The Chi Square Test

Diana Mindrila, Ph.D.
Phoebe Balentyne, M.Ed.

Based on Chapter 23 of The Basic Practice of Statistics (6th ed.)

Concepts:
- Two-Way Tables
- The Problem of Multiple Comparisons
- Expected Counts in Two-Way Tables
- The Chi-Square Test Statistic
- Cell Counts Required for the Chi-Square Test
- Uses of the Chi-Square Test
- The Chi-Square Distributions

Objectives:
- Construct and interpret two-way tables.
- Describe the problem of multiple comparisons.
- Calculate expected counts in two-way tables.
- Describe the chi-square test statistic.
- Describe the cell counts required for the chi-square test.
- Describe uses of the chi-square test.
- Describe the chi-square distributions.
- Perform a chi-square goodness of fit test.

References:
Moore, D. S., Notz, W. I, & Flinger, M. A. (2013). *The basic practice of statistics* (6th ed.). New York, NY: W. H. Freeman and Company.

Example

Question: Is there an association between students' preference for online or face-to-face instruction and their education level?

Survey Items:
Are you an undergraduate or graduate student?
- o Undergraduate
- o Graduate

Which method of instructional delivery do you prefer?
- o Face-to-face
- o Online

- ▪ The information gathered from this survey must be organized in a data file within the statistical software.
- ▪ For each question, a categorical (or nominal) variable is created.

Data File:

Student ID	Educational Level	Instructional Preference
1	Undergraduate	Online
2	Undergraduate	Face to Face
3	Undergraduate	Face to Face
4	Graduate	Online

Cross-tabulation

- *T* tests can be used to determine whether there are significant differences between undergraduate and graduate students or between face-to-face and online instruction.
- However, these procedures would be conducted separately by education level and then by instructional preference. This would not reveal an association between the two variables.
- Correlation analysis is typically used to measure the association between variables, but correlation can only be used with quantitative variables.
- In order to compare categorical variables, the data can be summarized into a table, which lists the options for one variable as the rows and the options for the other variable as the columns. This is called a crosstab because two variables are being tabulated at the same time, and the frequency, or the percentage of individuals in each subcategory, are being counted.

Cross-tabulation of the two qualitative (nominal) variables:

Educational Level			
Instructional Preference	**Undergraduate**	**Graduate**	**Total**
Face to face	?	?	?
Online	?	?	?
Total	60 (60%)	40 (40%)	100

- In this example, instructional preferences are listed as the rows and education levels are listed as the columns.
- The next step is to obtain the frequencies for each category, which can be done using statistical software, especially for a very large sample.
- Although a crosstab is a helpful descriptive statistic, it is also important to be able to determine if there is an association between the two variables and whether or not it is statistically significant.

Chi-Square Test

- ❖ To determine whether the association between two qualitative variables is statistically significant, researchers must conduct a test of significance called the **<u>Chi-Square Test</u>**. There are five steps to conduct this test.

Step 1: Formulate the hypotheses

- ❖ Null Hypothesis:

  H_0: There is no significant association between students' educational level and their preference for online or face-to-face instruction.

  or

  H_0: There is no difference in the distribution of instructional preferences between undergraduate and graduate students.

- ▪ If there is no association between the two variables, the individuals would be uniformly distributed across the cells of the table.

- ▪ The alternative hypothesis for a chi-square test is always two-sided. (It is technically multi-sided because the differences may occur in both directions in each cell of the table).

- ❖ Alternative Hypothesis:

  H_a: There is a significant association between students' educational level and their preference for online or face-to-face instruction.

  or

  H_a: There is a significant difference in the distribution of instructional preferences between undergraduate and graduate students.

Step 2: Specify the expected values for each cell of the table (when the null hypothesis is true)

- The **expected values** specify what the values of each cell of the table would be if there was no association between the two variables.
- The formula for computing the expected values requires the sample size, the row totals, and the column totals.

