6.1 The Chi-Square Distribution

chi-square (χ^2) distribution

The **chi-square (χ^2)** distribution is the distribution defined by

$$f(\chi^2) = \frac{1}{2^{k/2}\Gamma(k/2)}\chi^{2[(k/2)-1]}e^{-(\chi^2)/2}$$

This is a rather messy-looking function and most readers will be pleased to know that they will not have to work with it in any arithmetic sense. We do need to consider some of its features, however, to understand what the distribution of χ^2 is all about. The first thing that should be mentioned, if only in the interest of satisfying healthy curiosity, is that the term $\Gamma(k/2)$ in the denominator, called a **gamma function,** is related to what we normally mean by *factorial.* In fact, when the argument of gamma $(k/2)$ is an integer, then $\Gamma(k/2) = [(k/2) - 1]!$. We need gamma functions in part because arguments are not always integers. Mathematical statisticians have a lot to say about gamma, but we'll stop here.

gamma function

A second and more important feature of this equation is that the distribution has only one parameter (k). Everything else is either a constant or the value of χ^2 for which we want to find the ordinate $[f(\chi^2)]$. Whereas the normal distribution was a two-parameter function, with μ and σ as parameters, χ^2 is a one-parameter function with k as the only parameter. When we move from the mathematical to the statistical world, k will become our degrees of freedom. [We often signify the degrees of freedom by subscripting χ^2. Thus, χ^2_3 is read "chi-square with three degrees of freedom." Alternatively, some authors write it as $\chi^2(3)$.]

Figure 6.1 shows the plots for several different χ^2 distributions, each representing a different value of k. From this figure it is obvious that the distribution changes

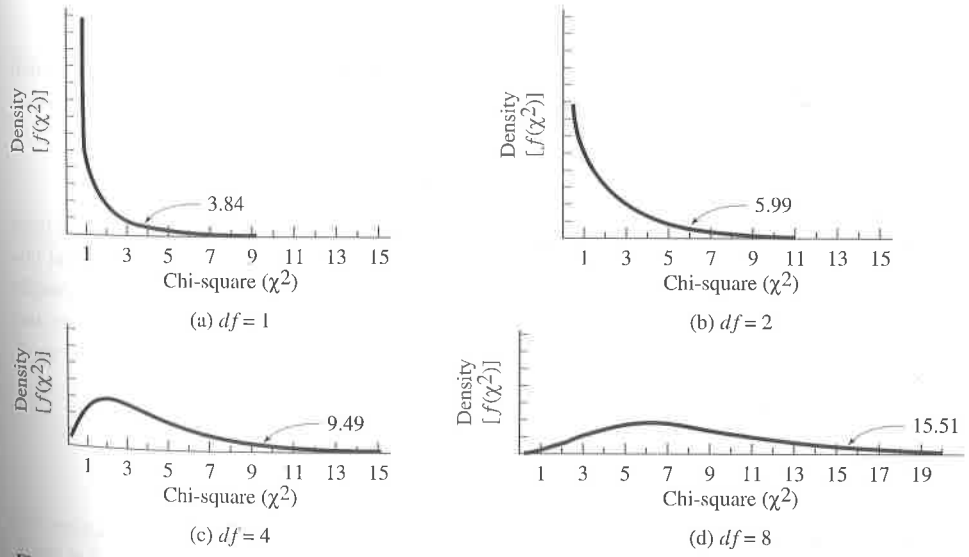

FIGURE 6.1 *Chi-square distributions for df = 1, 2, 4, and 8. (Arrows indicate critical values at alpha = 0.05.)*

markedly with changes in k, becoming more symmetric as k increases. It is also apparent that the mean and variance of each χ^2 distribution increase with increasing values of k and are directly related to k. It can be shown that in all cases

$$\text{Mean} = k$$
$$\text{Variance} = 2k$$

6.2 Statistical Importance of the Chi-Square Distribution

The χ^2 distribution is a mathematical distribution that exists independently of any particular set of statistical procedures. It would have no place in this book, however, if it was not somehow relevant to at least one statistical procedure. In fact, the χ^2 distribution is related to many of the statistics to be covered in this book, of which the chi-square test is only one. While the relationships that we are going to discuss in the next two sections may look like, and perhaps are, things that you could probably do without, they are important simply because they have ties to other material that we will cover. They are the kind of thing to keep in the back of your mind, not the stuff you use on a daily basis.

Chi-Square and z

Assume we have a normal population with known mean (μ) and variance (σ^2). From this distribution we will sample one observation (X), calculate

$$z^2 = \frac{(X - \mu)^2}{\sigma^2}$$

record z^2, and repeat the procedure an infinite number of times. Assume we finish this task; we can then plot the distribution of z^2, and we will find that this distribution looks exactly like the χ^2 distribution on 1 *df*. In fact,

$$\chi^2_1 = z^2$$

To carry the example further, assume that instead of sampling one score at a time from our normal population, we sampled N scores at a time. For each individual observation we calculated z^2, and then we calculated $\sum z^2$, summing over the sample of N observations. Again we repeat this procedure an infinite number of times and plot the resulting values of $\sum z^2$. The resulting distribution will be distributed as χ^2 on N *df*:

$$\chi^2_N = \sum_{i=1}^{N} z_i^2 = \sum \frac{(X_i - \mu)^2}{\sigma^2}$$

Since we have just seen that z_i^2 is itself distributed as χ^2, this last equation reveals an important property of χ^2: The sum of N independent values of χ^2 is its

distributed as χ^2. In this case, the degrees of freedom will equal the sum of the degrees of freedom for the separate χ^2s.

Two important restrictions on the previous two equations are the requirements that the observations be sampled independently of one another and that they be sampled from a normal population. These restrictions will be referred to again when we discuss the chi-square test.

The usefulness of these relationships probably is not immediately apparent, but a full discussion of their value must await a discussion of the chi-square test. The most important point to be made here is that the last two equations tell us what would happen *if* we were to draw an infinite number of samples under the specified conditions. In other words, we do not ever have to take on the Herculean task of drawing vast numbers of samples, because we already know what the resulting distribution would look like. This is true of most of the sampling distributions we will discuss in this book.

Chi-Square and Variance

One of the important but little discussed uses of χ^2 deals with its relationship to population and sample variances. This relationship forms the basis of many of our most important statistical tests and helps to explain several of the restrictions we place on the use of these tests.

Assume for the moment that we have a normally distributed population with a known variance (σ^2). From this population, we draw an infinitely large number of samples of N observations each and calculate the sample variance (s^2) for each sample. We could then plot the sampling distribution of the variance—the distribution of sample variances. You can see such a distribution plotted in Figure 7.4 (p. 184). This distribution would bear a direct linear relationship to the distribution of χ^2:

$$\chi^2_{N-1} = \frac{(N-1)s^2}{\sigma^2}$$

Turning this around, we have

$$s^2 = \chi^2 \frac{\sigma^2}{N-1}$$

Since the fraction $\frac{\sigma^2}{N-1}$ is a constant for any given population variance and sample size, the distribution of s^2 will differ by only a constant from the distribution of χ^2. In other words, the sampling distribution of the variance has a χ^2 distribution on $N-1$ *df*. Keep in mind that we have assumed that we are dealing with normal distributions. (For a more thorough treatment of this topic see Hays, 1994, pp. 350–358.)

We have just seen something concerning the sample variance that will become important later. We already know that s^2 is an unbiased estimate of σ^2, meaning that, *on the average,* s^2 will equal σ^2. We now know that the distribution of s^2 resembles that of χ^2, which we know to be skewed. This means that although the average s^2 will equal σ^2, more than half the time our particular value of s^2 will be smaller than σ^2, no matter how things average out in the end. This fact will have important implications when we discuss t tests in Chapter 7.

