3.4: Chapter 14- Chi-Square

Key Terms

chi-square

contingency table

expected values

marginal values

nonparametric tests

test for goodness of fit

test for independence

We come at last to our final topic: chi-square (χ^2). This test is a special form of analysis called a nonparametric test, so the structure of it will look a little bit different from what we have done so far. However, the logic of hypothesis testing remains unchanged. The purpose of chi-square is to understand the frequency distribution of a single categorical variable or find a relationship between two categorical variables, which is a frequently very useful way to look at our data.

Categories and Frequency Tables

Our data for the χ^2 test are categorical—specifically nominal—variables. Recall from Unit 1 that nominal variables have no specified order and can only be described by their names and the frequencies with which they occur in the dataset. Thus, unlike the other variables we have tested, we cannot describe our data for the χ^2 test using means and standard deviations. Instead, we will use frequencies tables.

Table 14.1 gives an example of a frequency table used for a χ^2 test. The columns represent the different categories within our single variable, which in this example is pet preference. The χ^2 test can assess as few as two categories, and there is no technical upper limit on how many categories can be included in our variable, although, as with ANOVA, having too many categories makes our computations long and our interpretation difficult. The final column in the table is the total number of observations, or N. The χ^2 test assumes that each observation comes from only one person and that each person will provide only one observation, so our total observations will always equal our sample size.

Table 14.1. Pet preferences

	Cat	Dog	Other	Total
Observed	14	17	5	36
Expected	12	12	12	36

There are two rows in this table. The first row gives the observed frequencies of each category from our dataset; in this example, 14 people reported preferring cats as pets, 17 people reported preferring dogs, and 5 people reported a different animal. The second row gives expected values; expected values are what would be found if each category had equal representation. The calculation for an expected value is:

$$E = \frac{N}{C}$$

where N is the total number of people in our sample and C is the number of categories in our variable (also the number of columns in our table). The expected values correspond to the null hypothesis for χ^2 tests: equal representation of categories. Our

first of two χ^2 tests, the test for goodness of fit, will assess how well our data lines up with, or deviates from, this assumption.

Test for Goodness of Fit

The test for goodness of fit assesses one categorical variable against a null hypothesis of equally sized frequencies. Equal frequency distributions are what we would expect to get if categorization was completely random. We could, in theory, also test against a specific distribution of category sizes if we have a good reason to. If we have information about how a population is distributed, we could compare our observed sample distribution to the expected values if the sample followed the same distribution as the population. For example, if we know that in the population of a small liberal arts college, 15% of students are international students, while 85% are domestic students, we would then calculate expected values for our sample using these percentages. In that case, we would be testing against the null hypothesis of 15% international students. This is less common, so we will not deal with more examples of this sort in this text.

Hypotheses

All χ^2 tests, including the test for goodness of fit, are nonparametric tests. This means that there is no population parameter we are estimating or testing against; we are working only with our sample data. Because of this, there are no mathematical statements for χ^2 hypotheses. This should make sense because the mathematical hypothesis statements were always about population parameters (e.g., μ), so if we are nonparametric, we have no parameters and therefore no mathematical statements.

We do, however, still state our hypotheses verbally. For χ^2 tests for goodness of fit, our null hypothesis is that there is an equal number of observations in each category. That is, there is no difference between the categories in how prevalent they are. Our alternative hypothesis says that the categories do differ in their frequency. We do not have specific directions or one-tailed tests for χ^2, matching our lack of mathematical statements.

Degrees of Freedom and the χ^2 Table

Our degrees of freedom for the χ^2 test are based on the number of categories we have in our variable, not on the number of people or observations like it was for our other tests. Luckily, they are still as simple to calculate:

$$df = C - 1$$

So for our pet preference example, we have 3 categories, thus we have 2 degrees of freedom. Our degrees of freedom, along with our significance level (still defaulted to a = .05) are used to find our critical values in the χ^2 table, a portion of which is shown in Table 14.2. (The complete χ^2 table can be found in Appendix E.) Because we do not have directional hypotheses for χ^2 tests, we do not need to differentiate between critical values for one- or two-tailed tests. In fact, just like our F tests for regression and ANOVA, all χ^2 tests are one-tailed tests.

