Chapter 2: Analysis of Categorical Data

2.1 Introduction

Categorical, or nominal data is most often encountered when observations are grouped into discrete, mutually exclusive categories (i.e., each observation can fall into one and only one category), and one examines the frequency of occurrence for each of the categories. The most common statistical test for such data is some form of a chi-square test. (The *ch* in *chi* is pronounced like a *k*, and *chi* rhymes with *eye*.)

2.2 One-way classification: Chi-square "goodness of fit" test

Let us begin with an example: You have been hired by a supermarket chain to conduct some "Pepsi-challenge" market research. Let us assume that you asked 150 shoppers to sample 3 different brands of Cola: Coke, Pepsi, and X (a No-Name brand produced for the supermarket). (We will assume that you were a careful experimenter, and counterbalanced the order in which the 3 Colas were presented to subjects, and that subjects were blind as to what they were drinking.) Each participant had to indicate which of the 3 Colas they preferred, and the data looked like this:

Table 2.1 Observed frequencies for Pepsi-challenge problem

Coke	Pepsi	Brand X	Total
45	40	65	150

So in your sample of 150, more people preferred Brand X to either Coke or Pepsi. But can we conclude from this that more people in the *population* of interest prefer Brand X? Not necessarily. It could be that there are no differences between Coke, Pepsi, and Brand X in the population, and that the differences we see in this sample are due to sampling error. Fortunately, we can evaluate this possibility with a *chi-squared* test.

The *chi-squared* test is based on the *difference between observed frequencies, and the frequencies that are expected if the null hypothesis is true*. The null hypothesis *often* states that the frequencies will be equal in all categories, but not always. In this case, let's assume that it does. Therefore, the expected frequencies would be:

Table 2.2 Expected frequencies for Pepsi-challenge problem

Coke	Pepsi	Brand X	Total
50	50	50	150

The X^2 statistic can be computed with the following formula:

$$X^2 = \sum \frac{(O-E)^2}{E} \tag{2.1}$$

Note that many textbooks call the statistic calculated with this formula χ^2 rather than X^2. Siegel and Castellan (1988) use X^2 to emphasise the distinction between the *observed value of the statistic* (X^2) and the *theoretical probability distribution that is its (asymptotic) sampling distribution under a true null hypothesis* (χ^2). That is, if H_0 is true (and certain conditions/assumptions[1] are met), the χ^2 distribution with df = k-1 provides a pretty good approximation to the sampling distribution of X^2. Therefore, we can use the χ^2 distribution with df = k-1 to obtain the probability of getting the observed value of X^2, or a larger value, given that the null hypothesis is true. This conditional probability is the p-value, of course. And if the p-value is small enough (.05 or less, usually), we can reject the null hypothesis.

Table 2.3 Calculation of X^2 for the Pepsi-challenge problem

Observed (O)	*Expected (E)*	*O-E*	*$(O-E)^2$*	*$(O-E)^2/E$*
45	50	-5	25	0.5
40	50	-10	100	2.0
65	50	15	225	4.5
150	150	0		**7.0**

For the Pepsi challenge example, X^2 (*df*=2, *n*=150) = 7.0, p = 0.030 (see Table 2.3 for the calculations). Therefore, we can reject the null hypothesis that all 3 brands are preferred equally in the population from which we have sampled.

2.3 Unequal proportions under the null hypothesis

As mentioned earlier, the null hypothesis does not always state that the same number of observations is expected in each category. For example, a geneticist might know that in the fruit fly population, 4 different sub-types of fruit flies appear in the ratio 4:3:2:1. So if a sample of 100 fruit flies was randomly selected from this population, the expected frequencies *according to the null hypothesis* would be 40, 30, 20, and 10. Would the geneticist be able to reject the null hypothesis (that the sample was randomly drawn from this population) if the observed frequencies were 44, 36, 12, and 8? Let's work it out and see.

The calculations shown in Table 2.4 reveal that X^2 (*df*=3, *n*=100) = 5.2, *p* = 0.158. Therefore, the geneticist would not have sufficient evidence to allow rejection of the null hypothesis.

[1] We will discuss those assumptions a bit later.

Table 2.4 Calculation of X^2 for the fruit fly problem

Observed (O)	*Expected (E)*	*O-E*	*(O-E)²*	*(O-E)²/E*
44	40	4	16	0.4
36	30	6	36	1.2
12	20	-8	64	3.2
8	10	-2	4	0.4
100	100	0		**5.2**

2.4 *Chi-square* test of independence (or association)

So far we have considered cases with one categorical variable. We now move on to another common use of *chi-squared* tests: A test of **independence between two categorical variables**. In this case, it is common to present the data in a **contingency table**. The levels of one variable are represented on different rows, and the levels of the other variable in different columns. Let us assume, for example, that we have a problem-solving task that requires subjects to use a screw-driver as a pendulum. We randomly assign 90 subjects to 3 groups of 30. One group is given *no special instructions* (a_1); a second group is asked to *list 5 common uses of screwdrivers* (a_2); and the third group is asked to *list 5 uncommon uses of screwdrivers* (a_3). Then all subjects are given the problem, and each one is categorised as to whether or not they solved it. The frequency data look like this:

Table 2.5 Observed frequencies for the "screwdriver" experiment

	a₁	***a₂***	***a₃***	**Total**
Solve	9	17	22	48
Fail to solve	21	13	8	42
Total	30	30	30	90

The null hypothesis states that the proportion of subjects solving the problem is **independent of instructions**.[2] In other words, we expect the proportions of solvers and non-solvers to be the same in all 3 groups. If that is so, then based on the marginal (row) totals, we expect 48/90, or about 53% of subjects from each group to solve the problem. The general rule for calculating the *E*, the expected frequency for a cell under the null hypothesis, is:

$$E = \frac{(\text{row total})(\text{column total})}{\text{grand total}} \tag{2.2}$$

[2] The null hypothesis states that the two variables are independent of each other, so the alternative must state that they are associated. Hence the two names for the same test: test of *independence*, and test of *association*.

In terms of frequencies then, we expect 30x48/90 = 16 subjects in each group to solve the problem; and 30x42/90 = 14 to **not** solve it. Using these expected frequencies, we can calculate X^2 as before: $X^2 = 11.5178$ (see Table 2.6).

For a *chi-square* test of independence, the number of degrees of freedom is given by:

$$df = (r-1)(c-1) \tag{2.3}$$

where r = the number of rows, and c = the number of columns. For this problem, therefore, df = (2-1)(3-1) = 2. So the p-value for this example can be obtained using the χ^2 distribution with α = .05 and df = 2. Using SPSS, I found that $p = 0.003$. Therefore, I would report the results of my analysis as follows: : X^2 (df=2, n=90) = 11.5178, p = 0.003. Because the p-value is lower than the conventional alpha level of 0.05, we can reject H_0.

Table 2.6 Calculation of X^2 for the fruit fly problem

Observed (O)	***Expected (E)***	***O-E***	***(O-E)^2***	***(O-E)^2/E***
9	16	-7	49	3.0625
17	16	1	1	0.0625
22	16	6	36	2.2500
21	14	6	49	3.5000
13	14	1	1	0.0714
8	14	-6	36	2.5714
90	90	0		**11.5178**

2.5 Alternative to Pearson's X^2: Likelihood ratio tests

There is an alternative to Pearson's *chi-square* that has been around since the 1950s. It is based on likelihood ratios (which we have not discussed), and is often referred to as the *maximum likelihood chi-square*. Note that this statistic is calculated by both Statview (where it is called the "G" statistic) and **SPSS** (where it is called the "Likelihood chi-square", and symbolized G^2 in some of the manuals). Following Howell (1992, 1997), I will use L^2 to symbolise the statistic that is calculated. There are various ways to calculate L^2, but the easiest is probably the way given by Howell. The formula is as follows:

$$L^2 = 2 \sum_{\text{all cells}} O \ln(O/E) \tag{2.4}$$

where: *ln* = natural logarithm (see Appendix A)
O = observed frequency for a cell
E = expected frequency for a cell.

L^2 and X^2 (computed with Pearson's formula) usually yield slightly different values. Nevertheless, for practical purposes, L^2 is distributed as χ^2 with $df = k\text{-}1$ for one-way classification (goodness of fit) problems, and $df = (r\text{-}1)(c\text{-}1)$ for tests of independence. Thus it can be used in any situation where Pearson's formula could be used. (Pearson's formula has remained prominent and popular because it was computationally easier in the days before electronic calculators and computers.)

One might wonder if there is any advantage to using L^2 rather than X^2. For the moment, let me to simply assure you that there are some advantages. We will be in a better position to discuss them after discussing two or three other topics. But first, let us redo two of the problems we looked at earlier using L^2 as our statistic.

As you can see by comparing Tables 2.3 and 2.7, L^2 and Pearson's X^2 yield slightly different values for the Pepsi-challenge data. L^2 (df=2, n=150) = 6.7736, p = 0.034; and X^2 (df=2, n=150) = 7.0, p = 0.030. The p-values differ only in the 3rd decimal place, and in both cases, we can reject the null hypothesis.

Table 2.7 Calculation of L^2 for the Pepsi-challenge problem

Observed (O)	*Expected (E)*	*ln(O/E)*	*O ln(O/E)*		
45	50	-0.1054	-4.7412		
40	50	-0.2231	-8.9257		
65	50	0.2624	17.0537		
150	150		3.3868	L^2 =	**6.7736**

The steps in calculating L^2 for the "screwdriver" problem are shown in Table 2.8. Again, the values of L^2 and X^2 are somewhat different: L^2 (df=2, n=90) = 11.866, p = 0.003; and X^2 (df=2, n=90) = 11.518, p = 0.003. In this case, the p-values for both tests are the same (to 3 decimal places), and both tests lead to rejection of the null hypothesis.

