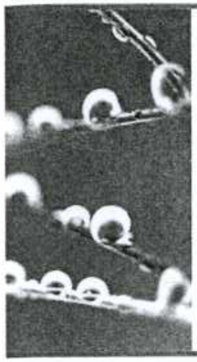
CHAPTER 6



Categorical Data and Chi-Square

Objectives

To present the chi-square test as a procedure for testing hypotheses when the data are categorical and to examine other measures that clarify the meaning of our results.

Contents

IN SAINT-EXUPÉRY'S, *The Little Prince,* the narrator, remarking that he believes the prince came from an asteroid known as B-612, explains his attention to such a trivial detail as the precise number of the asteroid with the following comment:

> Grown-ups love figures. When you tell them you have made a new friend, they never ask you any questions about essential matters. They never say to you, "What does his voice sound like? What games does he love best? Does he collect butterflies?" Instead they demand: "How old is he? How many brothers has he? How much does he weigh? How much does his father make?" Only from these figures do they think they have learned anything about him.[1]

In some ways, the first chapters of this book have concentrated on dealing with the kinds of numbers Saint-Exupéry's grown-ups like so much. This chapter will be devoted to the analysis of largely nonnumerical data.

In Chapter 1, I drew a distinction between measurement data (sometimes called quantitative data) and categorical data (sometimes called frequency data). When we deal with measurement data, each observation represents a score along some continuum, and the most common statistics are the mean and the standard deviation. When we deal with categorical data, on the other hand, the data consist of the frequencies of observations that fall into each of two or more categories ("Does your friend have a gravelly voice or a high-pitched voice?" or "Is he a collector of butterflies, coins, or baseball cards?").

In Chapter 5, we examined the use of the binomial distribution to test simple hypotheses. In those cases, we were limited to situations in which an individual event had one of only two possible outcomes, and we merely asked whether, over repeated trials, one outcome occurred (statistically) significantly more often than the other.

In this chapter, we will expand the kinds of situations that we can evaluate. We will deal with the case in which a single event can have two *or more* possible outcomes, and then with the case in which we have two variables and we want to test null hypotheses concerning their independence. For both of these situations, the appropriate statistical test will be the chi-square (χ^2) test.

chi-square (χ^2)

The term **chi-square** (χ^2) has two distinct meanings in statistics, which leads to some confusion. In one meaning, it is used to refer to a particular mathematical distribution that exists in and of itself without any necessary referent in the outside world. In the second meaning, it is used to refer to a statistical test that has a resulting test statistic distributed in approximately the same way as the χ^2 distribution. When you hear someone refer to chi-square, they usually have this second meaning in mind. (The test itself was developed by Karl Pearson [1900] and is often referred to as **Pearson's chi-square** to distinguish it from other tests that also produce a χ^2 statistic—for example, Friedman's test, discussed in Chapter 18, and the likelihood ratio tests discussed at the end of this chapter and in Chapter 17.) You need to be familiar with both meanings of the term, however, if you are to use the test correctly and intelligently and if you are to understand many of the other statistical procedures that follow.

Pearson's chi-square

6.1 The Chi-Square Distribution

chi-square (χ^2) distribution

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

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

[1] Antoine de Saint-Exupéry, *The Little Prince,* trans. Katherine Woods (New York: Harcourt Brace, 1943), pp. 15–16.

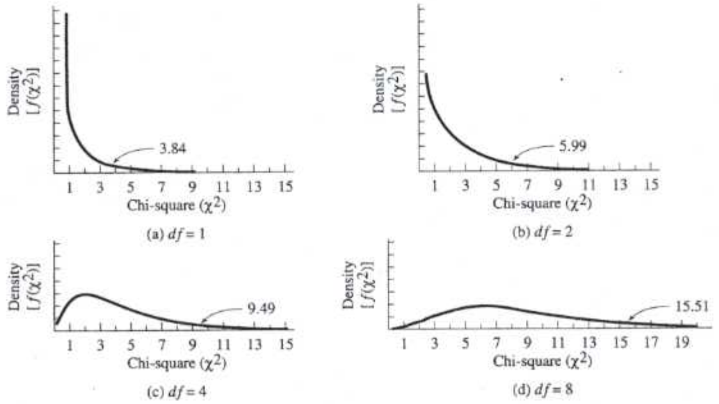
gamma function

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.

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 else 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, we can see that the distribution changes 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

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



(a) $df = 1$

(b) $df = 2$

(c) $df = 4$

(d) $df = 8$

Figure 6.1 Chi-square distributions for $df = 1, 2, 4$, and 8 (arrows indicate critical values at alpha = .05)

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

chi-square test

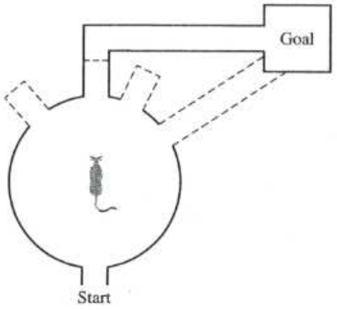
We now turn to what is commonly referred to as the **chi-square test**, which is based on the χ^2 distribution. We will first examine the test as it is applied to one-dimensional tables and then as applied to two-dimensional tables (contingency tables).

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?" Though that doesn't seem like such a radical question now, I assure you that it was a very radical question in the 1940s.) 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 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." As 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 the result of 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 they had learned something.

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?



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

Figure 6.2 Schematic diagram of Tolman's maze. The alleys are labeled A through D from left to right.

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 in Figure 6.2. 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 8 times, we would probably have strong reservations about this assumption of independence.) We have four alleys, so if the rats are responding at random, rather than on the basis of what they have learned about the maze, we would expect that one-quarter 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.

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

This 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 1,000 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 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.

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

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 if the null hypothesis of a chance distribution were true. 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 that we saw 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 as many as 30 *df*, which would require 30 separate tables, each the size of Appendix *z*. Such a procedure would be particularly wasteful because most users would need only a small fraction of each of these tables. If we want to reject H_0 at the .05 level, all that we really care about is whether 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% cutoff, for example. Thus, tables of χ^2 such as the one given in Appendix χ^2, part of which is reproduced 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	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.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. Because our obtained value $(\chi^2_{\text{obt}}) = 9.25$ is greater than $\chi^2_{.05} = 7.82$, 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.

6.3 Two Classification Variables: Contingency Table Analysis

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.

contingency table

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 about (1) whether the defendant was found innocent or 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 such a contingency table.