6.3 The Chi-Square Goodness-of-Fit Test—One-Way Classification

chi-square test

So far we have confined our discussion to the χ^2 distribution, examining its relationship to certain important statistics. We now turn to what is commonly referred to as the **chi-square test,** which is based on the χ^2 distribution. We will first examine the test as it is applied to one-dimensional tables and then as applied to two-dimensional tables (contingency tables).

The following example is based on one of the most famous experiments in animal learning, conducted by Tolman, Ritchie, and Kalish (1946). At the time of the original study, Tolman was engaged in a theoretical debate with Clark Hull and the latter's students on whether a rat in a maze learns a discrete set of motor responses (Hull) or forms some sort of cognitive map of the maze and responds on the basis of that map (Tolman). At issue was the fundamental question of whether animals learn by stimulus–response conceptions or whether there is room for a cognitive interpretation of animal behavior. (To put this in less academic language, "Do animals think?") The statistical test in question is called a **goodness-of-fit test** because it asks whether there is a "good fit" between the data (observed frequencies) and the theory (expected frequencies).

goodness-of-fit test

In a simple and ingenious experiment, Tolman and his colleagues first taught a rat to run down a starting alley of a maze into a large circular area. From the circular area another alley exited straight across from the entrance but then turned and ended up in a goal box, which was actually to the right of the circular area. After the rats had learned the task ("go to the circular area and exit straight across"), Tolman changed the task by making the original exit alley a dead end and by adding several new alleys, one of which pointed in the direction of the original goal box. Thus, the rat had several choices, one of which included the original alley and one of which included a new alley that pointed directly toward the goal. The maze is shown in Figure 6.2, with the original exit alley drawn with solid lines and the new alleys drawn

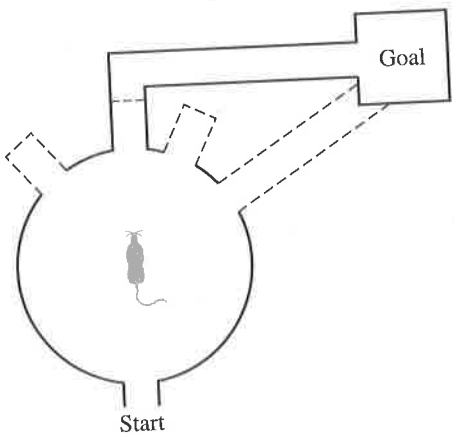

FIGURE 6.2 *Schematic diagram of Tolman's maze*

with dotted lines. If Hull was correct, the rat would learn a stimulus–response sequence during the first part of the experiment and would therefore continue to make the same set of responses, thus entering the now dead-end alley. If Tolman was right and the rat learned a cognitive map of the situation, then the rat would enter the alley on the *right* because it knew that the food was "over there to the right." Since Tolman was the one who published the study, you can probably guess how it came out—the rats chose the alley on the right more often than the others. But we still need some way of testing whether the preference for the alley on the right was due to chance (the rats entered the alleys at random) or whether the data support a general preference for the right alley. Do the data represent a "good fit" to a random choice model? Tolman certainly hoped not, because he wanted to show that the rats had learned something.

	Alley Chosen			
	A	B	C	D
Observed	4	5	8	15
Expected	8	8	8	8

It certainly looks as if animals were choosing Alley D much more than the others, which is what Tolman expected, but how can we be sure?

observed frequencies

expected frequencies

The most common and important formula for χ^2 involves a comparison of observed and expected frequencies. The **observed frequencies,** as the name suggests, are the frequencies you actually observed in the data—the numbers in the table above. The **expected frequencies** are the frequencies you would expect *if the null hypothesis were true.* We want to test the null hypothesis that rats enter alleys at random. In this case we have 32 rats, each making independent choices. (If we used the same four rats 32 times, we would probably have strong reservations about this assumption of independence.) Since we have four alleys, if the rats are responding at random, rather than on the basis of what they have learned about the maze, we would expect that ¼ of them would enter each alley. That means that we would expect frequencies of 8 for each alley. Instead we got frequencies of 4, 5, 8, and 15. The standard formula for the chi-square test looks at the difference between these observed and expected frequencies.

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

The above formula should make a certain amount of intuitive sense. Start with the numerator. If the null hypothesis is true, the observed and expected frequencies (O and E) would be reasonably close together and the numerator would be small, even after it is squared. Moreover, how large the difference between O and E would be ought to depend on how large a number we expected. If we were talking about 1000 animals entering each alley, an $O - E$ difference of 5 would be trivial. But if we expected 8 animals to enter each alley, an $O - E$ difference of 5 would be substantial. To keep the squared size of the difference in perspective relative to the number of observations we expect, we divide the former by the later. Finally, we sum over all of the alleys to combine these relative differences. (If you wonder why we square the numerator, work out what would happen with these, or any other data, if we did not.)

First I will go ahead and calculate the χ^2 statistic for these data using the observed and expected frequencies given in the table.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$
$$= \frac{(4-8)^2}{8} + \frac{(5-8)^2}{8} + \frac{(8-8)^2}{8} + \frac{(15-8)^2}{8}$$
$$= 9.25$$

The Tabled Chi-Square Distribution

Now that we have obtained a value of χ^2, we must refer it to the χ^2 distribution to determine the probability of a value of χ^2 at least this extreme. We can do this through the use of the standard tabled distribution of χ^2.

tabled distribution of χ^2

The **tabled distribution of χ^2,** like that of most other statistics, differs in a very important way from the *tabled* standard normal distribution, as was pointed out in Chapter 3. We will use a simple illustration. Consider the distribution of χ^2 for 1 *df* shown in Figure 6.1. Although it is certainly true that we could construct a table of exactly the same form as that for the standard normal distribution, allowing us to determine what percentage of the values are greater than any arbitrary value of χ^2, this would be tremendously time-consuming and wasteful. We would have to make up a new table for every reasonable number of degrees of freedom. It is not uncommon to want up to 30 *df,* which would require 30 separate tables, each the size of Appendix z. Such a procedure would be particularly wasteful since most users would need only a small fraction of each of these tables. If we wish to reject H_0 at the 0.05 level, all that we really care about is whether or not our value of χ^2 is greater or less than the value of χ^2 that cuts off the upper 5% of the distribution. Thus, for our particular purposes, all we need to know is the 5% cutoff point for each *df.* Other people might want the 2.5% cutoff, 1% cutoff, and so on, but it is hard to imagine wanting the 17% values, for example. Thus, tables of χ^2 such as the one given in Appendix χ^2, which is reproduced in part in Table 6.1, are designed to supply only those values that might be of general interest.

TABLE 6.1 Upper percentage points of the χ^2 distribution

df	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	**7.82**	9.35	11.35	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.54	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.66	23.59
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Look for a moment at Table 6.1. Down the leftmost column you will find the degrees of freedom. In each of the other columns, you will find the critical values of χ^2 cutting off the percentage of the distribution labeled at the top of that column. Thus, for example, you will see that for 3 *df* a χ^2 of 7.82 cuts off the upper 5% of the distribution. (Note the boldfaced entry in Table 6.1.)

Returning to our example, we have found a value of $\chi^2 = 9.25$ on 3 *df*. We have already seen that, with 3 *df*, a χ^2 of 7.82 cuts off the upper 5% of the distribution. Since our obtained value (χ^2_{obt}) = 9.25 is greater than $\chi^2_{0.05} = 7.82$, we reject the null hypothesis and conclude that the obtained frequencies differed from those expected under the null hypothesis by more than could be attributed to chance.[2] In other words, Tolman's rats were not behaving randomly—they look as if they knew what they were doing.

