Hypothesis Testing with Chi-Square

CHAPTER OBJECTIVES

After reading this chapter, you should be able to

- Understand the process of hypothesis testing
- Define and apply the concept of "statistical significance"
- Test relationships among categorical variables
- Evaluate chi-square test assumptions
- Discuss how sample size affects statistical significance
- Consider tests involving control variables

Descriptive analysis goes only so far. An important task of statistics is to provide statistical evidence for determining whether relationships exist. This is essential to public policy, for example, establishing whether a program or policy had any impact, such as whether an anger management program affected classroom violence. It is also essential to science, establishing whether or not two variables are related. This chapter discusses general procedures for testing whether a relationship exists. This is also called *hypothesis testing*. Different statistical tests for hypothesis testing are used for different measurement levels of variables involved in relationships. This chapter, using chi-square, shows how to test for

relationships between two categorical variables, but the process as described here is valid for other measurement levels, too. Only after the existence of any relationships has been established does it make sense to analyze them further.

WHAT IS CHI-SQUARE?

Chi-square (pronounced "ky-square") is a quantitative measure used to determine whether a relationship exists between two categorical variables. The Greek notation for chi-square is χ^2, which can be used interchangeably with its Latin alphabet spelling, chi-square. Many statistics quantify the relationship between variables in some way. We continue here with the example from Chapter 8 to illustrate the process of calculating chi-square and determining whether a relationship exists, but you are also encouraged to identify categorical variables in your field of interest.

In Chapter 8 we examined the relationship between two categorical variables, namely, gender and the year of promotion for a sample of employees. Managers are concerned that employees are promoted at unequal rates based on gender, raising the possibility of gender discrimination in the workplace. The data are shown again, in Table 11.1. We want to establish whether a relationship exists between gender and year of promotion. Table 11.1 shows both frequency counts and column percentages (in parentheses).

Chi-square provides a quantitative measure of the relationship between two categorical variables, first, by determining what the distribution of observations

Table 11.1 Year of Promotion by Gender: Frequencies and Percentages (frequency counts in parentheses)

	Gender		
Year	Male	Female	Total
1	32.6%	15.4%	23.2%
	(14)	(8)	(22)
2	37.2	26.9	31.6
	(16)	(14)	(30)
3	16.2	42.3	30.5
	(7)	(22)	(29)
4	14.0	15.4	14.7
	(6)	(8)	(14)
Total	100.0	100.0	100.0
	(43)	(52)	(95)

(frequencies) would look like if *no* relationship existed and, second, by quantifying the extent to which the observed distribution (such as in Table 11.1) differs from that determined in the first step. This section explains the calculation of chi-square, which is used in the next section for hypothesis testing (that is, determining whether a relationship exists).

What would the relationship in Table 11.1 look like if no relationship existed between gender and year of promotion? When no relationship exists between gender and the year of promotion, then men and women, by definition, do not differ in promotion rates. The column percentages in Table 11.1 will then be identical for men and women; they will not differ from the aggregate sample of all men and women. This distribution is shown in the "Total" column. When no relationship exists between men and women, both men and women will be promoted at those rates. Hence, 23.2 percent of both men and women will be promoted in their first year, 31.6 percent will be promoted in their second year, 30.5 percent will be promoted in their third year, and 14.7 percent will be promoted in their fourth year.

The frequencies associated with these rates when no relationship exists are called ***expected frequencies***. Table 11.2 shows these expected frequencies. For example, when no difference in promotion rates exists between men and women, 30.5 percent of 43 men, or 13.1 men, would have been promoted in their third year. Similarly, 30.5 percent of 52 women, or 15.9 women, would have been promoted in their third year. The other expected frequencies are calculated in similar fashion in Table 11.2.

Clearly, when the data indicate that no relationship exists between these variables, the values of observed and expected frequencies must be identical. Also, the greater the relationship, the greater the difference between the observed and expected frequencies. The chi-square statistic (χ^2) measures the difference

Table 11.2 Year of Promotion by Gender: Expected Frequencies

	A: Percentages			B: Counts	
	Gender			Gender	
Year	Male	Female	Total	Male	Female
1	23.2%	23.2%	23.2%	(23.2/100)*43 = 10.0	(23.2/100)*52 = 12.1
2	31.6	31.6	31.6	(31.6/100)*43 = 13.6	(31.6/100)*52 = 16.4
3	30.5	30.5	30.5	(30.5/100)*43 = 13.1	(30.5/100)*52 = 15.9
4	14.7	14.7	14.7	(14.7/100)*43 = 6.3	(14.7/100)*52 = 7.6
Total	100.0	100.0	100.0	43.0	52.0
(*n* =)	(43)	(52)	(95)		

between the expected and observed frequencies and is thus a quantitative measure of this relationship. Chi-square is defined in the following manner:

$$\sum_i \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency in a cell and E_i is the expected frequency in a cell. As is readily seen, when $E_i = O_i$, the chi-square value for that cell is zero. Using the frequencies shown in Tables 11.1 and 11.2 (part B), we find that the chi-square value of the first cell is $[(14 - 10)^2/10 = 4^2/10 = 16/10 =] 1.60$. Calculating chi-square for all of the cells yields 8.97, as shown in Table 11.3. *Of course, the value of chi-square is usually calculated by computer.*[1] (It should be noted that additional examples of chi-square calculations can be found online, e.g., at Khan Academy.[2])

Key Point

Chi-square is a quantitative measure of a relationship between two categorical variables.

In short, when no relationship exists between the variables, chi-square equals zero. The greater the relationship, the greater the value of chi-square. Finally, note also that chi-square is always positive and that it provides no information about the direction of the relationship.[3]

HYPOTHESIS TESTING

We now use chi-square to determine whether a relationship exists between gender and promotion. This is called *hypothesis testing.* In our example, the hypothesis is that a relationship exists between gender and the rate of promotion; a hypothesis is a tentative statement about some relationship or condition that is subject to subsequent verification. The ***purpose of hypothesis testing*** is, simply, to determine whether a relationship exists. Specifically, we ask, "What is the probability that the above distribution of promotion rates among 95 men and women is consistent with a distribution in which men and women are promoted at *equal* rates?" That is, is a chi-square value of 8.97 sufficiently large to conclude that men are promoted at a faster rate than women?[4] A *key task* in statistics is to determine how large any measure of a relationship must be in order to say that it is "statistically significant." This part of hypothesis testing involves

- The null hypothesis
- The concept of statistical significance
- Critical values
- Steps to determine statistical significance

These issues are relevant to all statistical tests, such as chi-square tests, *t*-tests, and others discussed in this book.

