11

The Chi-Square Test

Our trek through the world of hypothesis testing so far has involved procedures based on one or more means. For example, we used the *t* test to determine whether or not there was a significant difference between fraternity members and non-members in mean levels of alcohol consumption. We relied on the ANOVA procedure when we wanted to look at mean levels of unemployment in four regions. In each of those cases, one of the variables in the hypothesis was an interval/ratio level variable. The requirement of interval/ratio data is central

- *Key Terms*
- *Chapter Problems*

Our trek through the world of hypothesis testing so far has involved procedures based on one or more means. For example, we used the *t* test to determine whether or not there was a significant difference between fraternity members and non-members in mean levels of alcohol consumption. We relied on the ANOVA procedure when we wanted to look at mean levels of unemployment in four regions. In each of those cases, one of the variables in the hypothesis was an interval/ratio level variable. The requirement of interval/ratio data is central to any hypothesis test involving means. The reason should be obvious: You can't calculate a mean unless you have interval/ratio data.

As you might expect, though, not all research situations involve interval-level data. Social scientists often encounter research situations in which the variables are measured at the nominal or ordinal level. The term **categorical**

data is typically used to describe information of this sort, because the data represent simple categories. Consider the following examples:

A response to a question might be *yes, no,* or *undecided*.

A response to a question might be *strongly agree, agree, disagree,* or *strongly disagree*.

A person might be classified as *Republican, Democrat,* or *Independent*.

A university might be classified as *public* or *private*.

When faced with a hypothesis-testing situation involving categorical variables (nominal or ordinal data), statisticians often turn to the chi-square test. In this chapter, we'll consider the chi-square test of independence, a procedure that is very appropriate for situations involving categorical data.

LEARNING CHECK

Question: What does the term *categorical data* mean?
Answer: Data expressed in simple categories—nominal- or ordinal-level data.

Before We Begin

You were introduced to the notion of null hypotheses in Chapter 7, and you also learned that there were many ways to express a null hypothesis. As you moved through Chapters 7, 8, and 9, you were exposed to hypothesis testing in a variety of situations, but in most of those cases, you dealt with null hypotheses that were statements of *no difference*. In this chapter, though, you'll face something different.

First, you're going to be dealing with a different sort of data, and you won't be calculating any means. It follows, therefore, that you'll have to change your vocabulary. Instead of hypotheses about such notions as no difference between means, you'll be dealing with null hypotheses that speak in terms of no relationship or chance relationship. All of that will make more sense as we move forward. For the moment, simply prepare for a slight shift in perspective.

The Chi-Square Test of Independence

The **chi-square test of independence** is a test that allows us to determine whether or not two variables are associated in some way. For example, it allows us to answer the following sorts of questions:

Is political affiliation associated with attitude toward a certain issue?

Is gender associated with selection of an academic major?

Is place of residence associated with attitude on a certain issue?

As you explore the chi-square test of independence, you'll actually go beyond the specific test application. Indeed, you'll also learn quite a bit about how statisticians look at the association between variables in general. As always, we'll start with a look at the logic behind the test.

LEARNING CHECK

Question: What is the chi-square test of independence?

Answer: A hypothesis-testing procedure appropriate for categorical variables. It tests whether or not there is an association between two variables.

The Logic of the Test

Let's start with a simple example. Let's assume that we've set out to determine whether or not there is any association between a person's political party affiliation (Republican, Democrat, or Independent) and how that person views a downtown redevelopment proposal (for, against, or undecided). In other words, we want to know if respondents' attitudes toward the proposal vary according to political party affiliation.

Let's assume we have asked a random sample of 180 residents to tell us about their political party affiliation (Republican, Democrat, or Independent) and how they feel about the proposal (for, against, or undecided). We can record the results in what's known as a **contingency table**. A contingency table is a classification tool that reveals the various possibilities (contingencies) in the comparison of variables. In a moment, I'll ask you to take a look at some results displayed in a contingency table. First, though, let me urge you to study carefully the various tables I ask you to consider. Don't just take a brief look and move on; take the time to carefully consider the illustrations.

Now take a look at Table 11-1. It presents two contingency tables, each reflecting a rather extreme pattern of responses, based on a sample of 180 respondents. Each table shows the possible response combinations, along with totals. Different response combinations are presented in individual cells of the table. Because the totals are presented in the margins of the table, we refer to them as **marginal totals**. In the real world, it's doubtful that we'd get such extreme patterns of responses, but we can afford to take leave of the real world for a moment or two. The goal is to develop an understanding of the chi-square test of independence and the logic that underlies it.