Table 14.2. Critical Values for Chi-Square (χ^2 Table)

df	Proportion in Critical Region				
	.1	.05	.02	.01	.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548

df	Proportion in Critical Region				
	.1	.05	.02	.01	.005
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188

χ^2 Statistic

The calculations for our test statistic in χ^2 tests combine our information from our observed frequencies (O) and our expected frequencies (E) for each level of our categorical variable. For each cell (category) we find the difference between the observed and expected values, square them, and divide by the expected values. We then sum this value across cells for our test statistic. This is shown in the formula:

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

For our pet preference data, we would have:

$$\chi^2 = \frac{(14-12)^2}{12} + \frac{(17-12)^2}{12} + \frac{(5-12)^2}{12} = 0.33 + 2.08 + 4.08 = 6.49$$

Notice that, for each cell's calculation, the expected value in the numerator and the expected value in the denominator are the same value. Let's now take a look at an example from start to finish.

Example Pineapple on Pizza

There is a very passionate and ongoing debate about whether pineapple should go on pizza. Being the objective, rational data analysts that we are, we will collect empirical data to see if we can settle this debate once and for all. We gather data from a group of adults, asking for a simple yes-or-no answer.

Step 1: State the Hypotheses

We start, as always, with our hypotheses. Our null hypothesis of no difference states that an equal number of people will say they do and do not like pineapple on pizza, and our alternative hypothesis will be that one side wins out over the other:

H_0: An equal number of people do and do not like pineapple on pizza

H_A: A significant majority of people agree one way or the other

Step 2: Find the Critical Value

To avoid any potential bias in this crucial analysis, we will leave α at its typical level. We have two options in our data (Yes or No), which will give us two categories. Based on this, we will have 1 degree of freedom. From our χ^2 table, we find a critical value of 3.84.

Step 3: Calculate the Test Statistic and Effect Size

The results of the data collection are presented in Table 14.3. We had data from 45 people in all and 2 categories, so our expected values are E = 45/2 = 22.50.

Table 14.3. Pineapple-on-pizza preferences

	Yes	No	Total
Observed	26	19	45

	Yes	No	Total
Expected	22.50	22.50	45

We can use these to calculate our χ^2 statistic:

$$\chi^2 = \frac{(26-22.50)^2}{22.50} + \frac{(19-22.50)^2}{22.50} = 0.54 + 0.54 = 1.08$$

Effect Size for χ^2

Like all other significance tests, χ^2 tests—both for goodness of fit and for independence—have effect sizes that can and should be calculated. There are many options for which effect size to use, and the ultimate decision is based on the type of data, the structure of your frequency or contingency table, and the types of conclusions you would like to draw. For the purpose of our introductory course, we will focus only on a single effect size that is simple and flexible: Cramer's V.

Cramer's V is a type of correlation coefficient that can be computed on categorical data. Like any other correlation coefficient (e.g., Pearson's r), the cutoffs for small, medium, and large effect sizes of Cramer's V are .10, .30, and .50, respectively. The calculation of Cramer's V is very simple:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

For this calculation, k is the smaller value of either R (the number of rows) or C (the number of columns). The numerator is simply the test statistic we calculate during Step 3 of the hypothesis-testing procedure. For our example, we had 2 rows and 3 columns, so k = 2:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{1.08}{45(2-1)}} = \sqrt{\frac{1.08}{45}} = \sqrt{.024} = .15$$

So the statistically significant relationship between our variables was moderately strong.

Step 4: Make the Decision

Our observed test statistic had a value of 1.08 and our critical value was 3.84. Our test statistic was smaller than our critical value, so we fail to reject the null hypothesis, and the debate rages on. Figure 14.1 shows the output from JASP for this example.