Table 11.3 Year of Promotion by Gender: Expected Frequencies

	Gender						
	Male			Female			Total
Year	Obs.	Exp.	x^2	Obs.	Exp.	x^2	x^2
1	14	10.0	1.60	8	12.1	1.39	2.99
2	16	13.6	0.42	14	16.4	0.35	0.77
3	7	13.1	2.84	22	15.9	2.34	5.18
4	6	6.3	0.01	8	7.6	0.02	0.03
Total	43	43.0	4.87	52	52.0	4.10	8.97

Note: Obs. = observed frequency; exp. = expected frequency.

The Null Hypothesis

Since statistics is a careful and cautious discipline, we presume that no relationship between variables exists and that any relationship that is found may have been obtained purely by chance. The ***null hypothesis*** states that *any observed pattern is due solely to chance* and that, hence, no relationship exists. Thus, the null hypothesis (that is, that no relationship exists) is assumed, and an objective of statistical testing is to examine whether the null hypothesis can be rejected. This idea is similar to the court of justice in which individuals are presumed innocent until proven guilty beyond a reasonable doubt. In our example, we presume that no relationship exists between gender and the rate of promotion.

In statistics the specific concern is that we may find a relationship in our sample when in fact none exists in the population. This may occur because of a fluke in our random sample. We endeavor to disprove this possibility. Another way of looking at this issue is that if we assume that a relationship does exist, we might be guilty of not trying hard enough to prove that it doesn't exist. By assuming that a relationship doesn't exist, we need only satisfy the standard of "reasonable evidence" in order to claim that it does exist. That standard is that it should be *very unlikely to find a relationship among variables* (that is, a test-statistic value such as chi-square) *of a certain (large) magnitude when in fact no relationship exists in the population.*

The null hypothesis is stated as follows:

H_0: No relationship exists between gender and the rate of promotion.

The alternate hypothesis is stated as follows:

H_A: A relationship exists between gender and the rate of promotion.

H_0 is the null hypothesis, and H_A is called the *alternate hypothesis.* H_0 is also sometimes called the straw man because we endeavor to "strike it down" or disprove it. The ***alternate hypothesis*** is the logical opposite of the null hypothesis; all possibilities must be accounted for between the null hypothesis and the alternate hypothesis.

In most instances, the null hypothesis is that *no relationship exists* between two variables, and the alternate hypothesis is that *a relationship does exist* between two variables. However, if the researcher has a priori information that a relationship can exist only in one direction (for example, that men can be promoted faster than women but that women cannot be promoted faster than men), then it is appropriate to state the null hypothesis as "men are not promoted faster than women" and the alternate hypothesis as "men are promoted faster than women." However, because, as is often the case, we cannot a priori rule out the direction of the relationship (it could be that women are promoted faster than men), we use the customary approach indicating that no relationship exists. If a relationship exists, we later can determine its direction.

Many scholars prefer to state these hypotheses as follows:

H_0: No relationship exists between gender and the rate of promotion in the population.

H_A: A relationship exists between gender and the rate of promotion in the population.

This usage clearly indicates that we are using sample data to draw inferences about relationships in the population. Indeed, we are not interested in our sample per se. Who cares about the preferences of, say, 500 citizens? We care about them only to the extent that their opinions *represent* those of the entire population. In the end, we want to know how the population, not merely a sample of it, thinks about something. We use a sample to infer conclusions about the population. To distinguish conclusions about the sample from those of the population, we use Greek letters to refer to the population. Then, the hypotheses are also written as follows:

H_0: $\mu_m = \mu_f$

H_A: $\mu_m \neq \mu_f$

where μ is the rate of promotion in the population, and the *m* and *f* subscripts stand for "male" and "female," respectively. When we work with sample data, the purpose of hypothesis testing is to test the significance of the relationship in the population.

Statistical Significance

The phrase *statistically significant* often carries considerable weight in public discourse. To say that something is statistically significant is tantamount to throwing the weight of science behind a statement or fact. But what exactly does the phrase mean? ***Statistical significance*** simply refers to the probability of being wrong about stating that a relationship exists when in fact it doesn't. The phrase ***level of statistical significance*** refers to the level of that probability—in other words, *how often* we would be wrong to conclude that a relationship exists when in fact none exists, or how often we would incorrectly reject the null hypothesis when in fact it is true. One reason we might wrongly reject the null hypothesis is that our data are a random sample; had we drawn a different sample, we might have concluded otherwise.

The statistical standard for significance is 5 percent in the social sciences; we are willing to tolerate a 1-in-20 chance of being wrong in stating that a relationship exists (that is, concluding that the null hypothesis should be rejected when in fact it shouldn't). Many researchers also consider a 1-in-100 (1 percent) probability of being wrong as an acceptable standard of significance. The latter is a stricter standard. We are less likely to be wrong stating that a relationship exists (when in fact it doesn't exist) when it is significant at the 1 percent level than when it is significant at only the 5 percent level.

We could set the bar even higher—for example, by choosing a level of significance of one-tenth of 1 percent—but doing so may cause us to conclude that no relationship exists when in fact one does. A standard of less than 1 percent is thus thought to be too risk averse. Why not settle for a 10 percent level of significance? If we did so, we would be accepting a 10 percent chance of wrongfully concluding that a relationship exists when in fact none does. Usually, that is thought to be too risky.[5]

By convention, 5 percent is usually thought to be the uppermost limit of risk that we accept. Thus, relationships that are significant at more than 5 percent (say, 6 percent) are said to be *not significant.* Only relationships that are significant at 5 percent or less are considered significant, and relationships that are significant at 1 percent or less are said to be *highly significant.* Another convention is that most relationships are reported as being significant only at the 1 percent or the 5 percent level. Thus, a relationship that is statistically significant at the 3 percent level is reported as being significant at the 5 percent level but not at the 1 percent level. A relationship that is significant at one-tenth of 1 percent is reported as being significant at the 1 percent level.