6.4 Two Classification Variables: Contingency Table Analysis

contingency table

In the previous example we considered the case in which data are categorized along only one dimension (classification variable). Often, however, data are categorized with respect to two (or more) variables, and we are interested in asking whether those variables are independent of one another. To put this in the reverse, we often are interested in asking whether the distribution of one variable is *contingent* on a second variable. In this situation we will construct a **contingency table** showing the distribution of one variable at each level of the other. An excellent example is offered by a study by Pugh (1983) on the "blaming the victim" phenomenon in prosecutions for rape.

Pugh conducted a thorough and complex study examining how juries come to decisions in rape cases. He examined a number of variables, but we will collapse two of them and simply look at his data in terms of (1) whether the defendant was found guilty, and (2) whether the defense alleged that the victim was somehow partially at fault for the rape. Pugh's actual data are presented in Table 6.2 in the form of a contingency table.

TABLE 6.2 Pugh's data on decisions in rape cases

	Verdict		
Fault	Guilty	Not Guilty	Total
Low	153 (127.559)	24 (49.441)	177
High	105 (130.441)	76 (50.559)	181
Total	258	100	358

[2]Notice that here the subscripts for χ^2 (i.e., obt and 0.05) do not refer to the degrees of freedom, but designate either the obtained value of χ^2 [χ^2_{obt}] or the value of χ^2 that cuts off the largest 5% of the distribution [$\chi^2_{0.05}$]. When we wish to designate both the degrees of freedom and the level of alpha we write something like $\chi^2_{0.05}(1) = 3.84$.

This table shows some evidence that jurors assign guilt partly on the basis of the perceived faults of the victim. Notice that when the victim was portrayed as low in fault, approximately 86% (153/177) of the time the defendant was found guilty. On the other hand, when the victim was portrayed as high in fault, the defendant was found guilty only 58% (105/181) of the time.

Expected Frequencies for Contingency Tables

marginal totals
cell
row total
column total

For a contingency table the expected frequency for a given cell is obtained by multiplying together the totals for the row and column in which the cell is located and dividing by the total sample size (N). (These totals are known as **marginal totals,** because they sit at the margins of the table.) If E_{ij} is the expected frequency for the **cell** in row i and column j, R_i and C_j are the corresponding **row** and **column totals,** and N is the total number of observations, we have the following formula:[3]

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

For our example

$$E_{11} = \frac{177 \times 258}{358} = 127.559$$

$$E_{12} = \frac{177 \times 100}{358} = 49.441$$

$$E_{21} = \frac{181 \times 258}{358} = 130.441$$

$$E_{22} = \frac{181 \times 100}{358} = 50.559$$

Calculation of Chi-Square

Now that we have the observed and expected frequencies in each cell, the calculation of χ^2 is straightforward. We simply use the same formula that we have been using all along, although we sum our calculations over all cells in the table.

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

$$= \frac{(153-127.559)^2}{127.559} + \frac{(24-49.441)^2}{49.441} + \frac{(105-130.441)^2}{130.441} + \frac{(76-50.559)^2}{50.559}$$

$$= 35.93$$

[3]This formula for the expected values is derived directly from the formula for the probability of the joint occurrence of two *independent* events given in Chapter 5 on probability. For this reason the expected values that result are those that would be expected if H_0 were true and the variables were independent. A large discrepancy in the fit between expected and observed would reflect a large departure from independence, which is what we want to test.

Degrees of Freedom

Before we can compare our value of χ^2 to the value in Appendix χ^2, we must know the degrees of freedom. For the analysis of contingency tables, the degrees of freedom are given by

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

where

R = the number of rows in the table

and

C = the number of columns in the table

For our example we have $R = 2$ and $C = 2$; therefore, we have $(2 - 1)(2 - 1) = 1$ *df*. It may seem strange to have only 1 *df* when we have four cells, but you can see that once you know the row and column totals, you need to know only one cell frequency to be able to determine the rest.

Evaluation of χ^2

With 1 *df* and $\alpha = 0.05$, the critical value of χ^2, as found in Appendix χ^2, is 3.84. Because our value of 35.93 exceeds the critical value, we will reject the null hypothesis that the variables are independent of each other. In this case we will conclude that whether a defendant is found guilty depends in part on whether the victim is portrayed by the defending lawyer as being at fault for the rape. How do these results fit with how you think you would judge the case?

Correcting for Continuity

Yates' correction for continuity

Many books advocate that for tables with only two rows and two columns (2×2 tables) such as Table 6.2, we should employ what is called **Yates' correction for continuity,** especially when the expected frequencies are small. (The correction merely involves reducing the absolute value of each numerator by 0.5 unit before squaring.) There is an extensive literature on the pros and cons of Yates' correction, with firmly held views on both sides. Although I recommend that you not use the correction in most cases, it is important to understand what all the fuss is about.

A glance back at Figure 6.1 will show that χ^2 is a continuous function, with all values of χ^2 along the abscissa possible. If you go back to Pugh's example, however, you will discover that there is no way to rearrange the data, *keeping the marginal totals constant,* to come up with a slightly different χ^2 (e.g., 35.82). With finite sample sizes the obtained distribution of χ^2 for the chi-square test is discrete (especially for small samples), whereas the theoretical χ^2 distribution is continuous. This leads to a certain mismatch when we attempt to evaluate a χ^2 statistic against the χ^2 distribution.

The mismatch led Yates (1934) to devise a correction to be applied in the case of a 2×2 contingency table. The result of this correction is that the probability of the

corrected χ^2, taken from χ^2 tables, is quite close to the *true* probabilities calculated on the basis of the individual probabilities of all possible tables *with those marginal totals* (the R_i and C_j). Yates' correction in fact accomplishes his goal quite nicely, and it is recommended whenever it makes sense to calculate a probability given that the marginal totals really are fixed. Unfortunately, to speak about **fixed marginals** implies that if you repeated the experiment, the individual cell totals might well change but you would arrive at the same marginal totals.[4] This situation rarely exists, and when it does not exist it makes little sense to ask what the probability of the data would be if we assumed that it did. For this reason (the unreasonableness of a fixed-marginal assumption), many papers have argued against the use of Yates' correction (Bradley, Bradley, McGrath, and Cutcomb, 1979; Camilli and Hopkins, 1978, 1979; Overall, 1980). Furthermore, for the cases in which either only one or neither marginal total is fixed, the uncorrected chi-square provides a good approximation to the true probabilities—certainly a better approximation than is provided by Yates' correction.

fixed marginals

6.5 Chi-Square for Larger Contingency Tables

The Pugh example involved two variables (Verdict and Fault), each of which had two levels. We referred to this design as a 2×2 contingency table; it is a special case of the more general $R \times C$ designs, where, again, R and C represent the number of rows and columns. As an example of a larger contingency table, consider the study by Geller, Witmer, and Orebaugh (1976) mentioned in Chapter 5. These authors were studying littering behavior and were interested, among other things, in whether a message about not littering would be effective if placed on the handbills that are often given out in supermarkets advertising the daily specials. To oversimplify a fairly complex study, two of Geller's conditions involved passing out handbills in a supermarket. Under one condition (Control), the handbills contained only a listing of the daily specials. In the other condition (Message), the handbills also included the notation, "Please don't litter. Please dispose of this properly." At the end of the day, Geller and his students searched the store for handbills. They recorded the number that were found in trash cans; the number that were left in shopping carts, on the floor, and various places where they didn't belong (denoted litter); and the number that could not be found and were apparently removed from the premises. The data obtained under the two conditions are shown in Table 6.3 and are taken from a larger table reported by Geller et al. Expected frequencies are shown in parentheses and were obtained exactly as they were in the previous example $[E_{ij} = (R_i)(C_j)/N]$.

