Square values for each Cell

	Approve	Disapprove	No Opinion	Row Totals
Elementary	$= \frac{(22.09 - 28)^2}{22.09}$	$= \frac{(19.74 - 14)^2}{19.74}$	$= \frac{(5.17 - 5)^2}{5.17}$	47
High School	$= \frac{(24.91 - 19)^2}{24.91}$	$= \frac{(22.26 - 28)^2}{22.26}$	$= \frac{(5.83 - 6)^2}{5.83}$	53
Column Totals	47	42	11	100

For this example, chi-square equals the following:

$$\frac{(22.09 - 28)^2}{22.09} + \frac{(19.74 - 14)^2}{19.74} + \frac{(5.17 - 5)^2}{5.17} + \frac{(24.91 - 19)^2}{24.91} + \frac{(22.26 - 28)^2}{22.26} + \frac{(5.83 - 6)^2}{5.83}$$

which sums to 6.14. In the parlence of hypothesis testing, this is known as the “calculated” or “obtained” value or sometimes, the “p” value, in contrast to the “tabled” value — located in the appendix of critical values in statistical textbooks.

Degrees of Freedom. The next step is to calculate the degrees of freedom. The term “degrees of freedom” is used to describe the number of values in the final calculation of a statistic that are free to vary. It is a function of both the number of variables and number of observations. In general, the degrees of freedom is equal to the number of independent observations minus the number of parameters estimated as intermediate steps in the estimation (based on the sample) of the parameter itself. (Perhaps the most useful discussion I have seen can be found at <http://seamonkey.ed.asu.edu/~alex/computer/sas/df.html>).

For χ^2 , the degrees of freedom are equal to $(r-1)(c-1)$, where r is the number of rows and c is the number of columns. In the field trip example, $r = 2$ and $c = 3$, so $df = (2-1)(3-1) = 2$.

Choosing a significance level. The risk of making an incorrect decision is an integral part of hypothesis testing. Simply following the steps prescribed for hypothesis testing does not guarantee that the correct decision will be made. We cannot know with certainty whether any one particular sample mirrors the true state of affairs that exists in the population or not. Thus, before a researcher tests the null hypothesis, the researcher must determine how much risk of making an incorrect decision is acceptable. “How much risk of an incorrect decision am I willing to accept? One chance out of a hundred? Five chances out of a hundred? Ten? Twenty?”

The researcher decides, before testing, on the cutoff value. The convention, which the researcher is free to ignore, is 5 times out of a hundred. This value is known as the “significance level,” or “alpha” (α). After the researcher decides on the alpha level, the researcher looks at the table of critical values. With alpha set at 0.05, the researcher knows which column of the table to use. If the researcher chooses to set alpha at 0.01, then a different column in the table is used.

Critical values of χ^2 . The chi-square table in Hopkins, Hopkins, & Glass (1996) is on page 352. Note that the symbol used in this table for degrees of freedom is ν . If we enter the table on the row for two degrees of freedom and move to the column labelled $\alpha = 0.05$, we see the value of 5.99 in the intersecting cell. This value is known as the “tabled” χ^2 value with two degrees of freedom and $\alpha = 0.05$.

Logic of Hypothesis Testing

The last step is to make a judgment about the null hypothesis. The χ^2 statistic is large when some of the cells have large discrepancies between the observed and expected frequencies. Thus we reject the null hypothesis¹ when the χ^2 statistic is large. In contrast, a small calculated χ^2 value does not provide evidence for rejecting the null hypothesis. The question we are asking here: Is the calculated chi-square value of 6.14 sufficiently large (with $df = 2$ and $\alpha = 0.05$) to provide the evidence we need to reject the null hypothesis?

Suppose that the χ^2 statistic we calculated is so large that the probability of getting a χ^2 statistic at least as large or extreme (i.e., somewhere out in the tail of the chi-square distribution) as our calculated χ^2 statistic is very small, if the null hypothesis is true. If this is the case, the results from our sample are very unlikely if the null hypothesis is true, and so we reject the null hypothesis.

To test the null hypothesis we need to find the probability of obtaining a χ^2 statistic at least as extreme as the calculated χ^2 statistic from our sample, assuming that the null hypothesis is true. We use the critical values found in the χ^2 table to find the approximate probability.

1. The idea of “falsifiability” of the null hypothesis was introduced by the philosopher Karl Popper. He noted that we cannot affirm the truthfulness of a hypothesis beyond all doubt, but we can conclusively disprove a hypothesis.

In our example the calculated χ^2 value is 6.14. With 2 degrees of freedom and alpha set at 0.05, this calculated value is larger than the tabled value (which is 5.99).

Recall that the tabled value represents the following statement: For 2 degrees of freedom and with alpha = 0.05, a calculated χ^2 value of 5.99 or larger is sufficiently extreme that we can say the probability of obtaining a value this large, due to chance variation, is only 5 out of 100. Because this is unlikely, we have the evidence we need to reject the null hypothesis (the hypothesis of no dependence/relationship/association) among variables, and, instead, accept the alternative hypothesis; namely, dependence among variables that is unlikely due to chance variation.