Finally, the phrase *level of significance* should not be confused with the term *confidence level.* The confidence level refers to the probability that an unknown population parameter falls within a range of values calculated from the sample, as discussed in Chapter 10. Sometimes the phrase *level of confidence* is taken as

Key Point

The null hypothesis states that no relationship exists. The critical value is the minimum value that a test statistic must be to reject the null hypothesis.

being synonymous with 100 percent minus the level of statistical significance; for example, a 5 percent level of significance is said to be the same as a 95 percent confidence level. However, the phrase *level of significance* should be used in connection with matters of hypothesis testing.

The Five Steps of Hypothesis Testing

Recall the question asked earlier: How large should chi-square be so that we can conclude that a statistically significant relationship exists between gender and year of promotion or, in other words, so that we can reject the null hypothesis and accept the alternate hypothesis? All statistical tests follow the same ***five steps of hypothesis testing***:

1. State the null hypothesis (in Greek letters).
2. Choose a statistical test.
3. Calculate the test statistic (t.s.) and evaluate test assumptions.
4. Look up the critical value (c.v.) of the test.
5. Draw a conclusion:

   If $|\text{t.s.}| < \text{c.v.}$, do not reject the null hypothesis.
   If $|\text{t.s.}| \geq \text{c.v.}$, reject the null hypothesis.

We already discussed the first item and mentioned the second item in the introduction to Section III. Readers also may wish to consult the Statistics Roadmap at the beginning of this book for more detailed guidance on selecting test statistics. We have seen how to calculate the chi-square test statistic. Most statistical tests make assumptions about variables; we will soon address those of the chi-square test statistic. Now we discuss critical values. The ***critical value*** is the minimum value that a test statistic must be in order to rule out chance as the cause of a relationship. Technically, the critical value is the value above which the test statistic is sufficiently large to reject the null hypothesis at a user-specified level of significance.

The following discussion is provided to enhance conceptual understanding because, again, computers do most of the work. The critical value of any test statistic is determined by two parameters: (1) the desired level of statistical significance and (2) the number of degrees of freedom (df). As stated earlier, by convention, analysts are interested in rejecting the null hypothesis at the 1 percent and 5 percent levels. The ***degrees of freedom*** address the practical, statistical problem that the magnitude of most test statistics is affected by the number of observations or categories. For example, the formula for calculating the chi-square test statistic requires us to calculate a value for each cell and then

add them all up. All things being equal, the larger the number of cells, the larger the value of this test statistic. The degrees of freedom statistic controls for this problem.[6] (This also means that it is generally meaningless to compare the values of different chi-square test statistics based on tables of unequal sizes and, as we will soon see, unequal numbers of observations.)

Each type of statistical test has its own way of calculating degrees of freedom. The degrees of freedom for any chi-square test are defined by the formula $(c - 1)^*(r - 1)$, where c is the number of columns in a contingency table and r is the number of rows. In Table 11.1, df = $(2 - 1)^*(4 - 1) = 3$. If our table had six rows and four columns, the number of degrees of freedom would be $[(6 - 1)^*(4 - 1) =] 15$, and so on.

To determine the critical value of our test, we turn to a table of chi-square critical values (see Appendix B). The table shows the levels of significance in columns and the degrees of freedom in rows. Assume that we wish to test whether our previously calculated χ^2 test statistic (8.97) is statistically significant at the 5 percent level. The critical value at this level of significance and three degrees of freedom is shown to be 7.815. Thus, applying the very last step in the method for testing hypotheses, we evaluate the absolute value of 8.97 as indeed larger than the critical value. The absolute value is stated in step 5 because some test statistics, but not χ^2, can have negative values, and because the critical value is always positive. So we conclude that *a relationship exists between gender and the rate of promotion at the 5 percent level of significance.* Alternatively, we can write that *a statistically significant relationship exists between gender and the rate of promotion* ($\chi^2 = 8.97$, $p < .05$). This important language is found in most analyses.

But is this relationship also significant at the 1 percent level? The critical value of this chi-square test at the 1 percent level and three degrees of freedom is 11.341. We evaluate that the absolute value of 8.97 is less than the critical value at this level of significance, and so we conclude that the relationship between gender and years of promotion is significant at the 5 percent level but not at the 1 percent level. We should always identify the highest level of significance, which in this instance is the 5 percent level. But if the test statistics had also been greater than the critical value at the 1 percent level, then the 1 percent level would be concluded.[7]

Note some features of the table of chi-square critical values in Appendix B. First, at any given level of significance, the value of the chi-square critical values increases as the degrees of freedom increase. This is consistent with the problem mentioned earlier: contingency tables with more rows and columns will have larger test statistics simply as a result of having more cells. The degrees of freedom "compensate" for this fact. Second, at any given number of degrees of freedom, the value of the chi-square critical values increases as the level of significance decreases. This, too, makes sense because a 1 percent level of significance will have a higher threshold than a 5 percent level.

Statistical software programs calculate test statistics and report the level of statistical significance at which the test statistic is significant. For example, software output might have shown "p = .029," which indicates that the test statistic is statistically significant at the 5 percent level but not at the 1 percent level. The probability "p = .000" means that the relationship is highly significant, at better than the 1 percent level. The probability "p = .1233" or "p = .9899" indicates that the relationship is not significant. Software programs do not ordinarily report critical values at the 1 percent and 5 percent levels; rather, they show the level of significance at which test statistics are significant. Looking up critical values is a valuable exercise that increases conceptual understanding but one that you will need to do only sporadically.

Here is another example that you can follow to gain additional practice with hypothesis testing. Table 11.4 shows data related to the effectiveness of training for 20 qualified unemployed individuals. The second column ("Training participation") indicates the individuals' participation status, and the third column captures data regarding employment 2 years after the training session. You are asked to conduct a chi-square test to examine the relationship between training participation and employment status. What is your null hypothesis?

The null hypothesis is that there is no relationship between training participation and employment status. The alternate hypothesis is that training participation is related to employment. The calculation of chi-square is shown in Table 11.5.

Getting Started

Replicate these results on your computer.

The value of χ^2 is 5.05. This example has [(2 – 1)* (2 – 1) =] 1 degree of freedom. The critical value at the 0.05 level is 3.841 (see Appendix B). Because χ^2 (= 5.05) is larger than the critical value, we reject the null hypothesis and conclude that training participation is related to employee status at the 5 percent level.

Getting Started

Test whether a relationship exists between two variables of your choice.