Expected Counts

The expected count in any cell of a two-way table when H_0 is true is:

$$\text{expected count} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}}$$

Total Values:

Instructional Preference	Educational Level		
	Undergraduate	Graduate	Total
Face to face	?	?	55 (55%)
Online	?	?	45 (45%)
Total	60 (60%)	40 (40%)	100 (100%)

Expected Values:

Instructional Preference	Educational Level		
	Undergraduate	Graduate	Total
Online	33	22	55 (55%)
Face to face	27	18	45 (45%)
Total	60 (60%)	40 (40%)	100 (100%)

Step 3: To see if the data give convincing evidence against the null hypothesis, compare the observed counts from the sample with the expected counts, assuming H_0 is true.

- The **observed values** are the actual counts computed from the sample.
- Statistical software will compute both the expected and observed counts for each cell when conducting a chi-square test.
- The image below shows the table that SPSS creates for the two variables. In each cell, the expected and observed value is present.

Observed Counts:

	Educational Level		
Instructional Preference	**Undergraduate**	**Graduate**	**Total**
Online	20	35	55 (55%)
Face to face	40	5	45 (45%)
Total	60 (60%)	40 (40%)	100 (100%)

Statistical Software:

Course Preference * Student Educational Level Crosstabulation

			Student Educational Level		Total
			Undergraduate	Graduate	
Course Preference	Online	Count	20	35	55
		Expected Count	33.0	22.0	55.0
	Face to Face	Count	40	5	45
		Expected Count	27.0	18.0	45.0
Total		Count	60	40	100
		Expected Count	60.0	40.0	100.0

Step 4: Compute the test statistic

- The **chi-square statistic** compares the observed values to the expected values.
- This test statistic is used to determine whether the difference between the observed and expected values is statistically significant.

The **chi-square statistic** is a measure of how far the observed counts are from the expected counts. The formula for the statistic is:

$$x^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

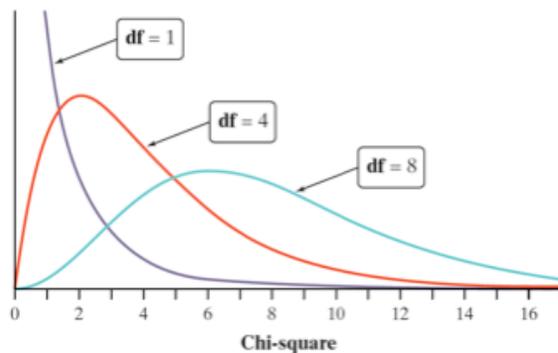where the sum is over all possible values of the categorical variable.

The Chi-Square Distributions

<div style="text-align: center">

The Chi-Square Distributions

</div>

The **chi-square distributions** are a family of distributions that take only positive values and are skewed to the right. A particular chi-square distribution is specified by giving its **degrees of freedom.**
The chi-square test for a two-way table with r rows and c columns uses critical values from the chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom. The P-value is the area under the density curve of this chi-square distribution to the right of the value of the test statistic.

df=(#rows-1)*(#columns-1)

Example:
df=(2-1)*(2-1)=1

- The image above shows that the distribution of the chi-square statistic starts at zero and can only have positive values.
- The shape of the distribution is much different than the t or z statistic and is skewed to the right.
- The shape of the distribution changes as the degrees of freedom increases.

Chi-Square Test – Test Statistic

Example:

Instructional Preference		Educational Level		
		Undergraduate	Graduate	Total
Online	Observed	20	35	55 (55%)
	Expected	33	22	
Face to face	Observed	40	5	45 (45%)
	Expected	27	18	
Total		60 (60%)	40 (40%)	100 (100%)

$$x^2 = \sum \frac{(\text{Observed - Expected})^2}{\text{Expected}}$$

$$x^2 = \frac{(20-33)^2}{33} + \frac{(35-22)^2}{22} + \frac{(40-27)^2}{27} + \frac{(5-18)^2}{18} = 28.451$$

- The above example shows the observed and expected values for the example problem.
- If these values are entered into the formula for the chi-square tests statistic, the value obtained is 28.451.
- Is this value high enough to reject the null hypothesis?