Figure 14.1. Output from JASP for the χ^2 test for goodness of fit described in the Pineapple on Pizza example. The output provides the χ^2 statistic (1.089), degrees of freedom (1) and the exact p value (.297, which is greater than .05). The output also provides the observed values and expected values (note that both expected values are 22.5, but decimals are not shown). Based on our sample of 45 people, there is no significant difference between the observed and expected values for preferring pineapple on pizza, χ^2 (1, N = 45) = 1.089, p = .297. ("JASP chi-square goodness of fit" by Rupa G. Gordon/Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

Multinomial Test

Multinomial Test

	χ^2	df	p
H_0 (a)	1.089	1	0.297

Descriptives

Choice	Observed	Expected: H_0 (a)
No	19	22
Yes	26	22

Contingency Tables for Two Variables

The test for goodness of fit is a useful tool for assessing a single categorical variable. However, what is more common is wanting to know if two categorical variables are related to one another. This type of analysis is similar to a correlation, the only difference being that we are working with nominal data, which violates the assumptions of traditional correlation coefficients. This is where the χ^2 test for independence comes in handy.

As noted above, our only description for nominal data is frequency, so we will again present our observations in a frequency table. When we have two categorical variables, our frequency table is crossed. That is, each combination of levels from each categorical variable is presented. This type of frequency table is called a contingency table because it shows the frequency of each category in one variable, contingent upon the specific level of the other variable.

An example contingency table is shown in Table 14.4, which displays whether or not 168 college students watched college sports growing up (Yes/No) and whether the students' final choice of which college to attend was influenced by the college's sports teams (Yes, primary; Yes, somewhat; No).

Table 14.4. Contingency table of college sports and decision making

		Affected Decision			
		Primary	Somewhat	No	Total
Watched as a child	Yes	47	26	14	87
	No	21	23	37	81
	Total	68	49	51	168

In contrast to the frequency table for our test for goodness of fit, our contingency table does not contain expected values, only observed data. Within our table, wherever our rows and columns cross, we have a cell. A cell contains the frequency of observing its corresponding specific levels of each variable at the same time. The top left cell in Table 14.4 shows us that 47 people in our study watched college sports as a child and had college sports as their primary deciding factor in which college to attend.

Cells are numbered based on which row they are in (rows are numbered top to bottom) and which column they are in (columns are numbered left to right). We always name the cell using (R,C), with the row first and the column second. A quick and easy way to remember the order is that the brand RC Cola exists but CR Cola does not. Based on this convention, the top left cell containing our 47 participants who watched college sports as a child and had sports as a primary criteria is cell (1,1). Next to it, which has 26 people who watched college sports as a child but had sports only somewhat affect their decision, is cell (1,2), and so on. We only number the cells where our categories cross. We do not number our total cells, which have their own special name: marginal values.

Marginal values are the total values for a single category of one variable, added up across levels of the other variable. In Table 14.4, these marginal values have been made bold for ease of explanation, though this is not normally the case. We can see that, in total, 87 of our participants (47 + 26 + 14) watched college sports growing up and 81 (21 + 23 + 37) did not. The total of these two marginal values is 168, the total number of people in our study. Likewise, 68 people used sports as a primary criterion for deciding which college to attend, 50 considered it somewhat, and 50 did not consider it at all. The total of these marginal values is also 168, our total number of people. The marginal values for rows and columns will always both add up to the total number of participants, N, in the study. If they do not, then a calculation error was made and you must go back and check your work.

Expected Values of Contingency Tables

Our expected values for contingency tables are based on the same logic as they were for frequency tables, but now we must incorporate information about how frequently each row and column was observed (the marginal values) and how many people were in the sample overall (N) to find what random chance would have made the frequencies out to be. Specifically:

$$E_{ij} = \frac{R_i C_j}{N}$$

The subscripts i and j indicate which row and column, respectively, correspond to the cell we are calculating the expected frequency for, and the Ri and Cj are the row and column marginal values, respectively. N is still the total sample size. Using the data from Table 14.4, we can calculate the expected frequency for cell (1,1), the college sport watchers who used sports at their primary criteria, to be:

$$E_{1,1} = \frac{(87)(68)}{168} = 35.21$$

We can follow the same math to find all the expected values for this table:

		Affected Decision			
		Primary	Somewhat	No	Total
Watched as a child	Yes	35.21	25.38	26.41	87
	No	32.79	23.62	24.59	81
	Total	68	49	51	168

Notice that the marginal values still add up to the same totals as before. This is because the expected frequencies are just row and column averages simultaneously. Our total N will also add up to the same value.