Chi-Square Test Assumptions

Nearly all test statistics make *assumptions* about the variables that are used. *Assessing test assumptions is a critical task in statistical testing,* because violations of test assumptions invalidate test results. Analysts need to be familiar with the assumptions of different tests and of ways for addressing violations of test assumptions when they occur. There are three ***chi-square test assumptions***. First, the variables must be categorical, which applies to our variables. Second, the observations are independent, as ours are. ***Independent samples*** are those in which each observation is independent of other observations in the sample. The concept of ***dependent samples*** is discussed more fully in Chapter 12 and typically involves experimental situations such as before-and-after measurement.

Table 11.4 Training Participation and Employment

ID	Training participation	Employment status 2 years after the training session
1	Not participating	Not employed
2	Not participating	Employed
3	Participating	Employed
4	Not participating	Not employed
5	Participating	Employed
6	Not participating	Employed
7	Not participating	Not employed
8	Participating	Employed
9	Not participating	Not employed
10	Participating	Employed
11	Not participating	Not employed
12	Participating	Not employed
13	Participating	Employed
14	Not participating	Not employed
15	Participating	Employed
16	Participating	Employed
17	Participating	Not employed
18	Participating	Not employed
19	Not participating	Not employed
20	Participating	Employed

Table 11.5 Calculating χ^2 for Job Training Performance

	Training participation						Total
	Not participating			Participating			
	Obs.	Exp.	χ^2	Obs.	Exp.	χ^2	χ^2
Not employed	7.00	4.50	1.39	3.00	5.50	1.14	2.53
Employed	2.00	4.50	1.39	8.00	5.50	1.14	2.53
Total	9.00	9.00	2.78	11.00	11.00	2.27	5.05

Note: Obs. = observed count; exp. = expected count.

Third, all cells must have a minimum of five expected observations. When this condition is not met, it is usually because the contingency table contains a large number of rows and columns relative to the number of observations. That is, the data are spread too thinly across too many cells. To correct this problem, simply

redefine the data categories (that is, combine adjacent rows or columns) to create a smaller number of cells. Examination of Table 11.2 shows that our data meet this third assumption, too. The smallest expected frequency count is 6.3. If our data had violated this assumption, we would have combined rows or columns, recalculated results, and reported the revised conclusions. Some analysts, however, feel that this third assumption is too strong.[8]

Although chi-square is useful for testing whether a relationship exists, we have also noted some limitations: chi-square provides no information about the direction or strength of the relationship, and the third assumption may be problematic at times. For this reason, analysts often consider an alternative statistic, Kendall's tau-c, discussed below, which offers information about significance, direction, and strength as well.

Statistical Significance and Sample Size

Most statistical tests are also affected by ***sample size***, which has implications for the likelihood of finding statistically significant relationships. Specifically, it is easier to find statistically significant relationships in large datasets than in small ones. This is more than a statistical artifact; rather, it reflects that having more information makes us more confident of our conclusions, and vice versa. The sample size affects the statistical significance of many widely used test statistics, including chi-square.

Getting Started

All test statistics have assumptions. It is essential to address them.

For example, assume we had a sample of 950 employees, rather than 95 employees, with the same relative distribution as shown in Table 11.1 (see Table 11.6). It is easy to verify that the data in Table 11.6 are distributed in the same exact manner as shown in Table 11.1. But the added observations affect the calculation of the chi-square test statistic. The value of the chi-square test statistic in the first cell is $(O_i - E_i)^2/E_i$, or $[(140 - 100)^2/100 =] 16$. This is exactly *10 times* that of the previously calculated value. Indeed, each cell value is 10 times larger, as is the chi-square test statistic, which now becomes 89.7. Yet the chi-square critic value is still defined as $(c - 1)(r - 1)$. The critical value for rejecting the null hypothesis at the 1 percent level is still 11.341. Whereas previously we could not reject the null hypothesis at this level, we now succeed in doing so by virtue of *having more observations.* This phenomenon occurs with many other widely used test statistics, too.

Of course, the opposite is also true: if we had tried to test for significance using only, say, 20 observations (instead of 95), we would have failed to reject the null hypotheses at even the 5 percent level. This reflects our having too little information to be sufficiently confident in our conclusions. By convention, many researchers prefer to test their null hypotheses on sample sizes of about 100 to a few hundred (say, 400). This is only a rough guideline.

One implication is that analysts are neither surprised to find statistically significant relations in large samples, nor are they surprised to find the lack of statistical significance in small samples. Another implication is that, when working with large samples, analysts can find minute differences between groups to be statistically significant, even when the differences have very little practical relevance. Bigger samples are not necessarily better; they merely increase the importance of questions about the practical significance of findings. Box 11.1 discusses statistical power, which often is used to determine a minimum sample size.

Finally, recall from Chapter 8 that once statistical significance has been established, analysts must turn to the task of establishing practical relevance. Are the differences between categories large or small? Are they large enough to warrant interest from policy makers? Are they large enough to conclude that programs and policy have a salient impact on society? This is the essential task that must follow after statistical hypothesis testing. The descriptive techniques discussed in Chapter 8 regarding the analysis of contingency tables and the use of column percentages are essential to providing these answers, building on the results established here.

Key Point

A larger sample makes it easier to reject the null hypothesis.

The above is the core of hypothesis testing, as illustrated by chi-square, and you will find these principles repeated in subsequent chapters. We know the above is a lot to take in. Still, we extend the above with two useful applications for public managers and analysts which, in their own way, can help consolidate the above through further examples and practice.

Table 11.6 Year of Promotion by Gender: Observed and Expected Counts

	A: Observed counts			B: Expected counts	
	Gender			Gender	
Year	Male	Female	Total	Male	Female
1	140	80	220	100	121
2	160	140	300	136	164
3	70	220	290	131	159
4	60	80	140	63	76
Total	430	520	950	430	520

In Greater Depth . . .

Box 11.1 Power and Sample Size

The level of statistical significance indicates how often we would be wrong to reject the null hypothesis when in fact it is true. However, another possible testing error occurs when we fail to reject the null hypothesis when in fact we should. The former is called a Type I (or α) error, wrongfully concluding that a relationship exists. The latter is called a Type II (or β) error, wrongfully concluding that a relationship does *not* exist.