[4] As an example of a table with fixed marginals, imagine that we designed a study to present a subject with handwriting samples from 10 physicians and 10 dentists (fixed row marginals of 10 and 10). We told the subject to sort them into piles on the basis of his judgment as to whether each sample came from a physician or a dentist. If we constrain the judge's sorting by saying that each pile must contain 10 samples, our column totals will also be fixed at 10 and 10.

TABLE 6.3 Data from Geller, Witmer, and Orebaugh (1976) (Expected frequencies in parentheses.)

	Location			
Instructions	**Trash Can**	**Litter**	**Removed**	
Control	41 (61.66)	385 (343.98)	477 (497.36)	903
Message	80 (59.34)	290 (331.02)	499 (478.64)	869
	121	675	976	1772

The calculation of χ^2 is carried out just as it was earlier:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$
$$= \frac{(41-61.66)^2}{61.66} + \frac{(385-343.98)^2}{343.98} + \cdots + \frac{(499-478.64)^2}{478.64}$$
$$= 25.79$$

There are 2 *df* for Table 6.3, since $(R-1)(C-1) = (2-1)(3-1) = 2$. The critical value is $\chi^2_{0.05} = 5.99$. Our value of 25.79 is larger than 5.99, so we are led to reject H_0 and to conclude that the location in which the handbills were left depended on the instructions given. In other words, Instructions and Location are not independent. From the data it is evident that when subjects were asked not to litter, a higher percentage of handbills were thrown in the trash can or taken out of the store, and fewer were left lying in shopping carts or on floors and shelves.

Computer Analyses

Chi-square statistics can be produced by computer programs in two different ways. The easier, though less common, way is to use a program in which you enter the cell totals for each row and column combination and ask for χ^2 to be computed. This can be done easily in most programs, and a sample Minitab printout is shown in Exhibit 6.1 (pages 154–155) for the data in Table 6.3. Here you can see Row and Column designations, followed by the cell frequencies. These same data are analyzed using SPSS in Exhibit 6.2 (pages 155–157).

Exhibit 6.2 contains several statistics we have not yet discussed. The three of them at the end (phi, Cramér's V, and the contingency coefficient) will be discussed later in this chapter. Under the Chi-Square heading are several statistics that have a chi-square distribution. The first entry is the standard (Pearson's) chi-square, and it is the statistic we seek. The likelihood ratio is an alternative, and related, test on contingency tables and its statistic is distributed as χ^2. This test will be discussed shortly, as will the measure of linear-by-linear association.

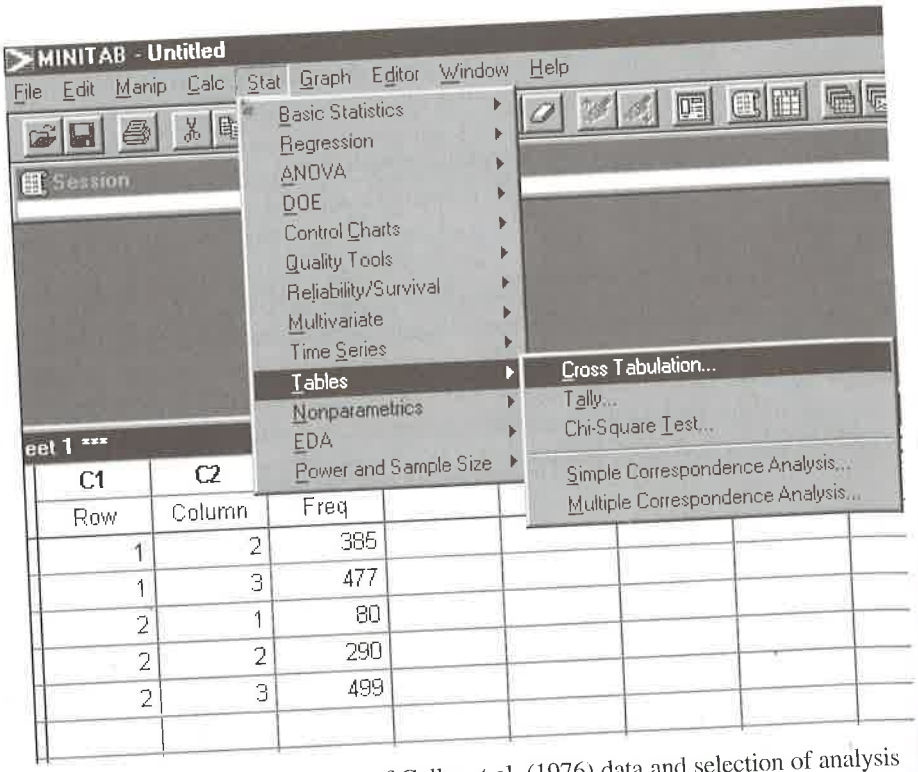

EXHIBIT 6.1a Minitab analysis of Geller et al. (1976) data and selection of analysis