Table 2.8 Calculation of L^2 for the "screwdriver" problem

Observed (O)	*Expected (E)*	*ln(O/E)*	*O ln(O/E)*		
9	16	-0.5754	-5.1783		
17	16	0.0606	1.0306		
22	16	0.3185	7.0060		
21	14	0.4055	8.5148		
13	14	-0.0741	-0.9634		
8	14	-0.5596	-4.4769		
90	90		5.9328	L^2 =	**11.8656**

2.6 Additivity of independent χ^2-distributed variables

It is a fact that the sum of two independent *chi-squares* with v_1 and v_2 degrees of freedom respectively is itself a *chi-square* with $df = v_1 + v_2$. Note also that this additivity can be extended to any number of *chi-squares*, provided that they are all independent of each other.

2.7 Partitioning an overall *chi-square*: One-way classification

From section 2.6, it follows that we should be able to take a chi-square with df =v (when $v > 1$), and partition it into v independent chi-squares, each with df = 1. I hope that the reason why we might want to do this will become clear through a couple of examples.

To begin, let us return again to the Pepsi-challenge problem. The L^2 value for that problem was 6.7736 (see Table 2.7) with df = 2. We ought to be able to partition this overall L^2 into two independent L^2 values with 1 degree of freedom each. With one of these partitions, we could compare Coke and Pepsi to see if the proportion of people choosing them is different; and with the second, we could compare Coke and Pepsi combined to Brand X.

Table 2.9 Calculation of L^2 for Coke vs Pepsi comparison

Observed (O)	*Expected (E)*	*ln(O/E)*	*O ln(O/E)*		
45	42.5	0.0572	2.5721		
40	42.5	-0.0606	-2.4250		
85	85		0.1471	L^2 =	**0.2942**

The observed frequencies for Coke and Pepsi are 45 and 40. Therefore the total number of observations for this sub-table is 85. The null hypothesis is that Coke and Pepsi are preferred by equal numbers of people. Therefore the expected frequency is 42.5 for both cells. Working it out, we find that L^2 (df=1, n = 85) = 0.2942, p = 0.588. (see Table 2.9). Therefore, we cannot reject the null hypothesis for this test.

Table 2.10 Calculation of L^2 for (Coke+Pepsi) vs Brand X comparison

Observed (O)	*Expected (E)*	*ln(O/E)*	*O ln(O/E)*		
85	100	-0.1625	-13.8141		
65	50	0.2624	17.0537		
150	150		3.2396	L^2 =	**6.4792**

For our second comparison, the observed frequencies are 85 (for Coke & Pepsi combined) and 65 (for Brand X). The original null hypothesis asserted that the proportions in the 3 categories were equal. Therefore, if we combine 2 of those categories, the expected proportion for the combined category is 2/3, and for the single category it is 1/3. Thus, according to the null hypothesis, the expected frequencies are 100 and 50. Working it through, we get L^2 (df=1, n =

150) = 6.4792, $p = 0.011$. This value of L^2 would allow us to reject the null, and we could conclude that more than one third of the population prefers Brand X.

The important point to be made here is that the sum of the L^2 values for these two orthogonal (or independent) comparisons is equal to the L^2 for the overall data:

0.2942 + 6.4792 = 6.7734

The degrees of freedom are also additive. For the overall data, $df = 2$; and for each of the analytical comparisons, $df = 1$.

2.8 Partitioning the overall *chi-square* for a contingency table

It is also possible to partition the overall *chi-square* for a contingency table, provided that $df > 1$. To see how, let us return to the "screwdriver" problem (Table 2.5). You may recall that there were 3 groups of subjects, and that group 1 was given "no special instructions", whereas groups 2 and 3 were given instructions. Therefore, it may be sensible to begin by comparing group 1 to groups 2 and 3 combined. This will allow us to assess the effect of no instructions versus instructions. The observed frequencies for this comparison are shown in Table 2.11. The expected frequencies can be calculated in the usual fashion.

For this comparison, L^2 (df=1, n = 90) = 10.0208, p = .002 (see Table 2.12). Therefore we would reject the null hypothesis, and conclude that there is a difference between *having* and *not having* instructions.

Table 2.11 Observed frequencies for comparison of a_1 to (a_2 + a_3)

	a_1	a_{2+3}	Total
Solve	9	39	48
Fail to solve	21	21	42
Total	30	60	90

Table 2.12 L^2 calculations for comparison of a_1 to (a_2 + a_3)

Observed (O)	*Expected (E)*	*ln(O/E)*	*O ln(O/E)*		
9	16	-0.5754	-5.1783		
39	32	0.1978	7.7152		
21	14	0.4055	8.5148		
21	28	-0.2877	-6.0413		
90	90		5.0104	L^2 =	**10.0208**

Having discovered that there is a difference between *having* and *not having* instructions, we might wish to go on and compare the 2 kinds of instructions that were given. The null hypothesis for this comparison is that the 2 instructional conditions do not differ. The observed frequencies for this comparison are shown in Table 2.13, and the calculations are summarised in Table 2.14. (Note that the expected frequencies are calculated in the usual fashion, but on the basis of this sub-table alone.) For this comparison, L^2 (1, n = 60) = 1.8448, p = 0.174. (see Table 2.14). Therefore we cannot reject the null hypothesis, and must conclude that there is no difference between the two instructional conditions.