First, take a close look at Pattern A in Table 11-1. Think about these questions:

How many Republicans are represented in the table?

How many Democrats are represented in the table?

How many Independents are represented in the table?

Table 11-1 Two Contingency Table Patterns

Pattern A

		Political Party Affiliation			
		Republican	*Democrat*	*Independent*	*Total*
View of Downtown Redevelopment Proposal	*For*	20	20	20	60
	Against	20	20	20	60
	Undecided	20	20	20	60
	Total	60	60	60	180

Pattern B

		Political Party Affiliation			
		Republican	*Democrat*	*Independent*	*Total*
View of Downtown Redevelopment Proposal	*For*	40	10	10	60
	Against	10	40	10	60
	Undecided	10	10	40	60
	Total	60	60	60	180

Just looking at the Republicans, how are they distributed in terms of the attitude variable? Are they fairly evenly distributed, or are they more or less concentrated in a particular cell of the table? In other words, could you say it looks as though Republicans are inclined toward a particular attitude?

What about the Democrats? Are they fairly evenly distributed across the attitude variable, or are they concentrated in a particular cell? Can you associate Democrats with a particular attitude?

What about the Independents? How are they distributed?

Given what you know so far about Pattern A, does there appear to be any association between political affiliation and attitude? (The answer is no.)

The answer is no because the overall pattern of the distribution is clearly even across the cells. Republicans are just as likely to be for the proposal as they are to be against the proposal or undecided. The same is true for Democrats and Independents. In a case like that, it would be difficult to say there is a difference between Republicans, Democrats, and Independents when it comes to the distribution of attitudes toward the redevelopment proposal.

Now look at Pattern B. Think about these questions:

How many Republicans are represented in the table?

How many Democrats?

How many Independents?

Interesting—you might say to yourself—the same number of people were represented in the previous response pattern (Pattern A). But now take a look at how the overall pattern has changed.

What about the Republicans? Are they concentrated in a particular cell?

What about the Democrats? Are they more likely to be associated with a particular attitude?

When it comes to the Independents, how are they distributed in terms of the attitude variable?

What does all of that suggest? Does it appear that there's an association between the variables? (The answer is yes.)

If you study Pattern B, you'll likely conclude that there appears to be some sort of association between political affiliation and attitude. Granted, the information at hand is only based on a sample of 180 respondents, but it still appears that there's some sort of association between the two variables (political affiliation and attitude toward the redevelopment proposal).

LEARNING CHECK

Question: What is a contingency table?
Answer: A table that presents data in terms of all combinations of two or more variables.

Before we go any further, let me reemphasize that these examples are extreme. The tables were constructed a certain way to demonstrate specific points. You *could* find results like those in the real world, but, as a rule, you're apt to find some pattern in between the two extremes. Let me explain.

First, the examples we've looked at reflect equal numbers of Republicans, Democrats, and Independents in the sample. Something like that is possible in the real world, to be sure; a community could be evenly divided among Republicans, Democrats, and Independents. More than likely, though, the actual distribution of political affiliation in a community won't be equal. Therefore, we'd expect a real-world sample to reflect the unequal distribution that actually exists in the community. Second, in a real-world instance, we'd likely get a more varied dispersion of responses over the entire table—neither completely even nor obviously concentrated in just a few cells.

That said, let me give you a general guideline to follow when looking at a contingency table: Always remember what the object of the analysis is. We want to know if the distribution of one variable seems to vary on the basis of the distribution of another variable. That, of course, is another way of saying that we want to know if there's any association between the two variables.

When there's a fairly even distribution of cases over all the cells, there's probably little, if any, association between the two variables. On the other hand, when there's a concentration of responses or cases in just a few cells, there's a greater chance that there's some sort of underlying connection between the two variables. If necessary, return to Table 11-1 to review the two patterns again. Think of Pattern A as one that reflects an even distribution of responses or cases over the table—a pattern that suggests no connection between the variables. Think of Pattern B as one that reflects a noticeable concentration of responses in just a few cells—a pattern that suggests the possibility of an association between the two variables.

The question of whether or not there's an association between two variables is something we've considered before. When we applied the difference of means test, we were actually examining the association between two variables. For example, the *t* test for the difference in alcohol consumption by fraternity members and non-members was actually a test to determine whether or not there was an association between fraternity membership status and level of alcohol consumption. When we used the ANOVA procedure to consider levels of unemployment by region, we were asking whether or not there was any association between region and unemployment level.