[2] Notice that here the subscript for χ^2 (i.e., obt and .05) do not refer to the degrees of freedom, but designate either the obtained value of $\chi^2[\chi^2_{\text{obt}}]$ or the value of χ^2 that cuts off the largest 5% of the distribution $[\chi^2_{.05}]$. When we want to designate both the degrees of freedom and the level of alpha we write something like $\chi^2_{.05}(1) = 3.84$.

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

For the moment, ignore the numbers in parentheses. 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 seen as low in fault, approximately 86% (153/177) of the time the defendant was found guilty. On the other hand, when the victim was seen 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 totals

The expected frequencies in a contingency table represent those frequencies that we would expect if the two variables forming the table (here, guilt and victim blame) were independent. 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

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

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

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

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

These values are shown in parentheses in Table 6.2.

[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.

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.

$$\begin{aligned}\chi^2 &= \sum \frac{(O-E)^2}{E} \\ &= \frac{(153-127.559)^2}{127.559} + \frac{(24-49.441)^2}{49.441} + \frac{(105-130.441)^2}{130.441} + \frac{(76-50.559)^2}{50.559} \\ &= 35.93\end{aligned}$$

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*, 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's correction for continuity

Many books advocate that for simple 2 × 2 tables such as Table 6.2, we should employ what is called **Yates's correction for continuity**, especially when the expected frequencies are small. (The correction merely involves reducing the absolute value of each numerator by 0.5 units before squaring.) There is an extensive literature debating the pros and cons of Yates's correction, with firmly held views on both sides. However, the common availability of Fisher's Exact Test, to be discussed next, makes Yates's correction superfluous.

Fisher's Exact Test

Fisher introduced what is called Fisher's Exact Test in 1934 at a meeting of the Royal Statistical Society. (Good [2001] has pointed out that one of the speakers who followed Fisher referred to Fisher's presentation as "the braying of the Golden Ass." Statistical debates at that time were far from boring, and no doubt Fisher had something equally kind to say about his critic.)

Without going into details, Fisher's proposal was to take all possible 2 × 2 tables that could be formed from the fixed set of marginal totals. He then determined the proportion of those tables whose results are as extreme, or more so, than the table we obtained from our data. If this proportion is less than α, we reject the null hypothesis that the two variables are independent and conclude that there is a statistically significant relationship between the two variables that make up our contingency table. (This is classed as a *conditional test* because it is conditioned on the marginal totals actually obtained, instead of all possible marginal totals given the total sample size.) I am assuming that you will do the calculations using statistical software rather than by hand.

fixed marginals

Fisher's Exact Test has been controversial since he proposed it. One problem concerns the fact that it is a conditional test (conditional on the **fixed marginals**). Some have argued that if you repeated the experiment exactly, you would likely find different marginal totals and have asked why those additional tables should not be included in the calculation. Making the test unconditional on the marginals complicates the calculations considerably. This may sound like an easy debate to resolve, but if you read the extensive literature surrounding fixed and random marginals, you will find that it is a difficult debate to follow and you will probably come away thoroughly confused. (An excellent discussion of some of the issues can be found in Agresti (2002), pages 95–96.)

Fisher's Exact Test also leads to controversy because of the issue of one-tailed versus two-tailed tests and what outcomes would constitute a "more extreme" result in the opposite tail. Instead of going into how to determine what is a more extreme outcome, I will avoid that complication by simply telling you to decide in advance whether you want a one- or a two-tailed test, and then to report the values given by standard statistical software. (I haven't given you any calculational formula for Fisher's Exact Test because I cannot imagine that you would ever do the calculations by hand.) Virtually all common statistical software prints out Fisher's Exact Test results along with Pearson's chi-square and related test statistics.

Fisher's Exact Test versus Pearson's Chi-Square

We now have at least two statistical tests for 2 × 2 contingency tables, and will soon have a third—which one should we use? Probably the most common solution is to go with Pearson's chi-square; perhaps because "that is what we have always done." In previous editions of this book I recommended against Fisher's Exact Test, primarily because of the conditional nature of it. However, in recent years there has been an important growth of interest in permutation and randomization tests, of which Fisher's Exact Test is an example. (This approach is discussed extensively in Chapter 18.) I am extremely impressed with the logic and simplicity of such tests and have come to side with Fisher's Exact Test. In most cases, the conclusion you will draw will be the same for the two approaches, though this is not always the case. When we come to tables larger than 2 × 2, Fisher's approach does not apply, without modification, and there we almost always use the Pearson chi-square. (But see Howell & Gordon, 1976.)

6.4 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.

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

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 [$E_{ij} = (R_i)(C_j)/N$].

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 two *df* for Table 6.3 because $(R - 1)(C - 1) = (2 - 1)(3 - 1) = 2$. The critical value of $\chi^2_{.05} = 5.99$. 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 to some extent 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.

As we have seen, the chi-square test can be applied to two-dimensional tables of any size (and, in some situations, tables of more dimensions). The calculations are always the same. The problem with larger tables, however, is one of interpretation. If a 2 × 2 chi-square is statistically significant, it is usually pretty obvious what the results mean. We just have to look at the contingency table. But with larger tables, it is not always clear. In the Geller et al. (1976) example, was chi-square significant because of a disparate distribution in the "litter" column, or the "trash" column, or in all three columns? There are statistical techniques to help tease this apart, but they are not common. Often larger contingency tables are collapsed back to 2 × 2 tables for ease of interpretation. We will see a similar kind of issue raised when we consider odds ratios shortly.

Computer Analyses

Chi-square statistics can be produced by computer programs in two different ways. Suppose that we had a data file containing Pugh's data on convictions for rape. One column (Fault) would contain a 1 if that defendant's lawyer had tried to assign high blame to the victim or a 0 if he assigned low blame. A second column (Guilt) would contain a 1 if the defendant was found guilty, and a 0 if not. (Alternatively, we could code the Fault variable as "Little" or "Much" depending on whether the victim was assigned little or much fault by the attorney. We could also code Guilt as "Guilty" and "Not Guilty." There would be 358 lines of data, one for each case. We could then ask SPSS (or almost any other program) to cross tabulate Fault against Guilt. This analysis is presented in Exhibit 6.1.

Exhibit 6.1 contains several statistics we have not yet discussed. In Exhibit 6.1b, the likelihood ratio test is one that we shall take up shortly and is simply another approach to calculating chi-square. The three statistics in Exhibit 6.1c (phi, Cramér's V, and the contingency coefficient) will also be discussed later in this chapter, as will the odds ratio shown in Exhibit 6.1d. Each of these four statistics is an attempt to assess the size of the effect.