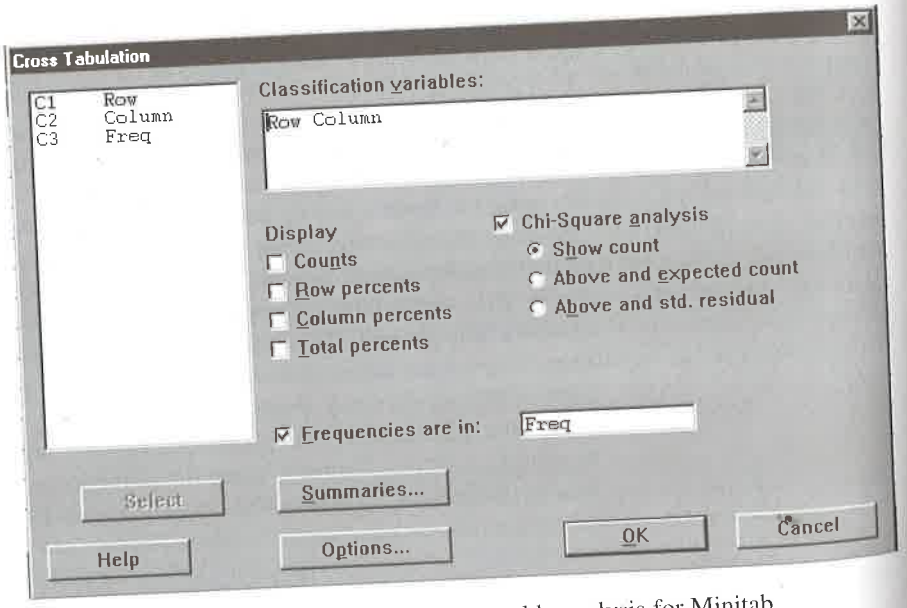

EXHIBIT 6.1b Specification of contingency table analysis for Minitab

Tabulated Statistics

Rows: Row Columns: Column

	1	2	3	All
1	41	385	477	903
2	80	290	499	869
All	121	675	976	1772

Chi-Square = 25.794, DF = 2, P-Value = 0.000

EXHIBIT 6.1c Results of Minitab analysis of Geller et al. data

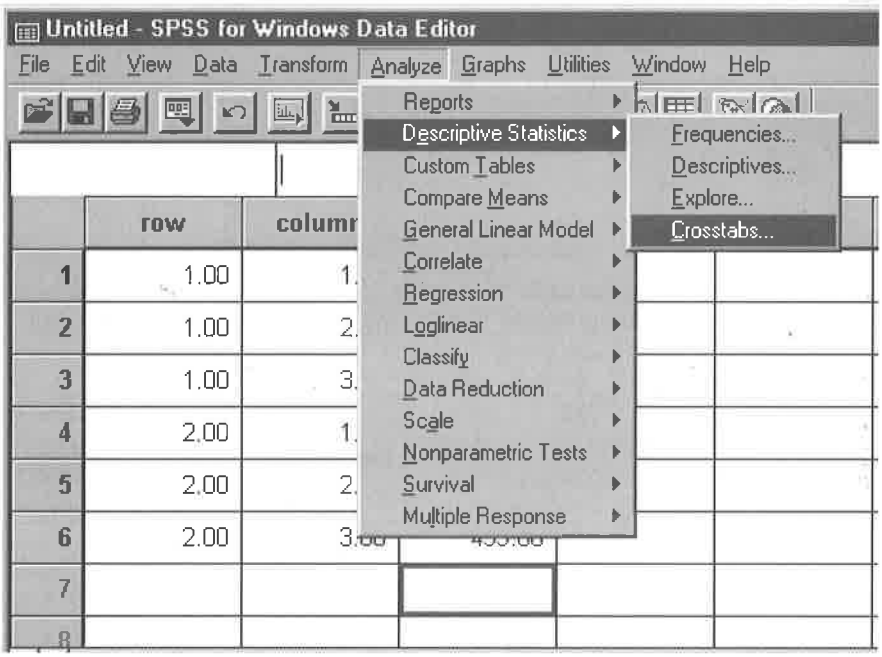

EXHIBIT 6.2a SPSS menus for analysis of Geller et al. data

Exhibit 6.3 on page 157 contains the SPSS printout for the data from Pugh's study of convictions for rape. I have shown this analysis because Pugh had a 2×2 table, which offers some additional considerations.

From Exhibit 6.3 you can see that we obtained the same value of χ^2 that we obtained earlier by hand. The next entry is the value of χ^2 with a continuity correction, as we discussed earlier. Because this result is meaningful only when both sets of marginal totals are fixed, we will ignore that statistic as being inappropriate. Fisher's Exact Test also assumes fixed marginals and we generally ignore it. (Guilt is a random variable, not under the control of the experimenter, and replications of the experiment would produce different column totals.) You did not see an "exact" test or the correction for continuity for Geller's data because they are produced only for 2×2 tables. For our situation the standard Pearson chi-square is most appropriate,

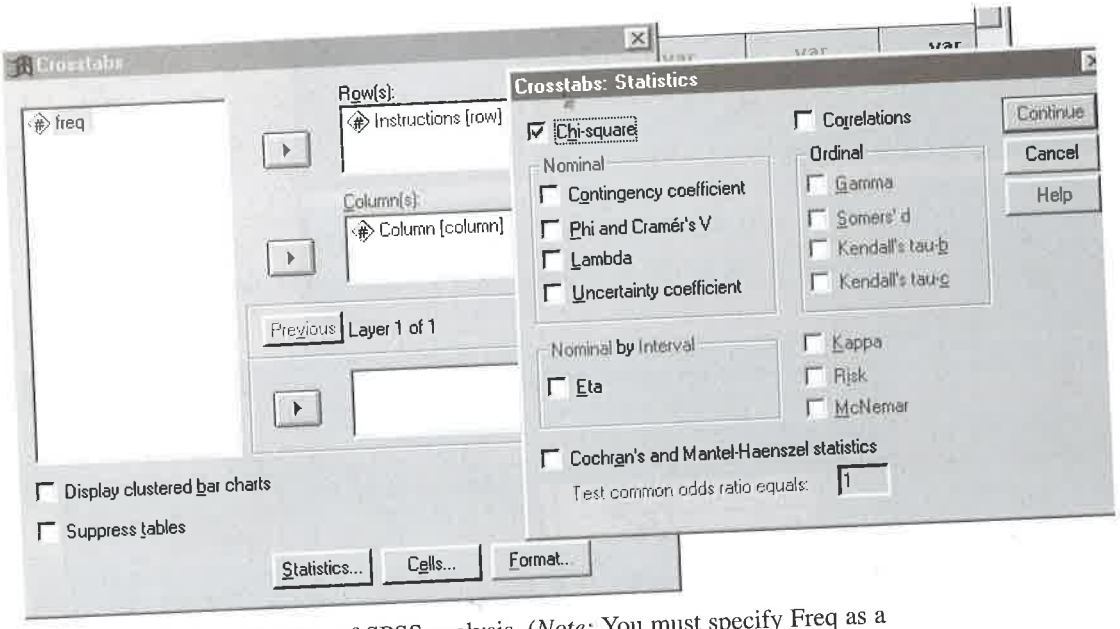

EXHIBIT 6.2b Specification of SPSS analysis. (*Note:* You must specify Freq as a weighting factor in the **Data/Weight Cases** menu before selecting Crosstabs, and you need to select Chi-square, as shown at right, using the Statistics button.)

Instructions * Column Crosstabulation

Count

		Column			Total
		Trashcan	Litter	Removed	
Instructions	Control	41	385	477	903
	Message	80	290	499	869
Total		121	675	976	1772

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	25.794[a]	2	.000
Likelihood Ratio	26.056	2	.000
Linear-by-Linear Association	.001	1	.982
N of Valid Cases	1772		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 59.34.

EXHIBIT 6.2c SPSS output of analysis of Geller et al. data

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.121	.000
	Cramer's V	.121	.000
	Contingency Coefficient	.120	.000
N of Valid Cases		1772	

EXHIBIT 6.2c SPSS output of analysis of Geller et al. data

Crosstabs

Fault * Verdict Crosstabulation

Count

		Verdict		Total
		Guilty	Not Guilty	
Fault	Low	153	24	177
	High	105	76	181
Total		258	100	358

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2 sided)	Exact Sig. (1 sided)
Pearson Chi-Square	35.930[b]	1	.000		
Continuity Correction[a]	34.532	1	.000		
Likelihood Ratio	37.351	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	35.830	1	.000		
N of Valid Cases	358				

a. Computer only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 49.44.

Tests for Homogeneity of the Odds Ratio

Statistics		Chi-Squared	df	Asymp. Sig. (2-sided)
Conditional Independence	Cochran's	35.930	1	.000
	Mantel-Haenszel	34.435	1	.000
Homogeneity	Breslow-Day	.000	0	
	Tarone's	.000	0	

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

EXHIBIT 6.3 SPSS printout for analysis of Pugh's data on convictions for rape

and we will ignore the other tests. Finally, the Mantel–Haenszel test is generally used when we have three or more variables. For example, if Pugh's experiment had been conducted on several distinct populations of subjects (e.g., male and female jurors) and we wanted to do the same analysis but control for population differences, their test would be appropriate.[5]

Small Expected Frequencies

small expected frequency

One of the most important requirements for using the chi-square test concerns the size of the expected frequencies. We have already met this requirement briefly in discussing corrections for continuity. Before defining more precisely what we mean by *small,* we should examine why a **small expected frequency** causes so much trouble. There are two ways of explaining the difficulty, and they are so closely related that it is difficult to speak about one without invoking the other.