When it comes to the chi-square test of independence, we're asking similar types of questions. In the present example, the question is whether or not there's an association between political affiliation and attitude toward a redevelopment proposal. In other words, are the variables associated in some way, or are they *independent* of one another?

To assert that there's an association between two variables is to say that

another or *why* you might be able to predict one variable from the other. Sometimes we're inclined to think in terms of causation—the idea that one variable *causes* the other—but I would caution you about that. As I'm fond of telling my students, causation is something that largely exists in our minds—it's a model or an explanation that we sometimes mistakenly impose on our data or results. Except in highly controlled experimental research situations, it's difficult to make legitimate claims of direct causation.

For example, some variables are associated only in the sense that they are expressions of a common concept. Consider the fact that many people who excel in the sport of football also excel in the sport of baseball. Just because people who are proficient in one sport are often proficient in the other sport doesn't mean that proficiency in one area causes proficiency in the other. In fact, both may be expressions of a common concept—namely, athletic ability. Being able to play football well probably doesn't cause someone to play baseball well (or vice versa). Instead, it's likely that people with a pronounced athletic ability tend to do well in almost any sport.

In short, the question is *whether* two variables are associated in some way—not *why* they're associated. Simply put, association doesn't necessarily imply causation. That said, we can consider the chi-square test of independence in the context of chance and a departure from chance.

A Focus on the Departure From Chance

Assuming you've gained an appreciation as to why it's a good idea to approach the notion of causality with caution, we can return to the fundamental logic behind the chi-square test of independence. To understand the logic, start with the idea that this procedure looks at the overall pattern in a contingency table and measures the extent to which the pattern reflected in the table departs from chance. To understand what this means, take another look at Pattern A in Table 11-1. One way to think about Pattern A is that it's a pattern you'd be likely to get if nothing but chance were at play. In other words, you'd be likely to get a pattern like this if the two variables were not tied together in any way.

Focus now on the marginal totals. When it comes to being a Republican, Democrat, or Independent (that is, the distribution of the political party affiliation variable), the picture reflected in Pattern A appears to be one of chance. Given the distribution of these 180 respondents, there appears to be an equal chance of being a Republican, a Democrat, or an Independent. By the same token, there appears to be an equal chance of someone's being for, against, or undecided regarding the redevelopment proposal. It seems to be mere chance whether Republicans are for, against, or undecided on the proposal. The same could be said for the Democrats and the Independents. The pattern may be extreme, but it should give you an idea of what a pattern of chance would look like in the context of a contingency table.

As you discovered before, though, Pattern B is very different. In fact, it's so different that it's reasonable to say that this response pattern represents a noticeable departure from chance. In fact, that's the meaning of the phrase *significant association*—an association that departs from chance.

In essence, that's what the chi-square test of independence is all about. It allows us to look at a pattern in a contingency table and determine whether or not the pattern we observe is one that departs from chance.

LEARNING CHECK

Question: What does it mean to say that two variables are associated?

Answer: The pattern exhibited by the association of the two variables represents a departure from chance.

The Null Hypothesis

In the case of the chi-square test, we move away from the symbolic or mathematical statements of a null hypothesis such as those we used with the *t* test or ANOVA. For this test, there are no statements about means being equal. Instead, we move to a statement about the association between two variables.

For example, let's say we wanted to explore the association between two variables: type of community (urban, suburban, or rural) and intention to vote (whether someone plans to vote in the next election—yes, no, or undecided). An appropriate statement of the null hypothesis would be as follows:

H_0: There is no association between type of community and intention to vote.

When we use the chi-square test of independence, we test the null hypothesis by examining the results obtained from sample data. But we do so with the idea that the sample patterns are representative of population patterns. We look at the pattern in the contingency table (the observed data), but our interest really goes beyond that.

If the pattern shows little, if any, departure from what would be expected by chance, we fail to reject the null hypothesis. In other words, we fail to reject the idea of no association between the variables. If, on the other hand, the pattern reflects a significant departure from what we would expect by chance (given the marginal totals of the variables in question), we reject the null. In doing so, we are suggesting that there is, in fact, some sort of association between the two variables in the population.