If you didn't already have a data file for Pugh's data, you would probably not be eager to create a file of 358 lines just to calculate a simple chi-square. Fortunately, there is an alternative approach that is much quicker. Basically, we create one line of data for each

Fault * Guilt Crosstabulation

Count

		Guilt		
		Guilty	Not Guilty	Total
Fault	Little	153	24	177
	Much	105	76	181
Total		258	100	358

Exhibit 6.1a Cross tabulation of Fault versus Guilt From Pugh's data on conviction for rape

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
N of Valid Cases	358				

[a] Computed only for a 2 × 2 table

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

Exhibit 6.1b Test statistics for analysis of Pugh's data

Symmetric Measures

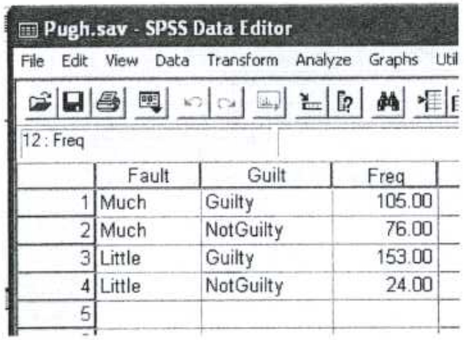
		Value	Approx. Sig.
Nominal by Nominal	Phi	.317	.000
	Cramer's V	.317	.000
	Contingency Coefficient	.302	.000
N of Valid Cases		358	

Exhibit 6.1c Measures of association for Pugh's data

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Fault (Little / Much)	4.614	2.738	7.776
For cohort Guilt = Guilty	1.490	1.299	1.709
For cohort Guilt = NotGuilty	.323	.214	.486
N of Valid Cases	358		

Exhibit 6.1d Risk estimates on Pugh's data



12 : Freq

	Fault	Guilt	Freq
1	Much	Guilty	105.00
2	Much	NotGuilty	76.00
3	Little	Guilty	153.00
4	Little	NotGuilty	24.00
5			

Exhibit 6.2 SPSS data file for analysis of Pugh's experiment

possible cell in the table, and then add a column (here labeled *Freq*) that reports how many observations fell in that cell. A screen shot of such a table is shown in Exhibit 6.2.

Once we have entered the frequencies, simply go to **Data/Weight cases** menu and instruct SPSS to weight each combination of Fault and Guilt by the Freq variable. Similar commands can be carried out in most software. The rest of the calculations can then be carried out just as we did earlier.

Exhibit 6.1b contains the printout of the test statistics for testing the null hypothesis of independence between Fault and Guilt. You can see that we obtained the same value of χ^2 (35.93) that we obtained earlier by hand. The next entry is the value of χ^2 with a continuity correction, as we discussed earlier. I suggest ignoring this. Fisher's Exact Test follows, and here it leads to the same conclusion as Pearson's chi-square. (You will not find Fisher's Exact Test printed out with larger tables because it was not designed for them nor will you see a chi-square value printed in column 2 because the test does not produce one.)

Small Expected Frequencies

small expected frequency

One of the most important requirements for using the Pearson 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.

For a given sample size, there are often a limited number of different contingency tables that you could obtain and, thus, a limited number of different values of chi-square. If only a few different values of χ^2_{obt} are possible, then the χ^2 distribution, which is continuous, cannot provide a reasonable approximation to the distribution of our statistic. We cannot closely 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 those with small expected frequencies in one or more cells. (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.)

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 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 five. This is a conservative position and I don't feel overly guilty when I violate it. Bradley and colleagues (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 .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 eight, 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 are inflated Type I error rates.

One major advantage of Fisher's Exact Test is that it is not based on the χ^2 distribution and, thus, is not affected by a lack of continuity. One of the strongest arguments for that test is that it applies well to cases with small expected frequencies.

6.5 Chi-Square for Ordinal Data

Chi-square is an important statistic for analyzing 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 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 relative 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, Section 10.4, and be able to follow the discussion.) I mention it here because it comes up most often when discussing χ^2.

6.6 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 more than 50 years since they were written. A somewhat more recent discussion of many of the issues Lewis and Burke (1949) raised 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 participant's choice among brands of coffee has no effect on another participant'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. So, we want the *observations* to be independent and we are testing the independence of *variables*.

It is not uncommon to find cases in which the assumption of independence of observations is violated, usually by having the same participant 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.

Inclusion of Nonoccurrences

Although the requirement that nonoccurrences be included has not yet been mentioned specifically, it is inherent in the derivation. It 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 if 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

nonoccurrences

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. However, the derivation of chi-square assumes that we have included both those opposed to DST and those in favor of it. So we need a table such as this one:

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

Now $\chi^2 = 4.29$, which is significant at $\alpha = .05$, 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 *2,000* 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 consider the nonoccurrences invalidates the test and 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.

6.7 One- and Two-Tailed Tests

People are often confused about 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 because either event would lead to a large value of χ^2.

The preceding discussion is not intended to start an argument about 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 weakness 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 want 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 $p < .05$.

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 you actually 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 Jennifer Mahon raised. 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.

6.8 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 size the standard Pearson chi-square is thought to be better approximated by the exact chi-square distribution than is the likelihood ratio chi-square (Agresti, 1990). Likelihood ratio tests are heavily used in log-linear models for analyzing contingency tables because of their additive 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, the general idea of a likelihood ratio can be described quite simply. Suppose we collect data and calculate the probability or 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 probable). 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," 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 6.16. It is a χ^2 on 1 *df*, and because it exceeds $\chi^2_{.05}(1) = 3.84$, it will lead to rejection of H_0. The decision of the juror depends in part on how the victim is portrayed.

6.9 Effect Sizes

The fact that a relationship is "statistically significant" doesn't tell us very much about whether it is of practical significance. The fact that two independent variables are not statistically independent does not mean that the lack of independence is important or worthy of our attention. In fact, if you allow the sample size to grow large enough, almost any two variables would likely show a statistically significant lack of independence.

d-family

r-family

measures of association

What we need, then, are ways to go beyond a simple test of significance to present one or more statistics that reflect the size of the effect we are looking at. There are two different types of measures designed to represent the size of an effect. One type, called the **d-family** by Rosenthal (1994), is based on one or more measures of the *differences* between groups or levels of the independent variable. For example, as we will see in a moment, the probability of being found guilty of rape is about 30% higher for dependents in Pugh's Low Fault condition than for those in the High Fault condition. The other type of measure, called the **r-family**, represents some sort of correlation coefficient between the two independent variables. We will discuss correlation thoroughly in Chapter 9, but I will discuss these measures here because they are appropriate at this time. Measures in the *r*-family are often called "measures of association."