Table 2.13 Observed frequencies for comparison of a_2 and a_3

	a₂	***a₃***	**Total**
Solve	17	22	39
Fail to solve	13	8	21
Total	30	30	60

Table 2.14 L^2 calculations for comparison of a_2 and a_3

Observed (O)	***Expected (E)***	***ln(O/E)***	***O ln(O/E)***		
17	19.5	-0.1372	-2.3324		
22	19.5	0.1206	2.6538		
13	10.5	0.2136	2.7765		
8	10.5	-0.2719	-2.1755		
60	60		0.9224	L^2 =	**1.8448**

Finally, it should be noted that the sum of the L^2 values for these two comparisons (each with *df* = 1) is equal to the L^2 value we calculated for the overall data (in Table 2.8). The sum of the values for our two comparisons is 10.0208 + 1.8448 = 11.8656; and the original L^2 value was 11.8656.

2.9 Assumptions/restrictions for use of *chi-square* based tests

The value of Pearson's X^2 will have a χ^2 distribution if the null hypothesis is true, and if the following conditions are met:

1. Each observation is independent of all the others (i.e., one observation per subject);

2. "No more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater" (Yates, Moore & McCabe, 1999, p. 734);

3. For 2x2 tables:

   a) All expected frequencies should be 10 or greater.

b) If any expected frequencies are less than 10, but greater than or equal to 5, some authors suggest that *Yates' Correction* for continuity should be applied. This is done by subtracting .5 from the absolute value of *(O-E*) before squaring (see equation). However, the use of Yates' correction is controversial, and is not recommended by all authors.
c) If any expected frequencies are smaller than 5, then some other test should be used (e.g., Fisher exact Test for 2x2 contingency tables).

$$\chi^2_{Yates} = \sum \frac{\left(|O - E| - 0.5\right)^2}{E} \tag{2.5}$$

2.10 Advantages of Likelihood Chi-Square Tests

It was suggested earlier that there are certain advantages to using L^2 rather than Pearson's X^2. First, according to Hays (1963), there is reason to believe that likelihood ratio tests are less affected by small sample sizes (and small expected frequencies) than are standard Pearson *chi-squares*, particularly when *df* > 1. In these circumstances, Hays suggests that the likelihood ratio test is superior. However, Alan Agresti (1990)—who is more of an authority on this topic, in my opinion—makes exactly the opposite claim:

> It is not simple to describe the sample size needed for the chi-squared distribution to approximate well the exact distributions of X^2 and G^2 [i.e., L^2]. For a fixed number of cells, X^2 usually converges more quickly than G^2. The chi-squared approximation is usually poor for G^2 when *n/IJ* < 5 [where n = the grand total and *IJ* = *rc* = the number of cells in the table]. When *I* or *J* [i.e., *r* or *c*] is large, it can be decent for X^2 for *n/IJ* as small as 1, if the table does not contain both very small and moderately large expected frequencies. (Agresti, 1990, p. 49)

Another advantage concerns partitioning of an overall contingency table into orthogonal components. Had we done this using Pearson's X^2 method, we would have needed to use two different expected frequencies for each cell. The expected frequency in the numerator (E_n) is based on the current sub-table; and the expected frequency in the denominator (E_d) is based on the original, overall table of frequencies. (For some comparisons, $E_n = E_d$, but this is not always the case.) Things are much simpler with L^2, however, because there is only one expected frequency for each cell, and it is always based on the current sub-table you are working with.

Also, when you partition a contingency table into as many orthogonal components as possible, with X^2, the orthogonal components usually do not add up exactly to the overall X^2. With L^2, on the other hand, rounding error aside, things *always* add up as they should. (I don't know about you, but this property of L^2 tests gives me a warm feeling inside. Things that are supposed to add up to a certain total do—and that gives me confidence in the test.)

Finally, L^2 tests of this sort can be viewed as the simplest form of *loglinear analysis*, which is increasing in popularity. We will not go into it here, but some of you may come across it in the future.

2.11 Fisher's exact test

As noted earlier, it may be more appropriate to use Fisher's exact test to analyze the data in a 2x2 contingency table if any of the expected frequencies are less than 5. Consider the following (slightly modified) example, taken from the BMJ's Statistics at Square One chapter on the Exact Probability Test.