First, consider the basic fact that we are using a mathematical distribution to approximate the distribution of the statistic resulting from a chi-square test. As you should recall, in deriving the test we first showed that when we sampled observations from a normal population,

$$\chi^2 = \sum z^2$$

We then used this relationship to show that $\sum z^2$ could be converted into our familiar chi-square statistic, invoking the binomial (or multinomial) distribution in the process. We were allowed to use the binomial (or multinomial) because these distributions approach the normal when Np is large, but we have invoked the assumption of normality in the process. Neither the binomial nor the multinomial, however, produces anywhere near normal distributions for small values of Np (e.g., see Figure 5.5). This means that we will have violated the assumption of **normality** for small expected frequencies (i.e., for small values of Np).

normality

Suppose we look at the problem from a different angle. If only a few different values of χ^2_{obt} are possible, then the χ^2 distribution cannot provide a reasonable approximation to the distribution of our statistic. We cannot fit a discrete distribution having relatively few values with a continuous one. Those cases that result in only a few possible values of χ^2_{obt}, however, are the ones with small expected frequencies. (This is directly analogous to the fact that if you flip a coin three times, there are only four possible values for the number of heads, and the resulting sampling distribution certainly cannot be satisfactorily approximated by the normal distribution.) If you select the marginal totals so as to have at least one small expected frequency, and then construct all possible data tables for that set of marginals, it will be immediately obvious what the problem is.

We have seen that difficulties arise when we have small expected frequencies, but the question of "How small is small?" remains. Those conventions that do exist are

[5]Although the Mantel–Haenszel test would be appropriate if we had a third variable, such as sex of juror, it assumes that whatever the relationship between Fault and Verdict, it is the same relationship for both levels of sex.

conflicting and have only minimal claims to preference over one another. Probably the most common is to require that all expected frequencies should be at least 5. This is a conservative position and I don't feel overly guilty when I violate it. Bradley et al. (1979) ran a computer-based sampling study. They used tables ranging in size from 2×2 to 4×4 and found that for those applications likely to arise in practice, the actual percentage of Type I errors rarely exceeds 0.06, even for *total* samples sizes as small as 20, unless the row or column marginal totals are drastically skewed. Camilli and Hopkins (1979) demonstrated that even with quite small expected frequencies, the test produces few Type I errors in the 2×2 case as long as the total sample size is greater than or equal to 8; but they, and Overall (1980), point to the extremely low power to reject a false H_0 that such tests possess. With small sample sizes, power is more likely to be a problem than inflated Type I error rates.

6.6 Chi-Square for Ordinal Data

Chi-square is an important statistic for the analysis of categorical data, but it can sometimes fall short of what we need. If you apply chi-square to a contingency table, and then rearrange one or more rows or columns and calculate chi-square again, you will arrive at exactly the same answer. That is as it should be, because chi-square does not take the ordering of the rows or columns into account.

But what do you do if the order of the rows and/or columns does make a difference? How can you take that ordinal information and make it part of your analysis? An interesting example of just such a situation was provided in a query that I received from Jennifer Mahon at the University of Leicester, in England.

Ms Mahon collected data on the treatment for eating disorders. She was interested in how likely participants were to remain in treatment or drop out, and she wanted to examine this with respect to the number of traumatic events they had experienced in childhood. Her general hypothesis was that participants who had experienced more traumatic events during childhood would be more likely to drop out of treatment. Notice that her hypothesis treats the number of traumatic events as an ordered variable, which is something that chi-square ignores. There is a solution to this problem, but it is more appropriately covered after we have talked about correlations. I will come back to this problem in Chapter 10 and show you one approach. (Many of you could skip now to Chapter 10 and be able to follow the discussion.) I mention it here because it comes up most often when discussing χ^2.

6.7 Summary of the Assumptions of Chi-Square

assumptions of χ^2

Because of the widespread misuse of chi-square still prevalent in the literature, it is important to pull together in one place the underlying **assumptions of χ^2.** For a thorough discussion of the misuse of χ^2, see the paper by Lewis and Burke (1949) and the subsequent rejoinders to that paper. These articles are not yet out of date, although it has been 50 years since they were written. A more recent discussion of many of the issues raised by Lewis and Burke (1949) can be found in Delucchi (1983).

The Assumption of Independence

At the beginning of this chapter, we assumed that observations were independent of one another. The word *independence* has been used in two different ways in this chapter, and it is important to keep these two uses separate. A basic assumption of χ^2 deals with the independence of *observations* and is the assumption, for example, that one subject's choice among brands of coffee has no effect on another subject's choice. This is what we are referring to when we speak of an assumption of independence. We also spoke of the independence of *variables* when we discussed contingency tables. In this case, independence is what is being tested, whereas in the former use of the word it is an assumption.

It is not uncommon to find cases in which the assumption of independence of observations is violated, usually by having the same subject respond more than once. A typical illustration of the violation of the independence assumption occurred when a former student categorized the level of activity of each of five animals on each of four days. When he was finished, he had a table similar to this:

Activity			
High	**Medium**	**Low**	**Total**
10	7	3	20

This table looks legitimate until you realize that there were only five animals, and thus each animal was contributing four tally marks toward the cell entries. If an animal exhibited high activity on Day 1, it is likely to have exhibited high activity on other days. The observations are not independent, and we can make a better-than-chance prediction of one score knowing another score. This kind of error is easy to make, but it is an error nevertheless. The best guard against it is to make certain that the total of all observations (N) equals precisely the number of participants in the experiment.

Normality

We discussed the assumption of normality (and therefore continuity) at length when we spoke of small expected frequencies. Nothing more needs to be said here.

Inclusion of Nonoccurrences

Although the requirement that nonoccurrences be included has not yet been mentioned specifically, it is inherent in the derivation and is probably best explained by an example. Suppose that out of 20 students from rural areas, 17 were in favor of having daylight savings time (DST) all year. Out of 20 students from urban areas, only 11 were in favor of DST on a permanent basis. We want to determine whether significantly more rural students than urban students are in favor of DST. One *erroneous* method of testing this would be to set up the following data table on the number of students favoring DST:

	Rural	Urban	Total
Observed	17	11	28
Expected	14	14	28

We could then compute $\chi^2 = 1.29$ and fail to reject H_0. This data table, however, does not take into account the *negative* responses, which Lewis and Burke (1949) call **nonoccurrences.** In other words, it does not include the numbers of rural and urban students *opposed* to DST. Look back at our original derivation of chi-square. Note that we began with the formula for the binomial, which included $N - X$ and Nq as well as X and Np. These values must also be taken into account in our analysis. Therefore, the data should really be cast in the form of a contingency table:

nonoccurrences

	Rural	Urban	
Yes	17	11	28
No	3	9	12
	20	20	40

Now $\chi^2 = 4.29$, which is significant at $\alpha = 0.05$, resulting in an entirely different interpretation of the results.

Perhaps a more dramatic way to see why we need to include nonoccurrences can be shown by assuming that 17 out of *2000* rural students and 11 out of 20 urban students preferred DST. Consider how much different the interpretation of the two tables would be. Certainly, our analysis must reflect the difference between the two data sets, which would not be the case if we failed to include nonoccurrences.

Failure to take the nonoccurrences into account not only invalidates the test, but also reduces the value of χ^2, leaving you less likely to reject H_0. Again, you must be sure that the total (N) equals the number of participants in the study.

6.8 One- and Two-Tailed Tests

People are often confused as to whether chi-square is a one- or a two-tailed test. This confusion results from the fact that there are different ways of defining what we mean by a one- or a two-tailed test. If we think of the sampling distribution of χ^2, we can argue that χ^2 is a one-tailed test because we reject H_0 only when our value of χ^2 lies in the extreme right tail of the distribution. On the other hand, if we think of the underlying data on which our obtained χ^2 is based, we could argue that we have a two-tailed test. If, for example, we were using chi-square to test the fairness of a coin, we would reject H_0 if it produced too many heads *or* if it produced too many tails, since either event would lead to a large value of χ^2.