An Example

prospective study

An important study of the beneficial effects of small daily doses of aspirin on reducing heart attacks in men was reported in 1988. More than 22,000 physicians were administered aspirin or a placebo, and the incidence of later heart attacks was recorded. The data follow in Table 6.4. Notice that this design is a **prospective study** because the treatments (aspirin versus no aspirin) were applied and then future outcome was determined. (A

Table 6.4 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

retrospective study

retrospective study would select people who had, or had not, experienced a heart attack and then look backward in time to see whether they had been in the habit of taking aspirin in the past.)

For these data, $\chi^2 = 25.014$ on one degree of freedom, which is statistically significant at $\alpha = .05$, indicating that there is a relationship between whether or not one takes aspirin daily and whether one later has a heart attack.[4]

d-family: Risks and Odds

Two important concepts with categorical data, especially for 2 × 2 tables, are the concepts of risks and odds. These concepts are closely related, and often confused, but they are basically very simple.

risk

For the aspirin data, 0.94% (104/11, 037) of people in the aspirin group and 1.71% (189/11, 034) of those in the control group suffered a heart attack during the study. (Unless you are a middle-aged male worrying about your health, the numbers look rather small. But they are important.) These two statistics are commonly referred to as **risk** estimates because they describe the risk that someone with, or without, aspirin will suffer a heart attack. Risk measures offer a useful way of looking at the size of an effect.

risk difference

The **risk difference** is simply the difference between the two proportions. In our example, the difference is 1.71% − 0.94% = 0.77%. Thus, there is about three-quarters of a percentage point difference between the two conditions. Put another way, the difference in risk between a male taking aspirin and one not taking aspirin is about three-quarters of 1%. This may not appear to be very large, but keep in mind that we are talking about heart attacks, which are serious events.

One problem with a risk difference is that its magnitude depends on the overall level of risk. Heart attacks are quite low risk events, so we would not expect a huge difference between the two conditions. (In contrast, when we looked at Pugh's data on convictions for rape, where the probability of being convicted was quite high, there was a lot of room for the two conditions to differ, and we saw a 30 percentage point difference. Does that mean that Pugh's study found a much larger effect size? Well, it depends—it certainly did with respect to risk difference.)

risk ratio

relative risk

Another way to compare the risks is to form a **risk ratio**, also called **relative risk**, that is just the ratio of the two risks. For the heart attack data, the risk ratio is

$$RR = Risk_{\text{no apsirin}}/Risk_{\text{aspirin}} = 1.71\%/0.94\% = 1.819$$

Thus, the risk of having a heart attack if you do not take aspirin is 1.8 times higher than if you do take aspirin. That strikes me as quite a difference.

odds

We must consider a third measure of effect size, and that is the odds ratio. At first glance, odds and odds ratios look like risk and risk ratios, and they are often confused, even by people who know better. (In a previous edition, I referred to odds, but described them as risks, much to my chagrin.) Recall that we defined the risk of a heart attack in the aspirin group as the number having a heart attack divided by the *total number of people in that group*. (e.g., $104/11{,}037 = 0.0094 = 0.94\%$.) The **odds** of having a heart attack for a member of the aspirin group is the number having a heart attack divided by the number *not having a heart attack*. (e.g., $104/10{,}933 = 0.0095$.) The difference (though very slight) comes in what we use as the denominator—risk uses the total sample size and is thus the proportion of people in that condition who experience a heart attack. Odds uses as a denominator the number not having a heart attack and is thus the ratio of the number having an attack versus the number not having an attack. Because the denominators are so much alike in this example, the results are almost indistinguishable. That is certainly not always the case. In Pugh's example, the risk of being convicted of rape in the low fault condition are $153/177 = 0.864$ (86% of the cases are convicted), whereas the odds of being convicted in the low fault condition are $153/24 = 6.375$ (the odds of being convicted are 6.4 times the odds of being found innocent).

odds ratio

Just as we can form a risk ratio by dividing the two risks, we can form an **odds ratio** by dividing the two odds. For the aspirin example, the odds of heart attack given that you did not take aspirin were $189/10{,}845 = 0.017$. The odds of a heart attack given that you did take aspirin were $104/10{,}933 = 0.010$. The odds ratio is simply the ratio of these two odds and is

$$OR = \frac{Odds \mid No\,Aspirin}{Odds \mid As\,pirin} = \frac{0.0174}{0.0095} = 1.83$$

Thus, the odds of a heart attack without aspirin are 1.83 times higher than the odds of a heart attack with aspirin.[5]

Why do we have to complicate things by having both odds ratios and risk ratios because they often look very much alike? That is a very good question, and it has some good answers. Risk is something that I think most of us understand. When we say the risk of having a heart attack in the No Aspirin condition is 0.0171, we are saying that 1.7% of the participants in that condition had a heart attack, and that is pretty straightforward. When we say that the odds of a heart attack in that condition are 0.0174, we are saying that the chances of having a heart attack are 1.7% of the chances of not having a heart attack. That may be a popular way of setting bets on race horses, but it leaves me dissatisfied. So why have an odds ratio in the first place?

The odds ratio has at least two things in its favor. In the first place, it can be calculated in situations in which a true risk ratio cannot be. In a retrospective study, where we find a group of people with heart attacks and another group of people without heart attacks, and look back to see if they took aspirin, we can't really calculate *risk*. Risk is future oriented. If we give 1,000 people aspirin and withhold it from 1,000 others, we can look at these people 10 years down the road and calculate the risk (and risk ratio) of heart attacks. But if

[4] It is important to note that, although taking aspirin daily is associated with a lower rate of heart attack, more recent data have shown that there are important negative side effects. Current literature suggests that Omega-3 fish oil is at least as effective with fewer side effects.