Step 5: Decide if chi-square is statistically significant

- The final step of the chi-square test of significance is to determine if the value of the chi-square test statistic is large enough to reject the null hypothesis.
- Statistical software makes this determination much easier.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	28.451[a]	1	.000		
Continuity Correction[b]	26.305	1	.000		
Likelihood Ratio	31.104	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	28.167	1	.000		
N of Valid Cases	100				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 18.00.
b. Computed only for a 2x2 table

- For the purpose of this analysis, only the Pearson Chi-Square statistic is needed.
- The p-value for the chi-square statistic is .000, which is smaller than the alpha level of .05.
- ❖ Therefore, there is enough evidence to reject the null hypothesis.
- ❖ Conclusion: Evidence from the sample shows that there is a significant difference in the distribution of instructional preference between undergraduate and graduate students.

- A table in a statistics textbook could also be used to conduct the test by hand:
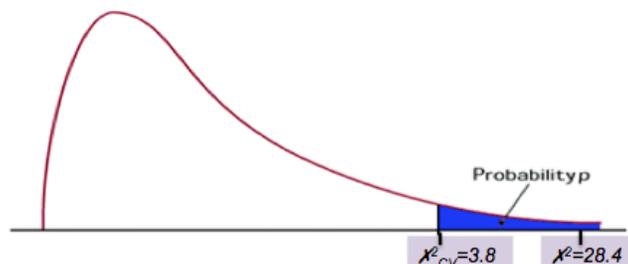
Table D (p. 680):

α=.05
Chi-square$_{CV}$=3.84
df=1

Chi-square=28.451
Chi-square$_{CV}$=3.84
Chi-square ≥ Chi-square$_{CV}$ => reject Ho and accept Ha:

Table D. Chi-square distribution critical values

		p	
df	.10	05	.25
1	2.71	3.84	5.02
2	4.61	5.99	7.38

Conclusion: Evidence from the sample shows that there is a significant difference in the distribution of instructional preferences between undergraduate and graduate students.

Probability p

x^2_{CV}=3.8 x^2=28.4

- The chi-square value of 28.451 is much larger than the critical value of 3.84, so the null hypothesis can be rejected.

The Chi-Square Test Interpretation

The chi-square test is an overall test for detecting relationships between two categorical variables. If the test is significant, it is important to look at the data to learn the nature of the relationship. There are three ways to look at the data:

1) **Compare selected percents:** which cells occur in very different percentages than the other cells?

2) **Compare observed and expected cell counts:** which cells have more or less observations than would be expected if H_0 were true?

3) **Look at the terms of the chi-square statistic:** which cells contribute the most to the value of λ^2?

Instructional Preference		Educational Level		
		Undergraduate	**Graduate**	**Total**
Online	Observed	20	35	55 (55%)
	Expected	33	22	
Face to face	Observed	40	5	45 (45%)
	Expected	27	18	
Total		60 (60%)	40 (40%)	100 (100%)

- In the example, only 5 graduate students preferred face-to-face instruction, compared with the expected value of 18.
- Forty undergraduate students preferred face-to-face instruction, while the expected value was 27.
- ❖ Therefore, researchers can conclude not only that there is an association between the variables, but they can also describe the association.
- ❖ Researchers can conclude that graduate students prefer online instruction, whereas undergraduate students prefer face-to-face instruction.

Cell Counts Required for the Chi-Square Test

❖ The chi-square test is an approximate method that becomes more accurate as the counts in the cells of the table get larger. Therefore, it is important to check that the counts are large enough to result in a trustworthy p-value. Fortunately, the chi-square approximation is accurate for very modest counts.

> **Cell Counts Required for the Chi-Square Test**
> You can safely use the chi-square test with critical values from the chi-square distribution when no more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater. In particular, all four expected counts in a 2×2 table should be 5 or greater.

Uses of the Chi-Square Test

❖ One of the most useful properties of the chi-square test is that it tests the null hypothesis "the row and column variables are not related to each other" whenever this hypothesis makes sense for a two-way variable.

> **Uses of the Chi-Square Test**
> Use the chi-square test to test the null hypothesis
> H_0: there is no relationship between two categorical variables
>
> when there is a two-way table from one of these situations:
> - Independent random samples from two or more populations, with each individual classified according to one categorical variable.
> - A single random sample, with each individual classified according to both of two categorical variables.