If β is the probability of wrongfully concluding that a relationship does not exist when in fact it does (Type II error), then 1 – β is the probability of correctly rejecting the null hypothesis when we should. This probability, 1 – β, is called ***statistical power***.

		Decision: Reject	Decision: Accept
Null hypothesis	True	Type I (α) error	Correct
	False	Correct (Power, 1 – β)	Type II (β) error

A typical reason for Type II errors is that the sample size is too small relative to the relationships or differences for which we are testing; we just don't have enough statistical evidence to reject the null hypothesis. The purpose of analyzing power is usually to determine minimum sample size. It has been suggested that the power of tests should be at least .80. Formulas for calculating power vary from test to test; they depend on the sample size, the level of statistical significance, and the effect size (for example, difference between means). Effect size, too, is defined differently for different tests.[9] Typically, analysts use tables or power calculators, many of which are now available on the Internet.[10] Analysts err on the side of caution by postulating small effect sizes (that is, small differences between means or large standard deviations), thereby indicating a need for larger samples.

THE GOODNESS-OF-FIT TEST

Chi-square is most commonly used to determine whether two variables are statistically related to each other. However, an interesting adaptation for managers and analysts is using chi-square for testing whether a program or policy exceeds a standard or norm. Assume we test 400 cars and find a 6 percent failure rate. Is that any different from a norm of 8 percent? You can readily think of other applications, such as students who pass a test or clients who succeed in a treatment, or you might think of water samples, housing, or anything else against a stated norm. What might be relevant in your area of interest?

When using chi-square for this purpose, program or policy outcomes are regarded as the observed frequencies, and the norm is used for calculating expected frequencies. This is called the ***goodness-of-fit test***, which tests whether these two distributions are significantly different (H_a). Let's work through the above example. The above two distributions for the cars example are shown in Table 11.7. The left data column shows the actual frequencies, and the right column shows the expected frequencies that would exist if the actual distribution was exactly consistent with the norm. Specifically, the actual frequencies are [0.06*400 =] 24 failed cars and [0.94*400 =] 376 passed cars. The expected frequencies are [0.08*400 =] 32 failed cars and [0.92*400 =] 368 passed cars. This calculation is different from Table 11.6, because the norm is not really an empirical variable.

The null hypothesis is that the two distributions are similar, and the alternate hypothesis is that they are dissimilar. Using the chi-square formula, $\Sigma(O_i - E_i)^2/E_i$, we calculate chi-square as $(24 - 32)^2/32 = 2.000$ for the failed category and as $(376 - 368)^2/368 = 0.174$ for the passed category. Thus, the chi-square test statistic is [2.000 + 0.174 =] 2.174. The degrees of freedom for this test is defined as the number of rows (r) minus 1, or [2 – 1 =] 1. From Appendix B, the chi-square critical value at the 5 percent level and df = 1 is 3.841. Because the test statistic is less than the critical value, |t.s.| < c.v., we fail to reject the null hypothesis. Hence, we conclude that the failure rate is *not different* from the prespecified norm of 8 percent. The failure rate is neither higher nor lower than the standard; it meets the standard.

The above example can be expanded by considering more than just two response categories such as pass or fail. Assume that we just completed a citizen survey yielding 1,034 responses. We next want to know whether the age distribution of these respondents is consistent with that of the U.S. Census for the area, checking for problems of under- or oversampling. Hence,

H_0: The age distribution of the sample is consistent with that of the population.

H_A: The age distribution of the sample is inconsistent with that of the population.

The results are shown in Table 11.8. Here, the census population frequencies are the expected frequencies, and the sample frequencies are the observed frequencies. With 1,034 completed survey responses, the expected frequency of the 18–45 age category is [1,034*0.623 =] 644. The expected frequencies of the other two categories are, respectively, [1,034*0.241 =] 249 and [1,034*0.136 =] 141; similarly, the observed (actual) frequencies are 649, 277, and 108. Using the usual chi-square formula, we find that the chi-square value for the first category (age 18–45) is $[(649 - 644)^2/644 =] 0.039$. The values for the second and third categories are calculated similarly and are, respectively, 3.149 and 7.723. Thus, the chi-square test statistic is 10.91 (with rounding). The number of degrees of freedom is $r - 1$, or [3 – 1 =] 2. The critical value at the 5 percent level of significance with df = 2 is 5.991 (see Appendix B); thus we conclude that the sample *is significantly different* from the population.[11] Further inspection of Table 11.8 suggests that the researchers undersampled older respondents. Perhaps they might want to reweight their findings to examine the effect or continue surveying older respondents.

The goodness-of-fit test is a useful extension of chi-square for public managers and analysts for comparing a variable against a stated norm.

A NONPARAMETRIC ALTERNATIVE

We now bring you one more piece of useful information. While chi-square is a widely known and popular statistic, it has some limitations. As we have seen, it provides no information about the direction or strength of relationships, and it is limited by some test assumptions (though they are not as cumbersome as some we will see later). Statisticians have developed alternative measures that overcome these limitations. For example, Kendall's tau-c belongs to the family

Table 11.7 Test Failure Rates

	Actual (observed)	Norm (expected)
Passed	376	368
Failed	24	32

Table 11.8 U.S. Census Response by Age Groups

Age	U.S. Census (%)	Survey sample (%)
18–45	62.3	62.8
46–65	24.1	26.8
66+	13.6	10.4

of nonparametric statistics, which derive their name from the fact that they have very few test assumptions. They are a bit less powerful, but quite useful.

Kendall's tau-c can be used as an alternative to chi-square and provides information about the level of significance as well as the strength and direction of relationships.[12] Kendall's tau-c can vary from +1.00 to –1.00. A positive sign indicates a positive relationship and a negative sign indicates a negative relationship (see Chapter 8). A value of zero indicates that no relationship exists between the variables, and a value of |1.00| indicates a perfect relationship. Although there are no absolute standards, many analysts regard scores of less than |0.25| as indicating weak relationships; scores of between |0.25| and |0.50|, moderate relationships; and scores of greater than |0.50|, strong relationships. Beyond this, computers readily calculate the level at which Kendall's tau-c is statistically significant. Thus, this statistic provides three important pieces of information about any relationship: significance, direction, and strength. Another advantage of Kendall's tau-c is that it does not have the third test assumption of chi-square, that all cells must have a minimum of five expected observations. This assumption is unnecessary given the way that Kendall's tau-c is calculated.