The Application

As we've done before, we'll put off any discussion of the formula until we've spent a bit of time with the problem at hand. As a start, take a look at the data in Table 11-2. Once again, we have a contingency table. This time, the contingency table shows responses from 98 people to questions about their type of community and their intention to vote. Since we're beginning the application at this point, we'll assume we've set the level of significance at .05.

Table 11-2 Responses to Survey: Voter Intention by Type of Community

		Type of Community			
		Urban	*Suburban*	*Rural*	*Total*
Voter Intention	*Yes*	8	17	7	32
	No	6	8	15	29
	Undecided	19	7	11	37
	Total	33	32	33	98

Remember what a contingency table is all about and what it allows us to do. It's a mechanism that allows us to see all possible combinations of variables in a given research situation. When we look at a contingency table, the numbers we see in the various cells (with the exception of the marginal totals) are referred to as the **observed frequencies**. You've seen observed frequencies before. That's really what you saw when you looked at Pattern A and Pattern B in Table 11-1. In Table 11-2, we're looking at a different contingency table and a different pattern. The observed frequencies are simply the results that are presented in Table 11-2.

LEARNING CHECK

Question: In the chi-square test of independence, what are the observed frequencies?

Answer: The frequencies (results) that appear in each cell of a contingency table (excluding the marginal totals).

Table 11-2 is known as a three-by-three contingency table; it has three rows and three columns. With three rows and three columns, the table has a

total of nine cells (exclusive of the cells associated with the marginal totals). The numbers in each of these nine cells are the observed frequencies or cases. Looking at the upper left-hand cell, for example, you see the number eight. The observed frequency for that cell is eight. This means that eight respondents reported that they are urban residents and that they intend to vote in the election. The concept of observed frequencies, as you've probably gathered by now, is quite straightforward. They are simply the numbers you see displayed in the contingency table. The table we're considering now has nine cells, so there are nine observed frequencies.

We turn now to the matter of **expected frequencies**. As with the observed frequencies, there will be nine expected frequencies—one for each cell. To obtain the value of the expected frequencies, though, we'll have to go through a few calculations. Let me explain.

The expected frequency for each cell is a statement of the frequency that we would expect to find, given the marginal distributions and the total number of cases in the table. More precisely, the expected frequency for a given cell is a function of the number of cases in the row in question times the number of cases in the column in question, divided by the total number of cases for the entire table. For the sake of simplicity, we can summarize the calculation as follows:

$$\text{Expected Frequency of Each Cell} = \frac{\text{Row Total} \times \text{Column Total}}{n}$$

For example, to calculate the expected frequency for the cell in the upper left-hand corner of the table (the cell that contains an observed frequency of 8), we would proceed as follows: We would multiply the row total (32) by the column total (33). Then we'd divide the product by the total number of cases in the sample ($n = 98$). The result would be 10.78. Moving to the next cell in that row (the cell with the observed frequency of 17), we would calculate the expected frequency by multiplying the row total (32) by the column total (32) and, as before, we'd divide the product by the total number of cases in the sample ($n = 98$). The result would be an expected frequency of 10.45. Calculating the expected frequencies for each cell, we'd obtain the information presented in Table 11-3. Note that there is an observed frequency (f_o) and an expected frequency (f_e) for each of the nine cells in the table.

The individual steps in the calculation of expected frequencies are shown below. Note how the individual steps correspond to the calculations presented in Table 11-3.

Upper-left f_e	$= (32 \times 33)/98$	$= 1056/98$	$= 10.78$
Upper-middle f_e	$= (32 \times 32)/98$	$= 1024/98$	$= 10.45$
Upper-right f_e	$= (32 \times 33)/98$	$= 1056/98$	$= 10.78$
Middle-left f_e	$= (29 \times 33)/98$	$= 957/98$	$= 9.77$
Middle-middle f_e	$= (29 \times 32)/98$	$= 928/98$	$= 9.47$
Middle-right f_e	$= (29 \times 33)/98$	$= 957/98$	$= 9.77$
Lower-left f_e	$= (37 \times 33)/98$	$= 1221/98$	$= 12.46$

Table 11-3 Calculation of Expected Frequencies

Cell	Observed Frequency (f_o)	Row Total	Column Total	Row Total × Column Total	Row Total × Column Total Divided by n
					Expected Frequency (f_e)
Upper-left	8	32	33	1056	10.78
Upper-middle	17	32	32	1024	10.45
Upper-right	7	32	33	1056	10.78
Middle-left	6	29	33	957	9.77
Middle-middle	8	29	32	928	9.47
Middle-right	15	29	33	957	9.77
Lower-left	19	37	33	1221	12.46
Lower-middle	7	37	32	1184	12.08
Lower-right	11	37	33	1221	12.46

$$\text{Lower-middle } f_e = (37 \times 32)/98 = 1184/98 = 12.08$$
$$\text{Lower-right } f_e = (37 \times 33)/98 = 1221/98 = 12.46$$

Table 11-4 presents the observed and expected frequencies for each cell in an illustration similar to Table 11-2.