[5] In computing an odds ratio, there is no rule about which odds go in the numerator and which in the denominator. It depends on convenience. Where reasonable, I prefer to put the larger value in the numerator to make the ratio come out greater than 1.0, simply because I find it easier to talk about that way. If we reversed them in this example, we would find OR = 0.546, and conclude that your odds of having a heart attack in the aspirin condition are about half of what they are in the No Aspirin condition. That is simply the inverse of the original OR (0.546 = 1/1.83).

we take 1,000 people with (and without) heart attacks and look backward, we can't really calculate risk because we have sampled heart attack patients at far greater than their normal rate in the population (50% of our sample has had a heart attack, but certainly 50% of the population does not suffer from heart attacks). But we can always calculate odds ratios. And, when we are talking about low probability events, such as having a heart attack, the odds ratio is usually a very good estimate of what the risk ratio would be.[6] The odds ratio is equally valid for prospective, retrospective, and cross-sectional sampling designs. That is important.

A second important advantage of the odds ratio is that taking the natural log of the odds ratio [ln(OR)] gives us a statistic that is extremely useful in a variety of situations. Two of these are logistic regression and log-linear models, both of which are discussed later in the book. I don't expect most people to be excited by the fact that a logarithmic transformation of the odds ratio has interesting statistical properties, but that is a very important point nonetheless.

r-family: Phi and Cramér's V

The measures that we have discussed are sometimes called *d*-family measures because they focus on comparing differences between conditions—either by calculating the difference directly or by using ratios of risks or odds. An older, and more traditional set of measures, sometimes called "measures of association," but now frequently called "*r*-family measures" looks at the correlation between two variables. We won't come to correlation until Chapter 9, but I would expect that you already know enough about correlation to understand what follows.

There are a great many measures of association, and I have no intention of discussing most of them. One of the nicest discussions of these can be found in Nie, Hull, Jenkins, Steinbrenner, and Bent (1970). (If your instructor is very old—like me—he or she probably remembers it fondly as the old "maroon SPSS manual." It is such a classic that it is very likely to be available in your university library or through interlibrary loan.)

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 (a dichotomy is a variable that takes on one of two distinct values.). If we coded Aspirin as 1 or 2, for Yes and No, and coded Heart Attack as 1 for Yes and 2 for No, and then correlated the two variables (see Chapters 9 and 10), the result would be phi. (It doesn't even matter what two numbers we use as values for coding, as long as one condition always gets one value and the other always gets a different, but consistent, value.)

An easier way to calculate ϕ for these data is by the relation

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

For the aspirin data in Table 6.4, $\chi^2 = 25.014$, $\phi = \sqrt{25.014/22{,}071} = .034$. That does not appear to be a very large correlation, but we are speaking about a major life-threatening event, and even a small correlation can be meaningful.

[6] The odds ratio can be defined as $\text{OR} = \text{RR}(\frac{1-p_2}{1-p_1})$, where OR = odds ratio, RR = relative risk, p_1 is the population proportion of heart attacks in one group, and p_2 is the population proportion of heart attacks in the other group. When these two proportions are close to 0, numerator and denominator nearly cancel each other and OR = RR.

Cramér's V

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

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

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

Cramér's V

Cramér's V can be seen as a simple extension of ϕ. Note that when $k = 2$, it *is* ϕ. Its usefulness applies to larger tables. We can calculate Cramér's V for the data on littering in Table 6.3 as follows:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{25.79}{(1772)(1)}} = .121$$

The problem with V is that it is hard to give it a simple intuitive interpretation when there are more than two categories and they do not fall on an ordered dimension. There is a fairly technical explanation, but I am not going into it here, and I doubt that it would be very enlightening at this point.

I am not happy with the *r*-family of measures simply because I don't think that they have a meaningful interpretation in most situations. It is one thing to use a *d*-family measure like the odds ratio and declare that the odds of having a heart attack if you don't take aspirin are 1.83 times higher than the odds of having a heart attack if you do not take aspirin. I think that most people can understand what that statement means. But to use an *r*-family measure, such as phi, and say that the correlation between aspirin intake and heart attack is .034 doesn't seem to be telling them anything useful. (And squaring it and saying that aspirin usage accounts for 0.1% of the variance in heart attacks is even less helpful.) I would suggest that you stay away from the older *r*-family measures uhless you really have a good reason to use them.

Effect Sizes for Larger Tables

Measures like odds ratios are most easily understood with 2 × 2 tables because it is clear what the odds represent. Things are very much messier with larger tables. We will see this distinction between two levels and multiple levels in several places in this book. If you think clearly about what it is you want to convey to your audience, I suspect that you will generally find that you really want to compare only two things. For example, in the littering study, you might want to compare the number of flyers littering the floor with the number of flyers that took themselves off some place—such as the trash or out of the store. I would suggest that after computing the overall chi-square for the 2 × 3 table, you simply recompile your contingency table into "Litter" and "Non-litter" and treat it as a 2 × 2. That is really what you probably want. And if that is the case, risk ratios and odds ratios will do very nicely. (When we come to the analysis of variance in Chapter 11, which looks a million miles away from contingency tables, you will see that frequently the questions we most care about also come down to comparing two groups or sets of groups.)

6.10 Measures of Agreement

We should discuss one more measure. It is not really a measure of effect size, like the previous measures, but it is an important statistic when you want to ask about the agreement between judges.

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 want 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 6.5.

Table 6.5 Classification of behavior problems by two judges

	Judge I			
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

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 column 1. 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 entries on the diagonal (15, 3, 3) represent agreement between the two judges, whereas the off-diagonal entries represent disagreement.

percentage of agreement

A simple (but unwise) 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 $21/30 = 0.70 = 70\%$ agreement. This measure has problems, however. Most 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 $16/30 = .53$. The probability of No Problem for Judge II would be estimated as $20/30 = .67$. If the two judges operated by pulling their diagnoses out of the air, the probability that they would both classify the same case as No Problem is $.53 \times .67 = .36$, which for 30 judgments would mean that $.36 \times 30 = 10.67$ 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 standard chi-square test. For example, the expected frequency of both judges assigning a classification of No Problem, assuming that they are operating at random, 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

$$\sum f_O = 15 + 3 + 3 = 21$$

and

$$\sum f_E = 10.67 + 1.20 + 1.07 = 12.94.$$

Then

$$\kappa = \frac{21 - 12.94}{30 - 12.94} = \frac{8.06}{17.06} = .47$$

Notice that this coefficient is considerably lower than the 70% agreement figure that we just calculated. 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 rarely 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.