Using the data from this chapter's example (see Table 11.1), the computer calculates the value of Kendall's tau-c as .269, which is significant at the .029 level, indicating a positive and moderately strong relationship that is significant at the 5 percent level. However, in this example, the positive sign has no inherent meaning because the variable "gender" is a nominal variable. This example provides a good reminder to interpret outcomes in appropriate (mindful) ways; it is senseless to describe the relationship in Table 11.1 as either a positive or a negative one. Also, the computer-generated value of chi-square is significant at the .016 level; Kendall's tau-c indeed determines the statistical significance of this relationship as a bit less than the chi-square test. But both statistics come to the same conclusion, namely, that this relationship is significant at the 5 percent level. Though Kendall tau-c is a bit less powerful than chi-square, it is a useful alternative as it adds information on the direction and strength of relationships. Appendix 11.2 offers a few more nonparametric statistics that, at times, may be useful.

SUMMARY

When researchers assess the existence and nature of relationships between two variables, hypothesis testing and chi-square applications are invaluable tools. Hypothesis testing is an important step in data analysis because it establishes whether a relationship exists between two variables in the population, that is, whether a relationship is statistically significant. Processes of hypothesis testing involve

1. Stating the null hypothesis
2. Choosing the appropriate test statistics
3. Ensuring that data meet the assumptions of the test statistics
4. Calculating the test statistic values
5. Comparing the test statistic values against critical values and determining at what level a relationship is significant (or relying on the computer to calculate test statistics and to state the level at which they are statistically significant)

When analysts are confronted with two categorical variables, which can also be used to make a contingency table, chi-square is a widely used test for establishing whether a relationship exists (see the Statistics Roadmap at the beginning of the book). Chi-square has three test assumptions: (1) that variables are categorical, (2) that observations are independent, and (3) that no cells have fewer than five expected frequency counts. Remember, violation of test assumptions invalidates any test result. Chi-square is but one statistic for testing a relationship between two categorical variables.

Once analysts have determined that a statistically significant relationship exists through hypothesis testing, they need to assess the practical relevance of their findings. Remember, large datasets easily allow for findings of statistical significance. Practical relevance deals with the relevance of statistical differences for managers; it addresses whether statistically significant relationships have meaningful policy implications.

KEY TERMS

Alternate hypothesis (p. 182)
Chi-square (p. 178)
Chi-square test assumptions (p. 186)
Critical value (p. 184)
Degrees of freedom (p. 184)
Dependent samples (p. 186)
Expected frequencies (p. 179)
Five steps of hypothesis testing (p. 184)
Goodness-of-fit test (p. 191)
Independent samples (p.186)
Kendall's tau-c (p.193)
Level of statistical significance (p. 183)
Null hypothesis (p. 181)
Purpose of hypothesis testing (p. 180)
Sample size (and hypothesis testing) (p. 188)
Statistical power (p. 190)
Statistical significance (p. 183)

APPENDIX 11.1: RIVAL HYPOTHESES: ADDING A CONTROL VARIABLE

We now extend our discussion to rival hypotheses. The following is but one approach (sometimes called the "elaboration paradigm"), and we provide other

(and more efficient) approaches in subsequent chapters. First mentioned in Chapter 2, *rival hypotheses* are alternative, plausible explanations of findings. We established earlier that men are promoted faster than women, and in Chapter 8 (see "Pivot Tables") we raised the possibility that the promotion rate is different between men and women because men are more productive than women. We can now begin to examine this hypothesis formally using chi-square. Again, managers will want to examine this possibility among several.

Assume that we somehow measured productivity. Variables associated with rival hypotheses are called control variables. The control variable "productivity" is added to our dataset. To examine the rival hypothesis, we divide the sample into two (or more) groups, namely, employees with high productivity and those with low productivity. For each of these groups, we make a contingency table analysis by gender. If it is true that productivity, and not gender, determines the rate of promotion, then we expect to find no differences in the rate of promotion within the *same* level of productivity (high or low) because the differences exist across levels of productivity, and not by gender. Next, we construct a table (see Table 11.A1.1). Note that the control variable "goes on top." We still have a total of 95 employees, 43 of whom are men and 52 of whom are women. For simplicity, and to avoid violating chi-square test assumptions (we must maintain a minimum of five expected frequencies in each cell), the variable "year of promotion" has been grouped, although this needn't be done in other instances. The relevant hypotheses are now as follows:

$H1_0$: No relationship exists between gender and rate of promotion among employees with high productivity.

$H1_A$: A relationship exists between gender and rate of promotion among employees with high productivity.

$H2_0$: No relationship exists between gender and rate of promotion among employees with low productivity.

$H2_A$: A relationship exists between gender and rate of promotion among employees with low productivity.

Chi-square test statistics are calculated for *each* of the two different productivity groups. We could find that one or both relationships are now statistically significant. When both relationships are not statistically significant, the result is called an *explanation* of the initial findings; that is, the statistically significant result has been explained away. Sometimes it is said that the previous relationship has proven to be *spurious.* When both relationships are statistically significant, the result is called a *replication* of the initial findings. When only one of the relationships is statistically significant, the result is called a *specification* of the initial findings. We would want to examine further the relationship that is not

explained away. Finally, rarely does using a control variable result in uncovering statistically significant relationships that are otherwise insignificant. When this does occur, however, the result is called a *suppressor effect*; that is, the existing relationship is suppressed in the absence of the control variable.

Through our data, we obtain the following results. The chi-square test statistic for the relationship between gender and year of promotion among employees with low productivity is 2.39, which is not statistically significant ($p = .117$). Thus, we conclude that gender does *not* discriminate in the rate of promotion among employees with low levels of productivity. But the chi-square test statistic for the relationship between gender and year of promotion among employees with high productivity is 6.65, which is statistically significant at the 1 percent level ($p = .010$). Gender differences continue to explain differences in the rate of promotion among employees with high levels of productivity. This type of finding is called a *specification*.[13]

Although this approach allows us to test rival hypotheses, two limitations may be noted: results are sometimes inconclusive (for example, in the case of specification), and the added cells require a larger number of observations. Table 11.A1.1 acknowledges this problem; rows were combined. In Chapter 15 we discuss multiple regression as an alternative for continuous dependent variables, which is a much more commonly used and efficient approach than discussed here.