> Some soldiers are being trained as parachutists. One rather windy afternoon 55 practice jumps take place at two localities, dropping zone A and dropping zone B. Of 40 men who jump at dropping zone A, two suffer sprained ankles, and of 15 who jump at dropping zone B, five suffer this injury. The casualty rate at dropping zone B seems unduly high, so the medical officer in charge decides to investigate the disparity. Is it a difference that might be expected by chance? If not it deserves deeper study. (From http://www.bmj.com/collections/statsbk/9.shtml , downloaded on 2-Mar-2001.)

The data are summarized in Table 2.15. Note that the smallest expected frequency is 7(15)/55 = 1.91, which is considerably less than 5. Therefore the sampling distribution of Pearson's X^2 will not be all that well approximated by a χ^2 distribution with df=1.

Table 2.15: Numbers of injured and uninjured men at two different drop zones.

	Injured	**Uninjured**	**Total**
Drop zone A	2	38	40
Drop zone B	5	10	15
Total	7	48	55

The medical officer's null hypothesis is that there is no (population) difference in the injury rate at the two drop zones, or that the difference between 2/40 (5.0%) and 5/15 (33.3%) is due to chance. So what we really want to know then, is how likely is it that we would see a discrepancy in injury rates **this large or larger** if there is really no difference in the population from which we've sampled?

Table 2.16: More extreme differences in injury rates at the two drop zones.

(a)

	Injured	Uninjured	Total
Zone A	1	39	40
Zone B	6	9	15
Total	7	48	55

(b)

	Injured	Uninjured	Total
Zone A	0	40	40
Zone B	7	8	15
Total	7	48	55

The first thing to note here is that we are concerned not only with the numbers shown in Table 2.15, but also with any tables that exhibit a larger discrepancy in injury rates[3] while **maintaining the same marginal (row and column) totals**.[4] The discrepancy would be larger if the number injured at Drop Zone A dropped to 1 or 0. These scenarios are shown in Table 2.16. Remember that the marginal totals must remain fixed.

In general terms, what we need to do next is calculate the probability of getting 2 or 1 or 0 as the *Zone A—Injured* cell count, given that the null hypothesis is true. Note that these outcomes are all **mutually exclusive**. In other words, we cannot observe more than one of these outcomes at the same time. If the *Zone A—Injured* cell count = 2, it cannot equal 1 or 0, and so on. This is to our advantage, it allows us to take advantage of the special addition rule for mutually exclusive events. Very briefly, if A, B, and C are mutually exclusive events, then the probability of A or B or C is the sum of their probabilities. In symbols: p(A or B or C) = p(A) + p(B) + p(C).[5] For the problem we are considering, we may define A, B, and C as follows: A = (cell count = 2), B = (cell count = 1), and C = (cell count = 0).

And now the only remaining challenge is how to calculate p(A), p(B), and p(C). Because the marginal totals are fixed, these probabilities can be calcuated using the *hypergeometric* probability function. Let us start with the observed outcome shown in Table 2.15. When the marginal (row and column) totals are fixed, the probability of this outcome can be calculated as follows:

$$p(\text{Table 2.15 outcome}) = \frac{\left(C_2^{40}\right)\left(C_5^{15}\right)}{C_7^{55}} \tag{2.6}$$

where:

$$\begin{aligned} C_2^{40} &= \text{ the \# of ways of choosing 2 from 40 at drop Zone A} \\ &= \frac{40!}{2!38!} = \frac{40(39)(38)...(1)}{(2)(1)(38)(37)...(1)} = \frac{40(39)}{2(1)} = 780 \end{aligned} \tag{2.7}$$

$$\begin{aligned} C_5^{15} &= \text{ the \# of ways of choosing 5 from 15 at drop Zone B} \\ &= \frac{15!}{5!10!} = \frac{15(14)(13)...(1)}{(5)(4)...(1)(10)(9)...(1)} = \frac{15(14)...(11)}{5(4)(3)(2)(1)} = 3003 \end{aligned} \tag{2.8}$$

---

[3] I will limit discussion to a larger discrepancy in the observed direction only. In other words, I will discuss the one-tailed version of the test here. Two-tailed tests are also possible, as you may notice in the output from the CROSSTABS procedure in SPSS.

[4] Fisher's exact probability test is a "conditional" test, because the p-value that is calculated is **conditional** on the row and column totals.