6.11 Writing Up the Results

We will take as our example Pugh's study of rape convictions (1983). If you were writing up these results, you would probably want to say something like the following:

> In examining the question of whether a defense lawyer's attempt to place blame on the victim of rape would influence a jury's decision in a rape case, jury participants were presented with a situation in which the victim was characterized by the defense as either partly responsible for the rape or not responsible. The jurors were then asked to make a judgment about whether the defendant was guilty or not guilty of the crime. When the victim was portrayed as low in fault, 86% of the time the defendant was judged to be guilty. When the victim was portrayed as high in fault, the defendant was judged guilty only 58% of the time. A chi-square test of the relationship between Fault and Guilt produced $\chi^2(1) = 35.93$, which is statistically significant at $p < .05$. This is associated with an odds ratio of 4.61, indicating that the odds of being found guilty of rape are more than 4.5 times higher in the condition in which the victim is portrayed as not bearing fault for the rape. The odds ratio would indicate that we are speaking of a meaningful difference between the two conditions.

Key Terms

Chi-square (χ^2) (Introduction)

Pearson's chi-square (Introduction)

Chi-square distribution (χ^2) (6.1)

Gamma function (6.1)

Chi-square test (6.2)

Goodness-of-fit test (6.2)

Observed frequencies (6.2)

Expected frequencies (6.2)

Tabled distribution of χ^2 (6.2)

Contingency table (6.3)

Marginal totals (6.3)

Cell (6.3)

Row total (6.3)

Column total (6.3)

Yates's correction for continuity (6.4)

Fixed marginals (6.4)

Small expected frequency (6.4)

Assumptions of χ^2 (6.6)

Nonoccurrences (6.6)

Likelihood ratios (6.8)

d-family (6.9)

r-family (6.9)
Measures of association (6.9)
Prospective study (6.9)
Retrospective study (6.9)
Risk (6.9)
Risk difference (6.9)
Risk ratio (6.9)
Relative risk (6.9)
Odds (6.9)
Odds ratio (6.9)
Phi (ϕ) (6.9)
Cramér's V (6.9)
Kappa (κ) (6.10)
Percentage of agreement (6.10)

Exercises

6.1 The chairperson of a psychology department suspects that some of her faculty members are more popular with students than are others. There are three sections of introductory psychology, taught at 10:00 a.m., 11:00 a.m., and 12:00 p.m. by Professors Anderson, Klatsky, and Kamm. The number of students who enroll for each is

Professor Anderson	Professor Klatsky	Professor Kamm
32	25	10

State the null hypothesis, run the appropriate chi-square test, and interpret the results.

6.2 From the point of view of designing a valid experiment (as opposed to the arithmetic of calculation), there is an important difference between Exercise 6.1 and the examples used in this chapter. The data in Exercise 6.1 will not really answer the question the chairperson wants answered. What is the problem, and how could the experiment be improved?

6.3 You have a theory that if you ask subjects to sort one-sentence characteristics of people (e.g., "I eat too fast") into five piles ranging from "not at all like me" to "very much like me," the percentage of items placed in each of the five piles will be approximately 10, 20, 40, 20, and 10. You have one of your friend's children sort 50 statements, and you obtain the following data: [8, 10, 20, 8, 4]. Do these data support your hypothesis?

6.4 To what population does the answer to Exercise 6.3 generalize? (*Hint:* From what population of observations might these observations be thought to be randomly sampled?)

6.5 In a classic study by Clark and Clark (1939), African American children were shown black dolls and white dolls and were asked to select the one with which they wanted to play. Of 252 children, 169 chose the white doll and 83 chose the black doll. What can we conclude about the behavior of these children?

6.6 Thirty years after the Clark and Clark study, Hraba and Grant (1970) repeated the study referred to in Exercise 6.5. The studies were not exactly equivalent, but the results were interesting. Hraba and Grant found that of 89 African American children, 28 chose the white doll and 61 chose the black doll. Run the appropriate chi-square test on their data and interpret the results.

6.7 Combine the data from Exercises 6.5 and 6.6 into a two-way contingency table and run the appropriate test. How does the question that the two-way classification addresses differ from the questions addressed by Exercises 6.5 and 6.6?

6.8 We know that smoking has all sorts of ill effects on people; among other things, there is evidence that it affects fertility. Weinberg and Gladen (1986) examined the effects of smoking and the ease with which women become pregnant. The researchers asked 586 women who had planned pregnancies how many menstrual cycles it had taken for them to become pregnant after discontinuing contraception. Weinberg and Gladen also sorted the women into whether they were smokers or nonsmokers. The data follow.

	1 cycle	2 cycles	3+ cycles	Total
Smokers	29	16	55	100
Nonsmokers	198	107	181	486
Total	227	123	236	586

Does smoking affect the ease with which women become pregnant? (I do not recommend smoking as a birth control device, regardless of your answer.)

6.9 In discussing the correction for continuity, we referred to the idea of fixed marginals, meaning that a replication of the study would produce the same row and column totals. Give an example of a study in which

a. No marginal totals are fixed.

b. One set of marginal totals is fixed.

c. Both sets of marginal totals (row and column) could reasonably be considered to be fixed. (This is a hard one.)

6.10 Howell and Huessy (1981) used a rating scale to classify children in a second-grade class as showing or not showing behavior commonly associated with attention deficit disorder (ADD). The researchers then classified these same children again when they were in fourth and fifth grades. When the children reached the end of the ninth grade, the researchers examined school records and noted which children were enrolled in remedial English. In the following data, all children who were ever classified as exhibiting behavior associated with ADD have been combined into one group (labeled ADD):

	Remedial English	Nonremedial English	
Normal	22	187	209
ADD	19	74	93
	41	261	302

Does behavior during elementary school affect class assignment during high school?

6.11 Use the data in Exercise 6.10 to demonstrate how chi-square varies as a function of sample size.

a. Double each cell entry and recompute chi-square.

b. What does your answer to (a) say about the role of the sample size in hypothesis testing?

6.12 In Exercise 6.10, children were classified as those who never showed ADD behavior and those who showed ADD behavior at least once in the second, fourth, or fifth grade. If we do not collapse across categories, we obtain the following data:

	Never	2nd	4th	2nd & 4th	5th	2nd & 5th	4th & 5th	2nd, 4th, & 5th
Remedial	22	2	1	3	2	4	3	4
Nonrem.	187	17	11	9	16	7	8	6

a. Run the chi-square test.

b. What would you conclude, ignoring the small expected frequencies?

c. How comfortable do you feel with these small expected frequencies? If you are not comfortable, how might you handle the problem?

6.13 In 2000, the State of Vermont legislature approved a bill authorizing "civil unions" between gay or lesbian partners. This was a very contentious debate with very serious issues raised by both sides. How the vote split along gender lines may tell us something important about the different ways in which males and females looked at this issue. The data follow. What would you conclude from these data?