APPENDIX 11.2: NONPARAMETRIC TESTS FOR SPECIFIC SITUATIONS

Kendall's tau-c is one of several nonparametric tests that computer programs routinely compute. Nonparametric tests have very few test assumptions, and many are based on the idea of proportional reduction in error (PRE), the improvement that is expressed as a fraction, in predicting a dependent variable

Table 11.A1.1 Year of Promotion by Gender: Controlling for Productivity

	Low productivity		High productivity	
	Gender		Gender	
Year	Male (%)	Female (%)	Male (%)	Female (%)
1–2 Years	47	22	85	52
3+ Years	53	78	15	47
Total	100	100	100	100
(*n* =)	(17)	(18)	(26)	(34)

due to knowledge of the independent variable.[14] This results in a measure of strength and direction of the relationship as described in the text. A plethora of PRE and nonparametric statistics exist. Computers often calculate ***Gamma* (γ), *Somers' d, Kendall's tau-b* (τ_b),** and *Kendall's tau-c* (τ_c), which are all PRE-based, nonparametric alternatives to chi-square. They typically produce quite similar results, but Kendall's tau-c is the most conservative (lower p-values) and hence most commonly used, and some of these have quite specific uses; Kendall's tau-b is only for square tables, and ***Goodman-Kruskal's tau*** and ***lambda*** are only for nominal level variables.

Various PRE- and nonparametric statistics have been developed to address some rather specific situations that public managers and analysts may face. For example, evaluators are sometimes used to assess program or agency performance. Then, the ***Kruskal-Wallis H** test* assesses whether programs differ in their ratings. Assume that 15 evaluators are each asked to evaluate one of three programs, and an index score is constructed of their evaluations. Each evaluator evaluates a different program and the null hypothesis is that, on average, each program has the same average ranking; the program ratings are not different. The data are shown in Table 11.A2.1 (for presentation, the variables are shown in separate columns, but the data are entered in statistical software programs as two variables and 15 observations only). The rating is a continuous variable, but Kruskal-Wallis *H* assigns ranks to the rating variable (thus creating an ordinal variable from the continuous variable). The computer calculated H test statistic for these data (which has a chi-square distribution) is 7.636 ($df = 2$, $p = .022 < .05$).[15] Thus, the three programs do have different mean rankings. Information provided with this result shows that the mean rankings are, respectively, 4.70, 7.00, and 12.30.

A variation on the above occurs when several evaluators assess different (multiple) program items, and we want to know whether evaluators agree in their ratings. This is an example of *dependent* sample, defined as samples or respondents that are connected or matched up in some way (for more on this, see Chapter 12, "T-test Assumptions"). Dependent samples typically involve repeated measures (the case here) or matched subjects; statistics have been

Table 11.A2.1 Ratings of Three Programs

Program	Rating	Rank	Program	Rating	Rank	Program	Rating	Rank
1	2.5	3	2	3.4	7.5	3	4.8	13
1	2.9	4	2	3.3	6	3	5.0	14.5
1	4.0	10.5	2	4.0	10.5	3	5.0	14.5
1	3.2	5	2	3.9	9	3	3.4	7.5
1	1.2	1	2	2.1	2	3	4.2	12
Mean		4.7			7.0			12.3

developed for these scenarios. The *Friedman test*, developed by the well-known economist Milton Friedman, addresses the above case. Considering the data of Table 11.A2.2, the computer-calculated Friedman test statistic for these data is 0.545 (df = 2, p = .761 > .05). Thus, we conclude that the ratings of the evaluators are not different; *the evaluators agree with each other.* When columns and rows are reversed, the Friedman test assesses whether differences exist among the mean rankings of items. This test can also be used to examine test score changes in before-and-after situations. Then, the rows are subjects and the columns are the subjects' before-and-after scores.[16] Hence, it is a quite versatile test.

Managers and analysts may also need to assess whether discrimination is occurring in programs or policies. Assume that we want to test whether program staff is discriminating against minority clients by failing to provide them with services that are provided to other, white clients. To examine this possibility, we match up pairs of minority and white clients; each pair has similar equivalent conditions and are trained to provide similar responses to questions; the main difference is race. This strategy is used in testing for discrimination in employment interviews or in bank lending practices; pairs of majority and minority job (or loan) seekers are sent to interviews (or to apply for loans), intermingled with other candidates. This scenario illustrates matched subjects, hence, again involving a dependent sample.

The *McNemar test* determines the level at which dissimilar outcomes are statistically significant. For example, consider Table 11.A2.3, in which each count compares the employment outcomes of the paired testers. The McNemar

Table 11.A2.2 Ratings of Three Evaluators

Item	Rater 1	Rater 2	Rater 3
1	5	3	4
2	4	2	2
3	3	3	3
4	2	4	3
5	1	1	1

Table 11.A2.3 Employment Discrimination Test

	White applicants		
Minority applicants	Hired	Not hired	Total
Hired	0	1	1
Not hired	8	2	10
Total	8	3	11

test compares whether the eight instances in which a white but not minority applicant received a job are significantly different from the one instance in which the minority candidate, not the white one, received the job. The test for these data is significant ($p = .039 < .05$), which means that this disparate outcome cannot be attributed to chance alone.[17] The McNemar test is also an example of a test designed for small samples (that is, those with small frequency counts).

Finally, a variety of small-sample tests exist for independent samples. Although chi-square can be used, researchers also argue for the *Fisher exact test* for 2-by-2 tables and the ***Chi-square with the Yates continuity correction.*** Small samples bias the expected frequencies slightly upward; the Yates continuity correction corrects this bias by subtracting 0.50 from the difference of expected and observed frequencies, while this produces a conservative test statistic, some argue that this correction overcorrects.[18,19]

These examples show how rather specialized statistics can be adapted to the practice of management and policy. The analytical task in these instances is to align one's problem with some method of statistical testing.