The observed and expected frequencies can be used to calculate the same χ^2 statistic as we calculated for the test for goodness of fit. Before we get there, though, we should look at the hypotheses and degrees of freedom used for contingency tables.

Test for Independence

The χ^2 test performed on contingency tables is known as the test for independence. In this analysis, we are looking to see if the values of each categorical variable (that is, the frequency of their levels) is related to or independent of the values of the other categorical variable. Because we are still doing a χ^2 test, which is nonparametric, we still do not have mathematical versions of our hypotheses. The actual interpretations of the hypotheses are quite simple: the null hypothesis says that the variables are independent or not related, and the alternative hypothesis says that they are not independent or that they are related. Using this setup and the data provided in Table 14.4, let's formally test for whether watching college sports as a child is related to using sports as a criteria for selecting a college to attend.

Example College Sports

We will follow the same four-step procedure as we have since Chapter 7.

Step 1: State the Hypotheses

Our null hypothesis of no difference will state that there is no relationship between our variables, and our alternative will state that our variables are related.

H_0: College choice criteria is independent of college sports viewership as a child

H_A: College choice criteria is related to college sports viewership as a child

Step 2: Find the Critical Value

Our critical value will come from the same table that we used for the test for goodness of fit, but our degrees of freedom will change. Because we now have rows and columns (instead of just columns) our new degrees of freedom use information from both:

$$df = (R-1)(C-1)$$

In our example:

$$df = (2-1)(3-1) = (1)(2) = 2$$

Based on our 2 degrees of freedom, our critical value from our table is 5.991.

Step 3: Calculate the Test Statistic and Effect Size

The same formula for χ^2 is used once again:

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

$$\chi^2 = \frac{(47-35.21)^2}{35.21} + \frac{(26-25.38)^2}{25.38} + \frac{(14-26.41)^2}{26.41} + \frac{(21-32.79)^2}{32.79} + \frac{(23-23.62)^2}{23.62} + \frac{(37-24.59)^2}{24.59}$$

$$\chi^2 = 3.94 + 0.02 + 5.83 + 4.24 + 0.02 + 6.26 = 20.31$$

Step 4: Make the Decision

The final decision for our test of independence is still based on our observed value (20.31) and our critical value (5.991). Because our observed value is greater than our critical value, we can reject the null hypothesis.

Reject H0. Based on our data from 168 people, we can say that there is a statistically significant relationship between whether someone watches college sports growing up and the influence a college's sports teams have on that person's decision on which college to attend, and the effect size was moderate, χ^2 (2, N = 168) = 20.31, p < .05, V < .348.

Figure 14.2 shows the output from JASP for this example.

Figure 14.2. Output from JASP for the χ^2 test for independence described in the College Sports example. The output provides the χ^2 statistic (20.309), degrees of freedom (2), and the p value of less than .001. The output also provides the observed count and expected count in the contingency table and Cramer's V (.348) in the nominal table. Based on our data from 168 people, we can say that there is a statistically significant relationship between whether someone watches college sports growing up and the influence a college's sports teams have on that person's decision on which college to attend, χ^2 (2, N = 168) = 20.31, p < .001, V = .348. ("JASP chi-square independence" by Rupa G. Gordon/Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

Contingency Tables ▼

Contingency Tables

Watched		Decision			
		Somewhat	Primary	No	Total
Yes	Count	26.000	47.000	14.000	87.000
	Expected count	25.375	35.214	26.411	87.000
No	Count	23.000	21.000	37.000	81.000
	Expected count	23.625	32.786	24.589	81.000
Total	Count	49.000	68.000	51.000	168.000
	Expected count	49.000	68.000	51.000	168.000

Chi-Squared Tests ▼

	Value	df	p
X^2	20.309	2	< .001
N	168		

Nominal

	Value[a]
Phi-coefficient	NaN
Cramer's V	0.348

[a] Value could not be calculated - At least one row or column contains all zeros

Exercises

1. What does a frequency table display? What does a contingency table display?
2. What does a test for goodness of fit assess?
3. How do expected frequencies relate to the null hypothesis?
4. What does a test for independence assess?
5. Compute the expected frequencies for the following contingency table:

	Category A	Category B
Category C	22	38
Category D	16	14

6. Test significance and find effect sizes for the following tests:
    1. N = 19, R = 3, C = 2, χ^2 (2) = 7.89, a = .05
    2. N = 12, R = 2, C = 2, χ^2 (1) = 3.12, a = .05
    3. N = 74, R = 3, C = 3, χ^2 (4) = 28.41, a = .01
7. You hear a lot of people claim that The Empire Strikes Back is the best movie in the original Star Wars trilogy, and you decide to collect some data to demonstrate this empirically (pun intended). You ask 48 people which of the original movies they liked best; 8 said A New Hope was their favorite, 23 said The Empire Strikes Back was their favorite, and 17 said Return of the Jedi was their favorite. Perform a χ^2 test on these data at the .05 level of significance.

8. A pizza company wants to know if people order the same number of different toppings. They look at how many pepperoni, sausage, and cheese pizzas were ordered in the last week. Fill out the rest of the frequency table and test for a difference.

	Pepperoni	Sausage	Cheese	Total
Observed	320	275	251	
Expected				

9. A university administrator wants to know if there is a difference in proportions of students who go on to grad school across different majors. Use the data below to test whether there is a relationship between college major and going to grad school.

		Major		
		Psychology	Business	Math
Graduate School	Yes	32	8	36
	No	15	41	12

10. A company you work for wants to make sure they are not discriminating against anyone in their promotion process. You have been asked to look across gender to see if there are differences in promotion rate (i.e., if gender and promotion rate are independent or not). The following data should be assessed at the normal level of significance:

		Promoted in Last Two Years?	
		Yes	No
Gender	Women	8	5
	Men	9	7

Answers to Odd-Numbered Exercises

Frequency tables display observed category frequencies and (sometimes) expected category frequencies for a single categorical variable. Contingency tables display the frequency of observing people in crossed category levels for two categorical variables, and (sometimes) the marginal totals of each variable level.

3)

Expected values are what we would observe if the proportion of categories was completely random (i.e., no consistent difference other than chance), which is the same was what the null hypothesis predicts to be true.

5)

Observed:

	Category A	Category B	Total
Category C	22	38	60
Category D	16	14	30
Total	38	52	90

Expected:

	Category A	Category B	Total
Category C	$\frac{(60)(38)}{90} = 25.33$		

$\frac{(60)(52)}{90} = 34.67$

60

Category D

$\frac{(30)(38)}{90} = 12.67$ $\frac{(30)(52)}{90} = 17.33$

30

Total

38

52

90

7)

Step 1:H0: “There is no difference in preference for one movie,” HA: “There is a difference in how many people prefer one movie over the others.”

Step 2: Three categories (columns) gives df = 2, $\chi^2_{crit} = 5.991$
Step 3: Based on the given frequencies:

	New Hope	Empire	Jedi	Total
Observed	8	23	17	48
Expected	16	16	16	

$\chi^2 = 7.13$. Since this is a statistically significant result, we should calculate an effect size:Cramer’s $V = \sqrt{\frac{7.13}{48(3-1)}} = .27$, which is a moderate effect size

Step 4: Our obtained statistic is greater than our critical value, reject H0. Based on our sample of 48 people, there is a statistically significant difference in the proportion of people who prefer one Star Wars movie over the others, χ^2 (2, N = 48) = 7.13, p < .05.

9)

Step 1:H0: “There is no relationship between college major and going to grad school,” HA: “Going to grad school is related to college major.”

Step 2:df= 2, $\chi^2_{crit} = 5.991$

Step 3:Based on the expected frequencies:

		Major		
		Psychology	Business	Math
Graduate School	Yes	24.81	25.86	25.33
	No	22.19	23.14	22.67

$\chi^2 = 2.09 + 12.34 + 4.49 + 2.33 + 13.79 + 5.02 = 40.05$Step 4: Obtained statistic is greater than the critical value, reject H0. Based on our data, there is a statistically significant relationship between college major and going to grad school, χ^2 (2, N = 144) = 40.05, p < .05, V = .53, which is a large effect.