LEARNING CHECK

Question: In the chi-square test of independence, what are the expected frequencies?

Answer: The frequencies that would be expected by chance in each cell of a contingency table, given the marginal totals.

The Formula

Given the observed frequencies, and having calculated the expected frequencies, we now have all the elements required by the formula for the chi-square test of independence. At first, the formula for chi-square (symbolized as χ^2) looks a little complicated, but keep in mind that there are really only two fundamental elements—observed frequencies and expected frequencies. Don't let

Table 11-4 Comparison of Observed and Expected Frequencies

Observed Frequencies

		Type of Community		
		Urban	*Suburban*	*Rural*
Voter Intention	*Yes*	8	17	7
	No	6	8	15
	Undecided	19	7	11

Expected Frequencies

		Type of Community		
		Urban	*Suburban*	*Rural*
Voter Intention	*Yes*	10.78	10.45	10.78
	No	9.77	9.47	9.77
	Undecided	12.46	12.08	12.46

the summation sign, the exponent, the division, or anything else throw you when you first look at the formula. The essence of the formula really has to do with the observed and expected frequencies.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Always remember what the expected frequencies represent—namely, the frequencies we'd expect if the pattern were due to chance alone (taking into account the marginal distributions of the two variables). If you examine the formula carefully, you'll note that it has to do with the *difference* between observed and expected frequencies. The formula reflects an overall summation of this difference.

Because the object of the chi-square test of independence is to determine if the pattern reflected in a contingency table departs from chance in a significant manner, it should make intuitive sense that the formula should involve a measure of the overall difference between observation and expectation (or chance). The larger the difference between the observed frequencies and the expected frequencies, the larger will be the calculated value of chi-square. With that as a background, we can now move to the specific steps in the calculation.

The Calculation

As we've done before, we'll approach the calculations in a step-by-step fashion. For the sake of review, here's the formula again, followed by the individual steps in the calculation.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = (8 - 10.78)^2/10.78 + (17 - 10.45)^2/10.45 + (7 - 10.78)^2/10.78 + (6 - 9.77)^2/9.77 + (8 - 9.47)^2/9.47 + (15 - 9.77)^2/9.77 + (19 - 12.46)^2/12.46 + (7 - 12.08)^2/12.08 + (11 - 12.46)^2/12.46$$

$$\chi^2 = 7.73/10.78 + 42.90/10.45 + 14.29/10.78 + 14.21/9.77 + 2.16/9.47 + 27.35/9.77 + 42.77/12.46 + 25.81/12.08 + 2.13/12.46$$

$$\chi^2 = 0.72 + 4.11 + 1.33 + 1.45 + 0.23 + 2.80 + 3.43 + 2.14 + 0.17$$

$$\chi^2 = 16.38$$

The formula instructs us first to find the difference between the observed and expected frequencies of each cell ($f_o - f_e$). Those differences are then squared ($f_o - f_e)^2$. The squared difference associated with each cell is then divided by the expected frequency of the cell ($f_o - f_e)^2/f_e$. For example, beginning with the cell in the upper left-hand corner of our contingency table, we note that the observed frequency is 8 and the expected frequency (for the same cell) is calculated as 10.78. The formula directs us first to find the difference between the two values ($f_o - f_e$ or 8 – 10.78, or –2.78). Next we square the difference, which gives us a value of 7.73. We then divide 7.73 by

Table 11-5 Calculation of Chi-Square Test of Independence Statistic

Observed Frequency	Expected Frequency	Observed Minus Expected	Observed Minus Expected Squared	Observed Minus Expected Squared and Divided by Expected
f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
8	10.78	−2.78	7.73	0.72
17	10.45	6.55	42.90	4.11
7	10.78	−3.78	14.29	1.33
6	9.77	−3.77	14.21	1.45
8	9.47	−1.47	2.16	0.23
15	9.77	5.23	27.35	2.80
19	12.46	6.54	42.77	3.43
7	12.08	−5.08	25.81	2.14
11	12.46	−1.46	2.13	0.17
				$\Sigma = 16.38$

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 16.38$$

the expected frequency of the cell in question (10.78). The result (7.73 divided by 10.78) is 0.72.