	Vote		
	Yes	No	Total
Women	35	9	44
Men	60	41	101
Total	95	50	145

6.14 Stress has long been known to influence physical health. Visintainer, Volpicelli, and Seligman (1982) investigated the hypothesis that rats given 60 trials of inescapable shock would be less likely later to reject an implanted tumor than would rats that had received 60 trials of escapable shock or 60 no-shock trials. The researchers obtained the following data:

	Inescapable Shock	Escapable Shock	No Shock	
Reject	8	19	18	45
No Reject	22	11	15	48
	30	30	33	93

What could Visintainer et al. conclude from the results?

6.15 Darley and Latané (1968) asked subjects to participate in a discussion carried on over an intercom. Aside from the experimenter to whom they were speaking, subjects thought that there were zero, one, or four other people (bystanders) also listening over intercoms. Partway through the discussion, the experimenter feigned serious illness and asked for help. Darley and Latané noted how often the subject sought help for the experimenter as a function of the number of supposed bystanders. The data follow:

		Sought Assistance		
		Yes	No	
Number of Bystanders	0	11	2	13
	1	16	10	26
	4	4	9	13
		31	21	52

What could Darley and Latané conclude from the results?

6.16 In a study similar to the one in Exercise 6.15, Latané and Dabbs (1975) had a confederate enter an elevator and then "accidentally" drop a handful of pencils. They then noted whether bystanders helped pick them up. The data tabulate helping behavior by the gender of the bystander:

	Gender of Bystander		
	Female	Male	
Help	300	370	670
No Help	1003	950	1953
	1303	1320	2623

What could Latané and Dabbs conclude from the data? (Note that when we collapse over gender, only about one-quarter of the bystanders helped. That is not relevant to the question, but it is an interesting finding that could easily be missed by routine computer-based analyses.)

6.17 In a study of eating disorders in adolescents, Gross (1985) asked each of her subjects whether they would prefer to gain weight, lose weight, or maintain their present weight. (*Note:* Only 12% of the girls in Gross's sample were actually more than 15% above their normative weight—a common cutoff for a label of "overweight.") When she broke down the data for girls by race (African American versus white), she obtained the following results (other races have been omitted because of small sample sizes):

	Reducers	Maintainers	Gainers	
White	352	152	31	535
African American	47	28	24	99
	399	180	55	634

a. What conclusions can you draw from these data?

b. Ignoring race, what conclusion can you draw about adolescent girls' attitudes toward their own weight?

6.18 Use the likelihood ratio approach to analyze the data in Exercise 6.10.

6.19 Use the likelihood ratio approach to analyze the data in Exercise 6.12.

6.20 Would it be possible to calculate a one-way chi-square test on the data in row 2 of the table in Exercise 6.12? What hypothesis would you be testing if you did that? How would that hypothesis differ from the one you tested in Exercise 6.12?

6.21 Suppose we asked a group of 40 subjects whether they liked Monday Night Football, made them watch a game, and then asked them again. We would record the data as follows:

	Pro	Con	
Before	30	10	40
After	15	25	40
	45	35	80

Would chi-square calculated on such a table be appropriate? Why or why not?

6.22 As an alternative approach to the data in Exercise 6.21, you might find that after watching the game 20 people switched from Pro to Con and 5 people switched from Con to Pro. Thus, you can run a one-way chi-square test on the $20 + 5 = 25$ subjects who changed their opinion. (This is a test suggested by McNemar (1969) and is often referred to as McNemar's test.)

a. Run the test.

b. Explain how this tests the null hypothesis that you wanted to test.

6.23 From the SPSS printout in Exhibit 6.3

a. Verify the answer to Exercise 6.17a.

b. Interpret the row and column percentages.

c. What are the values labeled "Asymp. Sig."?

d. Interpret the coefficients.

RACE*GOAL Crosstabulation

			Goal			Total
			Gain	Lose	Maintain	
RACE	African-Amer	Count	24	47	28	99
		Expected Count	8.6	62.3	28.1	99.0
		% within RACE	24.2%	47.5%	28.3%	100.0%
		% within GOAL	43.6%	11.8%	15.6%	15.6%
		% of Total	3.8%	7.4%	4.4%	15.6%
	White	Count	31	352	152	535
		Expected Count	46.4	336.7	151.9	535.0
		% within RACE	5.8%	65.8%	28.4%	100.0%
		% within GOAL	56.4%	88.2%	84.4%	84.4%
		% of Total	4.9%	55.5%	24.0%	84.4%
Total		Count	55	399	180	634
		Expected Count	55.0	399.0	180.0	634.0
		% within RACE	8.7%	62.9%	28.4%	100.0%
		% within GOAL	100.0%	100.0%	100.0%	100.0%
		% of Total	8.7%	62.9%	28.4%	100.0%

Exhibit 6.3 Continued

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	37.229[a]	2	.000
Likelihood Ratio	29.104	2	.000
N of Valid Cases	634		

[a] 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.59.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.242	.000
	Cramer's V	.242	.000
	Contingency Coefficient	.236	.000
N of Valid Cases		634	

Exhibit 6.3 Continued

6.24 A more complete set of data on heart attacks and aspirin, from which Table 6.4 was taken, follows. Here we distinguish not just between Heart Attacks and No Heart Attacks, but also between Fatal and NonFatal attacks.

	Myocardial Infarction			
	Fatal Attack	NonFatal Attack	No Attack	Total
Placebo	18	171	10,845	11,034
Aspirin	5	99	10,933	11,037
Total	23	270	21,778	22,071

a. Calculate both Pearson's chi-square and the likelihood ratio chi-square table. Interpret the results.

b. Using only the data for the first two columns (those subjects with heart attacks), calculate both Pearson's chi-square and the likelihood ratio chi-square and interpret your results.

c. Combine the Fatal and NonFatal heart attack columns and compare the combined column against the No Attack column, using both Pearson's and likelihood ratio chi-squares. Interpret these results.

d. Sum the Pearson chi-squares in (b) and (c) and then the likelihood ratio chi-squares in (b) and (c), and compare each of these results with the results in (a). What do they tell you about the partitioning of chi-square?

e. What do these results tell you about the relationship between aspirin and heart attacks?

6.25 For the results in Exercise 6.24, calculate and interpret

a. Cramér's V

b. Useful odds ratios

6.26 Compute the odds ratio for the data in Exercise 6.10. What do these values mean?

6.27 Compute the odds ratios for the data in Exercise 6.13. What do these ratios add to your understanding of the phenomena being studied?