[5] If you need to review the special addition rule or other aspects of basic probability, see Chapter 5 in Norman & Streiner (2000), or my chapter of notes on *Probability & Hypothesis Testing*.

$$C_7^{55} = \text{ the \# of ways of choosing 7 from 55 overall}$$
$$= \frac{55!}{7!48!} = \frac{55(54)(53)...(1)}{(7)(6)...(1)(48)(47)...(1)} = \frac{55(54)...(49)}{7(6)(5)...(1)} = 202,920,000 \quad (2.9)$$

Putting the pieces together, we get:

$$p(\text{Table 2.15 outcome}) = \frac{780(3003)}{202,920,000} = .011543 \quad (2.10)$$

Going through the same calculations for the two more extreme outcomes shown in Table 2.16, I obtained the following:

$$p(\text{Table 2.16a outcome}) = \frac{C_1^{40}C_6^{15}}{C_7^{55}} \frac{40(5005)}{202,920,000} = .000987 \quad (2.11)$$

$$p(\text{Table 2.16b outcome}) = \frac{C_0^{40}C_7^{15}}{C_7^{55}} \frac{1(6435)}{202,920,000} = .000032 \quad (2.12)$$

Finally, the exact (conditional) probability of the obtained discrepancy in injury rates or a more extreme discrepancy (in the same direction) is given by the sum of these probabilities:

$$p = 0.011543 + 0.000987 + 0.000032 = 0.012561 \quad (2.13)$$

Assuming that the medical officer is using the conventional alpha of 0.05, he or she would reject the null hypothesis, and conclude that the difference in injury rates for the two drop zones is not due to chance.

I should point out that the test as described above is a one-tailed test, because we did not consider outcomes at the other extreme. To make it two-tailed, we would have to check for possible arrangements of the 2x2 table (with the marginal totals still fixed) that showed a larger proportion of injuries at Site A. Any such arrangements that had a probability less than the probability for the observed arrangement would have to be included in the calculation of the overall p-value.

It is not necessary for us to go into the details of all those calculations. However, it is important to note that **the p-value for a two-tailed Fisher exact test is not obtained by doubling the p-value from a one-tailed test**. Why not? Because that approach works only if the sampling distribution of the statistic is symmetrical, and for Fisher's exact test, it is not symmetrical.

Furthermore, if there are no arrangements of the 2x2 table at the other extreme of the distribution that have a probability less than the probability of the observed arrangement of the table, the one-tailed and two-tailed p-values will be identical. This is the case for the example we have been using. Both p-values are equal to 0.013, as shown in the following SPSS output.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7.884[b]	1	.005		
Continuity Correction[a]	5.540	1	.019		
Likelihood Ratio	6.952	1	.008		
Fisher's Exact Test				**.013**	**.013**

a. Computed only for a 2x2 table

b. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 1.91.

2.12 McNemar's chi-square (or McNemar change test)

As discussed earlier, the χ^2 test of association and Fisher's exact test both require independence of observations. This assumption is most often met by having only one observation per subject (i.e., each subject falls into only one of the cells of the contingency table).

But there are situations where we may have matched pairs of subjects (e.g., the Tryptophan-EMS example in Norman & Streiner, pp. 209-210), or when we categorize subjects both before and after some intervention. In these cases, we the observations for those matched pairs, or for the same person on two occasions are not independent. And so, it would be inappropriate to use the chi-squared test of association.

Fortunately, a chap named McNemar worked out a way to handle this situation. Let's see how by looking at an example. Suppose you are the campaign manager for a local political candidate. You survey members of the audience before a live television debate (on Cable 14), and find that 24 of the 60 in attendance favour your candidate. You poll the same 60 people again immediately after the debate, and the outcome is as shown in Table 2.17.

Table 2.17: Summary of opinions before and after the debate.

		After Debate		
		For	Against	Total
Before Debate	For	a = 20	b = 4	24
	Against	c = 13	d = 23	36
	Total	33	27	60

As it turns out, only those folks who changed their minds are of interest when it comes to computing McNemar's chi-square. (This is why it is sometimes called McNemar's **change test**.) In this dataset, the 17 subjects falling in the shaded cells (b and c) changed their minds. The null hypothesis we wish to test is that in the population from which we've sampled, changes in one direction are equally likely to changes in the other direction. If that is so, then the expected frequency for both cells b and c will be equal to 17/2=8.5. In general, the expected frequency is equal to the sum of the off-diagonal (discordant) frequencies divided by 2. With these expected

and observed frequencies in hand, we can proceed with the usual formula for calculating X^2. But let's add a subscript 'M' to remind ourselves that this is McNemar's X^2, not Pearson's.

$$X_M^2 = \frac{\sum (O-E)^2}{E} = \frac{(4-8.5)^2}{8.5} + \frac{(13-8.5)^2}{8.5} = 4.765 \tag{2.14}$$

I presented equation (2.14) to help you see that McNemar's X^2 is calculated using the same basic approach as Pearson's X^2. However, the more common formula for McNemar's test statistic is as follows:

$$X_M^2 = \frac{(b-c)^2}{b+c} = \frac{(4-13)^2}{4+13} = 4.765 \tag{2.15}$$

Equation (2.15) is just an algebraic simplification of equation (2.14), and as you can see, both yield the same result.

Because there are only two cells, it should be clear that df=1 for McNemar's chi-square. So, X_M^2 (*df*=1, *n*=17) = 4.765, p = 0.029, we can reject the null hypothesis.