Hypothesis testing errors. Whenever we make a decision based on a hypothesis test, we can never know whether or decision is correct. There are two kinds of mistakes we can make:

- 1 we can fail to accept the null hypothesis when it is indeed true (Type I error), or
- 2 we can accept the null hypothesis when it is indeed false (Type II error)

The best we can do is to reduce the chance of making *either* of these errors. If the null hypothesis is true (i.e., it represents the true state of affairs in the population), the significance level (alpha) is the probability of making a Type I error. Because the researcher decides the significance level, we control the probability of making a Type I error.

The primary methods for controlling the probability of making a Type II error is to select an appropriate sample size. The probability of a Type II error decreases as the sample size increases. At first glance the best strategy might appear to be to obtain the largest sample that is possible. However, time and money are always limitations. We do not want a sample size that is larger than the minimum necessary for a small probability of a Type II error. How can we determine the minimum sample size needed?

Spss code for field trip example

```
* File: d:\y590\sps\fieldtrip-sps .
set printback=on header=on messages=on error=on results=on.
Title 'Field Trip Opinions' .

data list / level 1-1 vote 4-4 wgt 7-8 .
begin data
1 1 28
1 2 14
1 3 5
2 1 19
2 2 28
2 3 6
end data .

variable labels level 'Elem or HS'
                / vote 'approve or not'
                / wgt 'Cell count' .
value labels   level 1 'Elementary'
                2 'High School'
                / vote 1 'Approve'
                2 'Disapprove'
                3 'No Opinion' .

weight by wgt .
crosstabs tables = level by vote
                / cells = count expected resid
                / statistics = chisq .
```

Such data are often reported by listing the number or percent of individuals that have each value. Suppose a kindergarten program at a school rates the reading status of 69 students according to the following index:

- 0 = does not know colors or shapes
- 1 = recognizes colors and shapes
- 2 = recognizes individual letters
- 3 = demonstrates individual sounds
- 4 = demonstrates blends
- 5 = reads 90% of words on sight word list

Such data, often referred to as *count* data or as *frequency* data, are usually reported in one of two ways: either the number (frequency) or percent of subjects that have each value. The frequencies for the reading status of kindergarten students are shown in Table 1.

Table 7: Distribution of Reading Status Rating

Rating	Frequency
0	34
1	10
2	17
3	6
4	1
5	1

Thirty-four students were rated as unable to name colors or shapes, 10 recognized colors and shapes, and so forth.

Frequency data such as this is often used to test hypotheses. The kinds of hypotheses that are test include: (a) two or more population proportions (or counts), (b) association between two variables, and (c) differences between paired proportions.

Hypotheses about two or more populations

Suppose a principal believes that half of students entering kindergarten would have reading status ratings equal to zero. Further, the principal expects that one-tenth of students will have each of the other reading status ratings. How can the data in Table 1 be used to either support or refute the principal's expectations? First, we state the principal's expectations explicitly in terms of proportions (see table 2). Then we can state the null hypothesis as the sum of the proportions (note that they must sum to 1)

Table 8: Hypothesized Proportions: Reading Status Rating

Rating	Expected Proportion
0	.50
1	.10
2	.10
3	.10
4	.10
5	.10

For convenience, we abbreviate the expected population proportions with the letter “p” and a subscript that corresponds to the rating. The null hypothesis that we wish to test is:

$$H_0 : p_0 = 0.50, p_1 = 0.10, p_2 = 0.10, p_3 = 0.10, p_4 = 0.10, p_5 = 0.10, p_6 = 0.10$$

The alternative hypothesis is:

H_A : At least one of the expected population proportions is incorrect.

The appropriate procedure for testing this hypothesis is the chi-square (χ^2) test of hypothesized proportions. This test is appropriate when frequencies or counts are available for two or more mutually exclusive categories that include all of the data.

The hypothesized, or expected, population proportions are specified by the investigator. They are not the result of any statistical method. Sometimes the expected population proportions can be based on a theoretical model. In all cases the sum of the expected proportions must equal 1 ($p_0 + p_1 + \dots + p_k = 1$), because all of the categories are mutually exclusive and include all the data. None of the hypothesized population proportions can be zero (if so, why have such a category?).

To test the null hypothesis, we first calculate the *expected frequencies*; that is, the frequencies we would expect if all the hypothesized population proportions were, indeed, correct. Let n denote the total number of observations and p_{i0} (i.e., “p sub i nought”) denote the hypothesized population proportion for category i . Then the expected frequency for category i is denoted by E_i and calculated according to the formula:

$$E_i = n \cdot p_{i0}$$

For the reading status rating data, the expected frequencies are

$$E_0 = 69 \cdot 0.50 = 34.5$$

$$E_1 = E_2 = E_3 = E_4 = E_5 = 69 \times 0.10 = 6.9$$

The sum of the expected frequencies is always equal to n , taking rounding error into account. For the reading status data, the expected frequencies are $34.5 + 6.9 + 6.9 + 6.9 + 6.9 + 6.9 = 69$.

The *observed frequencies* are the actual frequencies obtained from the data and are denoted O_i . The observed frequencies for the data in table 1 is represented as:

$$O_0 = 34$$

$$O_1 = 10$$

$$O_2 = 17$$

$$O_3 = 6$$

$$O_4 = 1$$

$$O_5 = 1$$

To test the null hypothesis, we use the χ^2 statistic that takes into account the discrepancies between the observed and expected frequencies. The formula is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$