The preceding discussion is not intended to start an argument over semantics (it does not really matter whether you think of the test as one-tailed or two); rather, it is intended to point out one of the weaknesses of the chi-square test, so that you can take this into account. The weakness is that the test, as normally applied, is nondirectional. To take a simple example, consider the situation in which you wish to show

that increasing amounts of quinine added to an animal's food make it less appealing. You take 90 rats and offer them a choice of three bowls of food that differ in the amount of quinine that has been added. You then count the number of animals selecting each bowl of food. Suppose the data are

Amount of Quinine		
Small	**Medium**	**Large**
39	30	21

The computed value of χ^2 is 5.4, which, on 2 *df*, is not significant at $p < 0.05$.

The important fact about the data is that any of the six possible configurations of the same frequencies (such as 21, 30, 39) would produce the same value of χ^2, and you receive no credit for the fact that the configuration you obtained is precisely the one that you predicted. Thus, you have made a *multi-tailed* test when in fact you have a specific prediction of the direction in which the totals will be ordered. I referred to this problem a few pages back when discussing a problem raised by Jennifer Mahon. A solution to this problem will be given in Chapter 10 (Section 10.4), where I discuss creating a correlational measure of the relationship between the two variables.

6.9 Likelihood Ratio Tests

likelihood ratios

An alternative approach to analyzing categorical data is based on **likelihood ratios.** For large sample sizes the two tests are equivalent, though for small sample sizes the standard Pearson chi-square is thought to approximate the exact chi-square distribution better than the likelihood ratio chi-square (Agresti, 1990). Moreover, likelihood ratio tests are heavily used in log-linear models for analyzing contingency tables, because of their additivity properties. Log-linear models will be discussed in Chapter 17. Such models are particularly important when we want to analyze multidimensional contingency tables. Such models are being used more and more, and you should be exposed at least minimally to such methods.

Without going into detail, we can describe quite simply the general idea of a likelihood ratio. Suppose we collect data and calculate the likelihood of the data occurring given that the null hypothesis is true. We also calculate the likelihood that the data would occur under some alternative hypothesis (the hypothesis for which the data are most likely). If the data are much more likely for some alternative hypothesis than for H_0, we would be inclined to reject H_0. However, if the data are almost as likely under H_0 as they are for some other alternative, we would be inclined to retain H_0. Thus, the likelihood ratio (the ratio of these two likelihoods) forms a basis for evaluating the null hypothesis.

Using likelihood ratios, it is possible to devise tests, frequently referred to as "maximum likelihood χ^2 tests," for analyzing both one-dimensional arrays and contingency tables. For the development of these tests, see Mood (1950) or Mood and Graybill (1963).

For the one-dimensional goodness-of-fit case,

$$\chi^2_{(C-1)} = 2\sum O_i \ln\left(\frac{O_i}{E_i}\right)$$

where O_i and E_i are the observed and expected frequencies for each cell and ln denotes the natural logarithm (logarithm to the base e). This value of χ^2 can be evaluated using the standard table of χ^2 on $C - 1$ degrees of freedom.

For analyzing contingency tables, we can use essentially the same formula,

$$\chi^2_{(R-1)(C-1)} = 2\sum O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

where O_{ij} and E_{ij} are the observed and expected frequencies in each cell. The expected frequencies are obtained just as they were for the standard Pearson chi-square test. This statistic is evaluated with respect to the χ^2 distribution on $(R - 1)(C - 1)$ degrees of freedom.

As an illustration of the use of the likelihood ratio test for contingency tables, consider the data found in the Pugh (1983) study. The cell and marginal frequencies follow:

	Verdict		
Fault	Guilty	Not Guilty	
Low	153	24	177
High	105	76	181
	258	100	358

$$\begin{aligned}\chi^2 &= 2\sum O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right)\\ &= 2\left[153 \ln\left(\frac{153}{127.559}\right) + 24 \ln\left(\frac{24}{49.441}\right) + 105 \ln\left(\frac{105}{130.441}\right) + 76 \ln\left(\frac{76}{50.559}\right)\right]\\ &= 2[153(0.1819) + 24(-0.7227) + 105(-0.2170) + 76(0.4076)]\\ &= 2[18.6785] = 37.36\end{aligned}$$

This answer agrees with the likelihood ratio statistic found in Exhibit 6.3. It is a χ^2 on 1 *df*, and since it exceeds $\chi^2_{0.05}(1) = 3.84$, it will lead to rejection of H_0. The decision of the juror depends in part on how the victim is portrayed.

6.10 Measures of Association

A chi-square test is designed primarily to test a null hypothesis. When the data are in the form of a contingency table, the test tells us whether or not the two variables that serve as the basis of classification for that table are independent. But assuming that the test is significant, it still does not tell us much about the degree of relationship between the two variables—only that they are not independent.

For example, the contingency tables in Tables 6.4 and 6.5 will both lead to a significant chi-square test, but they clearly represent different degrees of relationship between the variables. (The examples are fictitious, but the numbers are reasonable.) In Table 6.4, males are somewhat more likely than females to identify as smokers. In Table 6.5, regardless of recent changes in the status of women, women are decidedly more likely to be the primary shopper for food for the family. Although the χ^2 statistics are approximately the same in these two situations, the magnitude of χ^2 is not a useful measure of association. (For example, doubling every number in one of these examples will double the magnitude of χ^2, but it will obviously not change the relative magnitude of the differences among cells.) However, a number of statistics have been developed to address the question of the degree of relationship between variables. These coefficients are generally referred to as **measures of association.** Although none is entirely satisfactory as a measure, you should be familiar with the most common of them. The three coefficients we will consider are based on the χ^2 statistic and are very easy to calculate.

measures of association

TABLE 6.4 The relationship between smoking and gender

	Smoking Behavior		
	Nonsmoker	Smoker	
Male	350	150	500
Female	400	100	500
	750	250	1000

TABLE 6.5 The relationship between the primary shopper and gender

	Primary Food Shopper		
	Yes	No	
Male	15	4	19
Female	4	15	19
	19	19	38

Contingency Coefficient (C)

When data appear in the form of a contingency table of any dimension, one of the commonly employed coefficients is the **contingency coefficient (C).** For all tables, C is defined as

contingency coefficient (C)

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Pearson originally devised this statistic as an approximation of the correlation between two artificially dichotomized variables (see Chapter 10). However, several difficulties

arise with this coefficient. The first is rather obvious; namely, because N is always greater than 0, C can never equal 1. More important, however, the maximum value of C is dependent on the dimensions of the table from which it is computed. Thus for a 2×2 table, the maximum possible value for C is 0.707. Because of these problems, I do not recommend this coefficient. I cover it here only for completeness.

Phi (ϕ)

phi (ϕ)

In the case of 2×2 tables, a correlation coefficient that we will consider in Chapter 10 serves as a good measure of association. This coefficient is called **phi** (ϕ), and it represents the correlation between two variables, each of which is a dichotomy (i.e., takes on one of two distinct value.). If we coded Gender as 1 or 2, for male and female, and coded Smoking as 1 for nonsmoker and 2 for smoker, and then determined the correlation between the two variables (see Chapters 9 and 10), the result would be phi. An easier way to calculate ϕ for these data is by the relation

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

For the data in Table 6.4, $\chi^2 = 13.333$ and $\phi = \sqrt{\frac{13.333}{1000}} = 0.12$.

Cramér's Phi (ϕ_c)

The difficulty with phi is that it applies only to 2×2 tables, and therefore is not of any use with larger contingency tables. Cramér (1946) proposed a way around this problem by defining

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

where N is the sample size and k is defined as the smaller of R and C.

Cramér's ϕ_c
Cramér's V

Cramérs ϕ_c, often known as **Cramér's V** (I suspect because mainframe computer line printers did not usually print Greek characters), can be seen as a simple extension of ϕ. Note that when $k = 2$, it *is* ϕ. Unlike the contingency coefficient, the maximum value is 1 regardless of the dimensionality of the table.