However, note that it is possible to apply Yates' correction when calculating McNemar's X^2, just as one may do for Pearson's X^2. Applying Yates' correction to equation (2.14) yields this:

$$X_M^2 = \frac{\sum (|O-E|-.5)^2}{E} = \frac{(|4-8.5|-.5)^2}{8.5} + \frac{(|13-8.5|-.5)^2}{8.5} = 3.765 \tag{2.16}$$

which simplifies to the following:

$$X_M^2 = \frac{(|b-c|-1)^2}{b+c} = \frac{(|4-13|-1)^2}{4+13} = 3.765 \tag{2.17}$$

With Yates' correction, X_M^2 (*df*=1, *n*=17) = 3.765, p = 0.052. So in this particular example, X_M^2 tests with and without Yates' correction yield p-values on opposite sides of the conventional .05 alpha level.

McNemar test of marginal homogeneity. Notice that if b = c (i.e., the number of changes in one direction equals the number of changes in the other direction), then (a + b) = (a + c), and (c + d) = (b + d). Putting it another way, if b = c, the row 1 total equals the column 1 total, and the row 2 total equals the column 2 total. When that condition holds, the table is said to exhibit *marginal homogeneity*. And so, testing the null hypothesis that b = c is *equivalent* to testing the null hypothesis that marginal homogeneity exists in the population from which the sample was drawn.

2.13 Measures of association for categorical data

To this point, I have discussed methods that can help one decide whether or not there is a relationship between 2 categorical variables. If there is a relationship, it may be desirable to then assess the **strength** of the relationship.

For 2x2 contingency tables that could be analyzed with the chi-square test of association (i.e. with independent observations), one such measure of association is the phi coefficient. There are various ways to calculate phi, and many of them are computational shortcuts that were useful when folks did most of their computations by hand. But conceptually, the phi is really just **Pearson r calculated on 2 dichotomous variables**.

Table 2.18: Association between predicted and actual criminal status.

PRED * STATUS Crosstabulation

Count

		Actual Status		
		Youthful Legally Challenged	Other	Total
Predicted Status	bars = yes	36	24	60
	bars = no	40	100	140
Total		76	124	200

This can be illustrated with the data in Table 2.18, which is a reproduction of Table 21-1 from Norman & Streiner (2000). Table 2.19 shows chi-square test of association results for these data. It is clear that there is a relationship between predicted and actual criminal status. The phi coefficient shown in Table 2.20 provides a measure of the strength of that association.

I said a moment ago that phi is simply a Pearson r calculated on 2 dichotomous variables. This fact is illustrated in Table2.21. The Pearson r shown here was calculated with categories scored as follows:

- (Bars=Yes) = 0
- (Bars=No) = 1
- YLC = 0
- Other = 1

Table 2.19: Chi-square tests for data in Table 2.18

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	**17.609**[b]	1	.000		
Continuity Correction[a]	16.300	1	.000		
Likelihood Ratio	17.349	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	17.520	1	.000		
N of Valid Cases	200				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 22.80.

Table 2.20: Phi coefficient (and Cramer's V) from SPSS CROSSTABS.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.297	.000
	Cramer's V	.297	.000
N of Valid Cases		200	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Table 2.21: Pearson r calculated on Table 2.18 data.

Correlations

		Predicted Status	Actual Status
Predicted Status	Pearson Correlation	1.000	.297
	Sig. (2-tailed)	.	.000
	N	200	200
Actual Status	Pearson Correlation	.297	1.000
	Sig. (2-tailed)	.000	.
	N	200	200

Some textbooks include this formula for phi:

$$\phi = r_\phi = \sqrt{\frac{\chi^2}{N}} \tag{2.18}$$

The notation r_ϕ is used by some authors (e.g., Glass & Hopkins) to remind readers that the phi coefficient is nothing more or less than Pearson r calculated on 2 dichotomous variables. Because phi is equal to Pearson r, it can range in value from -1 to +1. But as you know, *the value of* χ^2 *cannot be negative*. The equation shown above cannot generate a negative number. So that equation really gives the *absolute value* of phi. If you plug in the values of N and Pearson chi-square from Tables 2.18 and 2.19 respectively, you should once again arrive at phi = 0.297. Bearing in mind that this is really just a Pearson r, the strength of the relationship is not all that impressive.

Contingency tables are sometimes used to display information about how well 2 (or more) raters agree. See Table 21-2 in Norman & Streiner (2000), for example. In this case the most commonly used measure of association—or measure of agreement—is Cohen's kappa (κ). I refer you to Norman & Streiner's excellent discussion of kappa.

Finally, as Norman & Streiner point out, there are a host of other measures of association for 2x2 tables, but most of these are variations on the same theme, and many of them are rarely seen these days (except by people like yourselves, who take great pleasure in pouring over statistics textbooks.)