The same process is followed for each cell in the table—finding the difference between the observed and expected frequencies, squaring the difference, and dividing the difference by the expected frequency of the cell in question. You can follow this sequence for each cell by examining Table 11-5. Once that process is completed for each cell, the results from all the cells are summed to obtain the calculated value of the chi-square statistic. This is the statistic we'll compare to a critical value as we work our way toward a conclusion.

Conclusion and Interpretation

The calculated value of chi-square ($\chi^2 = 16.38$) is shown in the lower right corner of Table 11-5. Once again, we're right back where we've been many times before. We have a calculated test statistic ($\chi^2 = 16.38$), and now we are faced with arriving at a conclusion. As before, our conclusion will be based on a comparison of our calculated test statistic to a critical value, given a certain level of significance, and taking into account a certain number of degrees of freedom.

Appendix H is a table of critical values for the chi-square test of independence. Different levels of significance are shown across the top row of the table, and the first column lists the degrees of freedom. We set the level of significance at .05 at the beginning of our application, so we'll be working with that column. Now we come to the matter of the degrees of freedom.

The number of degrees of freedom *(df)* for the chi-square test of independence is related to the number of cells in the contingency table. More specifically, *df* is determined by multiplying the number of rows in the table, minus 1, by the number of columns in the table, minus 1. The formula is stated as follows:

$$df = (r - 1) \times (c - 1)$$

where r = number of rows and c = number of columns

Since our example involves a contingency table with three rows and three columns, the calculation is as follows:

$$df = (r - 1) \times (c - 1)$$
$$df = (3 - 1) \times (3 - 1)$$
$$df = 2 \times 2$$
$$df = 4$$

Given four degrees of freedom ($df = 4$) and the .05 level of significance, we find that the critical value is 9.49. Our calculated test statistic (our chi-square value) is 16.38. Because our calculated test statistic exceeds the critical value, we're in a position to reject the null hypothesis. In doing so, we reject the idea of no association between the two variables type of community and intention to vote.

As always, there's a known probability (in this case, a 5% chance or less) that we've rejected the null hypothesis when, in fact, it is true. In other words, there's always a chance that our sample suggested that the two variables are associated when they really aren't. The good news, of course, is that we know what the probability is—it's simply the level of significance. When all is said and done, we're on fairly safe ground in our assertion that type of community appears to be associated with intention to vote.

Having explored the chi-square test of independence, it's time for a little reflection. Think about how the various tests of significance have been presented—how you've been introduced to one test after another, yet the underlying logic of hypothesis testing remains the same.

Chapter Summary

In this chapter, we made a major transformation. We moved from consideration of interval data and the calculation of means to the world of categorical data and the analysis of contingency tables. In doing so, we broadened our understanding of the types of situations that are suitable for statistical analysis.

Equally important, we examined the matter of chance, particularly as it relates to the portrayal of research results presented in a contingency table. In doing so, we began to think in terms of both chance and a departure from chance. Moreover, we learned to think of a departure from chance as a suggestion that two variables are associated with each other.

Finally, we looked at what it really means to assert that two variables are associated. With a word of caution, we explored the idea of causation, noting that causation—the idea that a given measurement or response on one variable somehow *causes* a given measurement or response on another variable—can be a tricky matter. In the process, you should have gained some understanding of a larger issue—one that goes beyond the specifics of any particular statistical procedure. In short, you should have gained even more understanding of the logic of scientific research.

Some Other Things You Should Know

The chi-square test of independence is widely used, but it is subject to certain limitations. For example, problems can arise when the number of cases is small, relative to the number of cells in a table. In short, the idea behind the chi-square test of independence is to analyze the pattern of a distribution, but it's difficult to see a pattern when there are just a few cases spread over a lot of cells.

There are two ways to deal with this problem. The table can be restructured so that it has a smaller number of cells—something you could accomplish by combining categories for either or both variables. That approach, however, should always be accompanied by sound justification. It's not something you should do just for the sake of statistical analysis. A more acceptable approach, if possible, is to simply increase the size of the sample. By increasing the sample size, you end up with more cases available to distribute over the same number of cells. That, in turn, increases the likelihood that a pattern of association will emerge (assuming there's a true pattern of association between the variables in the population).