6.28 Compute the odds in favor of seeking assistance for each of the groups in Exercise 6.15. Interpret the results.

6.29 Dabbs and Morris (1990) examined archival data from military records to study the relationship between high testosterone levels and antisocial behavior in males. Of 4016 men in the Normal Testosterone group, 10.0% had a record of adult delinquency. Of 446 men in the High Testosterone group, 22.6% had a record of adult delinquency. Is this relationship significant?

6.30 What is the odds ratio in Exercise 6.29? How would you interpret it?

6.31 In the study described in Exercise 6.29, 11.5% of the Normal Testosterone group and 17.9% of the High Testosterone group had a history of childhood delinquency.

a. Is there a significant relationship between these two variables?

b. Interpret this relationship.

c. How does this result expand on what we already know from Exercise 6.29?

6.32 In a study examining the effects of individualized care of youths with severe emotional problems, Burchard and Schaefer (1990, personal communication) proposed to have caregivers rate the presence or absence of specific behaviors for each of 40 adolescents on a given day. To check for rater reliability, the researchers asked two raters to rate each adolescent. The following hypothetical data represent reasonable results for the behavior of "extreme verbal abuse."

	Rater A		
Rater B	Presence	Absence	
Presence	12	2	14
Absence	1	25	26
	13	27	40

a. What is the percentage of agreement for these raters?

b. What is Cohen's kappa?

c. Why is kappa noticeably less than the percentage of agreement?

d. Modify the raw data, keeping N at 40, so that the two statistics move even farther apart. How did you do this?

6.33 Many school children receive instruction on child abuse around the "good touch-bad touch" model, with the hope that such a program will reduce sexual abuse. Gibson and Leitenberg (2000) collected data from 818 college students, and recorded whether they had ever received such training and whether they had subsequently been abused. Of the 500 students who had received training, 43 reported that they had subsequently been abused. Of the 318 who had not received training, 50 reported subsequent abuse.

a. Do these data present a convincing case for the efficacy of the sexual abuse prevention program?

b. What is the odds ratio for these data, and what does it tell you?

Computer Exercises

6.34 In a data set named Mireault.dat and described in Appendix: Computer Data Sets, Mireault (1990) collected data from college students on the effects of the death of a parent. Leaving the critical variables aside for a moment, let's look at the distribution of students. The data set contains information on the gender of the students and the college (within the university) in which they were enrolled.

a. Use any statistical package to tabulate Gender against College.

b. What is the chi-square test on the hypothesis that College enrollment is independent of Gender?

c. Interpret the results.

6.35 When we look at the variables in Mireault's data, we will want to be sure that there are no systematic differences of which we are ignorant. For example, if we found that the gender of the parent who died was an important variable in explaining some outcome variable, we would not like to later discover that the gender of the parent who died was in some way related to the gender of the subject, and that the effects of the two variables were confounded.

a. Run a chi-square test on these two variables.

b. Interpret the results.

c. What would it mean to our interpretation of the relationship between gender of the parent and some other variable (e.g., subject's level of depression) if the gender of the parent is itself related to the gender of the subject?

6.36 Zuckerman, Hodgins, Zuckerman, and Rosenthal (1993) surveyed more than 500 people and asked a number of questions on statistical issues. In one question a reviewer warned a researcher that she had a high probability of a Type I error because she had a small sample size. The researcher disagreed. Subjects were asked, "Was the researcher correct?" The proportions of respondents, partitioned among students, assistant professors, associate professors, and full professors, who sided with the researcher and the total number of respondents in each category were as follows:

	Students	Assistant Professors	Associate Professors	Full Professors
Proportion	.59	.34	.43	.51
Sample size	17	175	134	182

(*Note*: These data mean that 59% of the 17 students who responded sided with the researcher. When you calculate the actual obtained frequencies, round to the nearest whole person.)

a. Would you agree with the reviewer or with the researcher? Why?

b. What is the error in logic of the person you disagreed with in (a)?

c. How would you set up this problem to be suitable for a chi-square test?

d. What do these data tell you about differences among groups of respondents?

6.37 The Zuckerman et al. paper referred to in the previous question hypothesized that faculty were less accurate than students because they have a tendency to give negative responses to such questions. ("There must be a trick.") How would you design a study to test such a hypothesis?

Discussion Questions

6.38 Hout, Duncan, and Sobel (1987) reported data on the relative sexual satisfaction of married couples. They asked each member of 91 married couples to rate the degree to which they agreed with "Sex is fun for me and my partner" on a four-point scale ranging from "never or occasionally" to "almost always." The data appear here:

	Wife's Rating				
Husband's Rating	Never	Fairly Often	Very Often	Almost Always	TOTAL
Never	7	7	2	3	19
Fairly Often	2	8	3	7	20
Very Often	1	5	4	9	19
Almost Always	2	8	9	14	33
TOTAL	12	28	18	33	91

a. How would you go about analyzing these data? Remember that you want to know more than just whether or not the two ratings are independent. Presumably you would like to show that as one spouse's ratings go up, so do the other's, and vice versa.

b. Use both Pearson's chi-square and the likelihood ratio chi-square.

c. What does Cramér's V offer?

d. What about odds ratios?

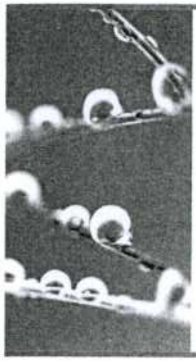
e. What about kappa?

f. Finally, what if you combined the Never and Fairly Often categories and the Very Often and Almost Always categories? Would the results be clearer, and under what conditions might this make sense?

6.39 In the previous question, we were concerned with whether husbands and wives rate their degree of sexual fun congruently (i.e., to the same degree). But suppose that women have different cut points on an underlying scale of "fun." For example, maybe women's idea of Fairly Often or Almost Always is higher than men's. (Maybe men would rate "a couple of times a month" as "Very Often" whereas women would rate "a couple of times a month" as "Fairly Often.") How would this affect your conclusions? Would it represent an underlying incongruency between males and females?

6.40 Use SPSS or another statistical package to calculate Fisher's Exact Test for the data in Exercise 6.13. How does it compare to the probability associated with Pearson's chi-square?

SIXTH EDITION



Statistical Methods for Psychology

David C. Howell
University of Vermont

THOMSON
WADSWORTH

Australia • Brazil • Canada • Mexico • Singapore • Spain
United Kingdom • United States