2.14 How to perform these tests using SPSS

The following syntax files show how to perform the analyses described in this chapter using SPSS:

- http://www.angelfire.com/wv/bwhomedir/spss/goodfit.SPS
- http://www.angelfire.com/wv/bwhomedir/spss/association.SPS
- http://www.angelfire.com/wv/bwhomedir/spss/fisher.SPS
- http://www.angelfire.com/wv/bwhomedir/spss/mcnemar.SPS
- http://www.angelfire.com/wv/bwhomedir/spss/phikappa.SPS

References

Agresti, A. (1990). *Categorical data analysis*. New York, NY: John Wiley & Sons.

Hays, W.L. (1963). *Statistics*. New York, NY: Holt, Reinhart & Winston.

Howell, D.C. (1997). *Statistical methods for Psychology* (4th edition). Boston, MA: Duxbury.

Norman, G.R., & Streiner, D.L. (2000). Biostatistics: The bare essentials (2nd ed). Hamilton, ON: B.C. Decker Inc.

Siegel, S., & Castellan, J.J. (1988). *Nonparametric statistics for the behavioral sciences*. New York, NY: McGraw-Hill.

Yates, D., Moore, Moore, D., McCabe, G. (1999). *The Practice of Statistics* (1st Ed.). New York: W.H. Freeman.

Appendix: Natural Logarithms

It may be that some of you are uncomfortable with this talk of logarithms. Let us try to clarify things by starting with some fairly simple examples. I hope you will all agree with the following:

$10^1 = 10$
$10^2 = 10 \times 10 = 100$
$10^3 = 10 \times 10 \times 10 = 1{,}000$
$10^4 = 10 \times 10 \times 10 \times 10 = 10{,}000$

It is also true that:

$\log_{10}(10) = 1$
$\log_{10}(100) = 2$
$\log_{10}(1{,}000) = 3$
$\log_{10}(10{,}000) = 4$

The subscript "10" in these equations is the **base** of the logarithm. It is possible to have any number (other than zero) as the base, but the most common bases are 10 and the constant *e*. (More on *e* in a moment.) If you see "log" without a subscript, it is usually safe to assume that the base is 10. It would not be unusual therefore, to see these equations written as:

$\log(10) = 1$
$\log(100) = 2$
$\log(1{,}000) = 3$
$\log(10{,}000) = 4$

The second of these equations (in bold type) asserts that the logarithm (with base = 10) of 100 is equal to 2. In other words, a logarithm is the **power** or **exponent** to which you must raise the base in order to produce the number in parentheses. Another way to think of this is that you need to solve for *x* in the following equation:

$$10^x = 100$$

Clearly, $10^2 = 100$, and so $x = 2$. The base 10 must be raised to the power of 2 to produce a result of 100.

As mentioned earlier, the *other* most common base for logarithms is the constant *e*. The value of *e* (to 7 decimal places) is 2.7182818. (You can find this number on your calculator by pressing 1, and then the e^x key.) For our purposes, it is not necessary to understand where this number comes from. (In other words, I can't remember enough high school physics to explain it!) However, it is important to know that logarithms with a base of *e* are called **natural logarithms**. Note as well that it is common to use *ln* rather than *log* for natural logarithms:

log_e = 'natural' logarithm = *ln*

Finally, for illustrative purposes, here are the base-10 and natural logs (to 4 decimal places) for the same set of numbers:

log(10) = 1	*ln*(10) = 2.3026	$e^{2.3026} = 10.0001$
log(100) = 2	*ln*(100) = 4.6052	$e^{4.6052} = 100.0030$
log(1,000) = 3	*ln*(1,000) = 6.9078	$e^{6.9078} = 1{,}000.0447$
log(10,000) = 4	*ln*(10,000) = 9.2103	$e^{9.2103} = 9{,}999.5962$

You may wish to check these calculations using the *ln* key on your calculator. (Most spreadsheet programs also have a built in *ln* function.) The column on the right shows the constant *e* raised to the power of the natural logarithm. As you can see, apart from some rounding error, the result is the original number whose natural log was calculated.

For anyone who is still unclear about logarithms, the following online tutorial may be helpful:

http://www.phon.ucl.ac.uk/cgi-bin/wtutor?tutorial=t-log.htm

Review Questions

1. What are the assumptions underlying the use of χ^2 tests?

2. What are the advantages of L^2 tests over Pearson's χ^2?

3. What is the formula for computing expected frequencies for a χ^2 test (or L^2 test) of independence? Explain the rationale underlying this formula.

4. What is the null hypothesis for a test χ^2 test (or L^2 test) of independence?

5. What is the null hypothesis for a test χ^2 (or L^2) goodness of fit test?

6. How many degrees of freedom are there for a χ^2 (or L^2) goodness of fit test?

7. What is the critical value of *z* (in the table of the standard normal distribution) with $\alpha = .05$ (2-tailed)? What is the critical value of χ^2 with $\alpha = .05$ and $df = 1$? What do you conclude about the relationship between a χ^2 distribution with $df = 1$ and the standard normal distribution?