If I were going to retain only one measure of association, it would be Cramér's ϕ_c. It is not constrained by the size of the table; it reduces to ϕ for 2×2 tables; and, especially with 2×2 tables, it has a certain reasonable interpretation.

For Tables 6.4 and 6.5 given at the beginning of this section, $\chi^2 = 13.333$ and 12.737 respectively. Then $\phi_c = 0.12$ and 0.58. If you think of these as coefficients with a possible range of 0 to +1.00, you can see that they clearly reflect the differences apparent in the two tables.

Odds Ratios

odds ratio

A useful statistic, especially for 2×2 tables, that makes clear the degree to which one variable influences another is the **odds ratio.** Odds ratios have the distinct advantage of being unaffected by sample size and by unequal row or column totals.

A good example of the use of an odds ratio is cited in a paper by Rosenthal (1990), who was actually using the data for a different purpose. An important study of the beneficial effects of small daily doses of aspirin on reducing heart attacks in men was reported in 1988. Over 22,000 physicians were administered daily doses of aspirin or a placebo, and the incidence of heart attacks was recorded. The data are presented in Table 6.6.

TABLE 6.6 The effect of aspirin on the incidence of heart attacks

	Outcome		
	Heart Attack	No Heart Attack	
Aspirin	104	10,933	11,037
Placebo	189	10,845	11,034
	293	21,778	22,071

For these data, 0.94% of people in the aspirin group and 1.71% of those in the control group suffered a heart attack during the course of the study, a difference of 0.77 percentage point. (Unless you are a middle-aged male worrying about your health, the numbers look rather small.) Instead of simply looking at percentages, we can look at the odds in favor of heart attacks over no heart attacks. The odds of having a heart attack *given that you were in the aspirin group* are equal to the number of people in that group who had a heart attack divided by the number in that group who did not have a heart attack, which equals $104/10{,}933 = 0.0095125$. If you were in the control group, the odds of your having a heart attack are $189/10{,}845 = 0.0174274$. Both of these odds are quite low. However, if we form a ratio of these two odds, the ratio is $0.0174274/0.0095125 = 1.83$. This means that a person in the control group is 1.83 times more likely to have a heart attack than is a person in the aspirin group. Put in the reverse, you are about half as likely to have a heart attack if you take an aspirin a day than if you don't take an aspirin. That effect is impressive concerning something as serious as a heart attack. Odds ratios are often a clear and effective way of presenting data in 2×2 tables.

As a result of this study, many physicians recommend that their male patients over the age of 40 take an aspirin every morning. Since the study did not involve women, and since women's rate of heart attacks is somewhat different from that of men, the data do not speak clearly to what women should do.

It is possible to extend the discussion of odds ratios to larger contingency tables. In some cases the interpretation is awkward, but in many cases it is clear. For example, if we return to the study by Geller et al. (1976) on littering in supermarkets, it is fairly clear that the real question is whether people litter the store with advertising brochures. This means that we could collapse over the "Trash can" and "Removed" categories. If we do this we find that the odds of littering for people in the control (no

message) group are 385/(41 + 477) = 0.7432. The odds of littering in the Message group are 290/(80 + 499) = 0.5009. Then the odds ratio of littering given no message to littering given the message is 0.7432/0.5009 = 1.48. In other words, someone is about one-and-one-half times as likely to litter if no message is included on the flier. To look at this the other way around, the odds ratio of littering given the message versus not being given the message is 0.5009/0.7432 = 0.67. Thus a shopper is only about two-thirds as likely to litter if a message is included on the flier.

Although chi-square is often the test of choice when testing for the significance of departures from independence, and although measures of association speak to the relative magnitude of effects, odds ratios are often a much better way of reporting what the data mean than are the more traditional measures of association discussed in the previous sections.

Kappa (κ)—A Measure of Agreement

kappa (κ)

An important statistic that is not based on chi-square but that does use contingency tables is **kappa** (κ), commonly known as Cohen's kappa (Cohen, 1960). This statistic measures interjudge agreement and is often used when we wish to examine the reliability of ratings.

Suppose we asked a judge with considerable clinical experience to interview 30 adolescents and classify them as exhibiting (1) no behavior problems, (2) internalizing behavior problems (e.g., withdrawn), and (3) externalizing behavior problems (e.g., acting out). Anyone reviewing our work would be concerned with the reliability of our measure—how do we know that this judge was doing any better than flipping a coin? As a check we ask a second judge to go through the same process and rate the same adolescents. We then set up a contingency table showing the agreements and disagreements between the two judges. Suppose the data are those shown in Table 6.7.

Ignore the values in parentheses for the moment. In this table, Judge I classified 16 adolescents as exhibiting no problems, as shown by the total in the first column. Of those 16, Judge II agreed that 15 had no problems, but also classed 1 of them as exhibiting internalizing problems and 0 as exhibiting externalizing problems. The

TABLE 6.7

	Judge 1			
Judge II	No Problem	Internalizing	Externalizing	Total
No Problem	15 (10.67)	2	3	20
Internalizing	1	3 (1.20)	2	6
Externalizing	0	1	3 (1.07)	4
Total	16	6	8	30

entries on the diagonal (15, 3, 3) represent agreement between the two judges, whereas the off-diagonal entries represent disagreement.

percentage of agreement

The simplest approach to these data is to calculate the **percentage of agreement.** For this statistic all we need to say is that out of 30 total cases, there were 21 cases $(15 + 3 + 3)$ where the judges agreed. Then the percentage of agreement is $21/30 = 0.70 = 70\%$. This measure has problems, however. The majority of the adolescents in our sample exhibit no behavior problems, and both judges are (correctly) biased toward a classification of No Problem and away from the other classifications. The probability of No Problem for Judge I would be estimated as $16/30 = 0.53$. The probability of No Problem for Judge II would be estimated as $20/30 = 0.67$. If the two judges operated independently, the probability that they would both classify a case as No Problem is $0.53 \times 0.67 = 0.36$, which for 30 judgments would mean $0.36 \times 30 = 10.67$ agreements on No Problem alone, purely by chance.

Cohen (1960) proposed a chance-corrected measure of agreement known as kappa. To calculate kappa we first need to calculate the expected frequencies for each of the diagonal cells assuming that judgments are independent. We calculate these the same way we calculate the expected frequencies for the standard chi-square test. For example, the expected frequency of both judges assigning a classification of No Problem is $(20 \times 16)/30 = 10.67$. For Internalizing it is $(6 \times 6)/30 = 1.2$, and for Externalizing it is $(4 \times 8)/30 = 1.07$. These values are shown in parentheses in the table.

We will now define kappa as

$$\kappa = \frac{\sum f_O - \sum f_E}{N - \sum f_E}$$

where f_O represents the observed frequencies on the diagonal and f_E represents the expected frequencies on the diagonal. Thus

$$\sum f_O = 15 + 3 + 3 = 21 \quad \text{and} \quad \sum f_E = 10.67 + 1.20 + 1.07 = 12.94$$

Then

$$\kappa = \frac{21 - 12.94}{30 - 12.94} = \frac{8.06}{17.06} = 0.47$$

Notice that this coefficient is considerably lower than the 70% agreement figure that we calculated above. Instead of 70% agreement, we have 47% agreement after correcting for chance.

If you examine the formula for kappa, you can see the correction that is being applied. In the numerator we subtract, from the number of agreements, the number of agreements that we would expect merely by chance. In the denominator we reduce the total number of judgments by that same amount. We then form a ratio of the two chance-corrected values.

Cohen and others have developed statistical tests for the significance of kappa. However, its significance is often not the issue. If kappa is low enough for us to even question its significance, the lack of agreement among our judges is a serious problem.