In some cases, certain correction factors are suggested when working with the chi-square test of independence. For example, a 2×2 contingency table typically calls for the use of the Yate's correction for continuity. This involves decreasing the difference between the observed and expected frequencies by .5 for each cell. Similar corrections are often used when the expected frequency in any cell (of any contingency table, not just 2×2 tables) is less than 5.

Finally, you should be aware that the chi-square test of independence only indicates whether or not there is an association between variables. It doesn't say anything about the *strength* of the association. In other words, the test can point to an association or link between two variables, but it says nothing about

how strong that association or link might be. To explore the matter of association strength, a separate procedure (a measure of association application) is required. For a wide-ranging discussion of some of the more commonly used measures of association, see Healy (2002).

Key Terms

categorical data
chi-square test of independence
contingency table
expected frequency
marginal totals
observed frequency

Chapter Problems

Fill in the blanks, calculate the requested values, or otherwise supply the correct answer.

General Thought Questions

1. A _____ table is a classification tool that reveals the various possibilities in the comparison of variables.
2. Information obtained on variables measured at the nominal or ordinal level is said to be _____ data.
3. _____ frequencies are the frequencies presented in the cells of a table.
4. The _____ frequency is the frequency that would be expected to occur in a particular cell, based upon chance and the marginal distributions.
5. The equation for expected frequency for the chi-square test of independence is _____.
6. The equation for degrees of freedom for the chi-square test of independence is _____.
7. There are _____ cells in a 2×2 contingency table.
8. There are _____ cells in a 3×4 contingency table.
9. A 4×6 contingency table has _____ degrees of freedom.
10. A 3×5 contingency table has _____ degrees of freedom.

Application Questions/Problems

1. A chi-square test of independence value of $\chi^2 = 9.26$ is calculated from data in a 3×3 contingency table. Assuming a .05 level of significance, identify the critical value and state your conclusion about the null hypothesis.

2. A chi-square test of independence value of $\chi^2 = 24.05$ is calculated from data in a 4×5 contingency table. Assuming a .05 level of significance, identify the critical value and state your conclusion about the null hypothesis.
3. A chi-square test of independence value of $\chi^2 = 4.28$ is calculated from data in a 2×2 contingency table. Assuming a .05 level of significance, identify the critical value and state your conclusion about the null hypothesis.
4. A chi-square test of independence value of $\chi^2 = 12.26$ is calculated from data in a 3×3 contingency table. Assuming a .05 level of significance, identify the critical value and state your conclusion about the null hypothesis.
5. A chi-square test of independence value of $\chi^2 = 6.15$ is calculated from data in a 4×5 contingency table. Assuming a .05 level of significance, identify the critical value and state your conclusion about the null hypothesis.
6. You are interested in whether there is any association between gender and academic major. Questioning 75 students, you obtain the following results:

		Academic Major				
		Business	*Science*	*Liberal Arts*	*Other*	*Total*
Gender	*Female*	10	9	9	7	35
	Male	12	11	10	7	40
	Total	22	20	19	14	75

   a. How many degrees of freedom are involved?
   b. What is the calculated value of χ^2?
   c. Assuming the .05 level of significance, what would you conclude?
7. You are interested in whether there is any association between attitude (favorable, unfavorable, or undecided) toward Candidate Busk and place of residence (urban, suburban, or rural). Questioning 95 potential voters, you obtain the following results:

		Attitude Toward Candidate			
		Favorable	*Unfavorable*	*Undecided*	*Total*
Place of Residence	*Rural*	19	7	8	34
	Suburban	9	14	6	29
	Urban	6	8	18	32
	Total	34	29	32	95

a. How many degrees of freedom are involved?
b. What is the calculated value of χ^2?
c. Assuming the .05 level of significance, what would you conclude?

8. You are interested in whether there is any association between gender and perception of movie plots. You show a movie that contains both action and love themes to a group of 70 research participants. You ask each participant to categorize the plot as either love, action, or both. Consider the following table of results:

		Perception of Movie Plot			
		Love	*Action*	*Both*	*Total*
Gender	*Female*	7	12	15	34
	Male	9	11	16	36
	Total	16	23	31	70

a. How many degrees of freedom are involved?
b. What is the calculated value of χ^2?
c. Assuming the .05 level of significance, what would you conclude?