Notes

1. The companion website (http://study.sagepub.com/bermaness4e) replicates these calculations on an Excel spreadsheet called "Chi-Square." The computer-calculated value of chi-square is slightly higher, 9.043, due to rounding errors in calculating the expected frequency counts. This same result is achieved when using expected frequency counts with three decimal places. The expected frequency counts, then, for men are 9.976, 13.588, 13.115, and 6.321; and for women 12.064, 16.432, 15.860, and 7.644. Of course, maintaining three decimal places is more labor intensive for the illustrative, manually calculated example in the text, which retains only one decimal place in calculating the expected frequencies.
2. See "Pearson's Chi-Squared Test," https://www.khanacademy.org/math/probability/statistics-inferential/chi-square/v/pearson-s-chi-square-test-goodness-of-fit.
3. See Chapter 8 for a discussion of the direction of relationships.
4. It is commonly said that inferential statistics state the degree of certainty by which we can say that a relationship exists beyond chance alone. This is plainly said, and some academics will take issue with how this is phrased. People are free to make their own plain-sense interpretation of statistical formulas.
5. Such a level might be acceptable at times in administration, and scientists occasionally report a 10 percent level, too.
6. The concept of degrees of freedom is not easy to explain. Some texts explain it as the number of calculations that are not predetermined after others have already occurred. Succinctly, if an array (or column) has four data elements,

and the sum total is also known, then after choosing the first three elements, the fourth element is predetermined; hence, we are free to choose only three elements, and the array is said to have three degrees of freedom, or $c - 1$.

7. For example, if the test statistic of our data had been, say, 15.0, then $p < .01$ rather than $p < .05$ would be concluded and reported. This is not the case here, though.
8. The rationale is to ensure that chi-square calculations are not unduly affected by small differences in cells with low counts. Note that the expected frequency is in the denominator of the chi-square formula. Some analysts feel that the standard of no cells with expected frequencies below 5.0 is too strict. They feel that (1) all cells should have greater expected frequency counts than 1.0 and (2) no more than 20 percent of cells should have expected frequency counts lower than 5.0. The standard adopted in the text is more conservative. The point is, of course, that test statistics should not be affected by a few sparse cells.
9. This is defined as $(\mu_1 - \mu_2)/\sigma_{pooled}$, where $\sigma_{pooled} = \sqrt{[(\sigma_1^2 + \sigma_2^2)/2]}$. Small effect sizes are defined as those for which $\mu_1 - \mu_2$ is about .2 σ_{pooled}, medium effect sizes are about .5 σ_{pooled}, and large effect sizes are .8 σ_{pooled}. For a chi-square test, effect size is defined as the *Phi coefficient*, z, for two-by-two tables, $\sqrt{(\chi^2/N)}$, and as the contingency coefficient, C, for larger tables, $\sqrt{[\chi^2/(\chi^2 + N)]}$. Some of these measures are discussed in later chapters.
10. For example, see http://www.dssresearch.com/toolkit/spcalc/power.asp or http://power.education.uconn.edu/otherwebsites.htm.
11. Note that if the sample had consisted of only 300 completed responses, then the chi-square would have been 3.16, which is not significant. Also, you can verify, by redoing the preceding calculations (using a spreadsheet), that completing another 12 surveys among the 66+ group (increasing the sample size to 1,046) reduces the chi-square test statistic to 5.955, which provides a sample that no longer is significantly different from that of the population.
12. The formula for Kendall's tau-c is quite different from chi-square and is a based on a concept of 'similar' and 'dissimilar' pairs of data: the formula is $2m(Ns - Nd)/N^2(m - 1)$, where m is the smaller number of rows or columns, N is the sample size, Ns is the number of similar pairs and Nd is the number of dissimilar pairs. In short, *similar pairs* are pairs of observations that each rank similarly low (or high), and *dissimilar pairs* are pairs in which one observation scores high and the other low. The direction of relationships is determined by comparing the number of similar pairs against the number of dissimilar pairs. When there are more similar pairs than dissimilar pairs, the relationship is said to be positive. When there are more dissimilar pairs than similar pairs, the relationship is negative.
13. This approach is rather inefficient. Note that we had to combine categories in order to preserve an adequate number of observations in each cell. In subsequent chapters, we examine approaches that are more efficient and more conclusive. Of course, when productivity is found to cause explanation

or specification, you subsequently want to report on the bivariate relationship between the rate of promotion and productivity. That, of course, is a different relationship from the one discussed here.

14. PREs can be calculated in different ways, and the following is purely illustrative of the concept. Assume that in a sample of 160 people, 90 people are not on welfare and 70 are on welfare. If we guess that each person is not on welfare, we will be wrong 70 times, which is better than guessing that each person is on welfare, in which case we would be wrong 90 times. However, if the number of welfare recipients at each level of income is also known, then we can make even fewer wrong guesses (or, errors). Assume that 60 of 100 welfare program participants also have low incomes; we then make only 40 errors guessing that people with low income also receive welfare. Likewise, if 10 of 60 people with high incomes also are known to get welfare, we make 10 errors guessing that those with high incomes are not on welfare. The total number of errors taking income into account is thus (40 + 10 =) 50, which is slightly less than the earlier 70 wrong guesses. The proportional reduction in errors (wrong guesses) can be defined as follows:

$$\frac{\left(\begin{array}{l}\text{Errors without knowledge of}\\ \text{the independent variable}\end{array}\right)-\left(\begin{array}{l}\text{Errors with knowledge of}\\ \text{the independent variable}\end{array}\right)}{\text{Errors without knowledge of the independent variable}}$$

Thus, our PRE is [(70 – 50)/70 =] 0.286. In other words, as a result of knowing respondents' incomes, we improved our guesses of their welfare situation by 28.6 percent. As discussed in the text, this is evidence of a moderately strong relationship.

15. The formula for H is

$$\frac{12}{n(n+1)}\left(\frac{T_2^2}{n_1}+\frac{T_1^2}{n_1}+\ldots\right)-3(n+1),$$

where T_i is the sum of ranks in group 1, and so on.

16. The Friedman test is quite sensitive to the number of items; it is best to have at least 10 rows.
17. The McNemar test statistic is defined as $X^2_{McNemar} = (|f_{0,1} - f_{1,0}|)^2/(f_{0,1} + f_{1,0})$.
18. The chi-square statistic with Yates continuity correction is defined as

$$\sum_i \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

19. The following measures may also be considered. Phi (ϕ) is defined as $\sqrt{X^2/n}$, and ranges from zero to one for two-by-k tables ($k \geq 2$). Phi-squared (ϕ^2) has a "variance-explained" interpretation; for example, a ϕ^2 value of 0.35 (or ϕ = 0.59) means that 35 percent of the variance in one variable is explained by the other. Yule's Q is a measure of association with a PRE interpretation but without a test of statistical significance.