Chapter 7
Statistical Approaches for Nominal Data: Chi-Square Tests

In the previous chapter, we considered statistical tests that involve (1) a single continuous numeric variable, the sample mean for which is tested against a hypothetical population value, and (2) a single continuous numeric variable in which we compare means across a two category nominal level grouping variable like sex. Sometimes, however, we are faced with two nominal level variables we think may be related, like marital status and region of the country. In other cases, we may have a pair of ordinal variables, or a combination of nominal and ordinal variables, we expect to be associated but perhaps not linearly. For example, we may be interested in the relationship between religious affiliation (e.g., Protestant, Catholic, Jewish, Other, None) and educational degree attainment (e.g., high school diploma, 2-year college degree, 4-year college degree, more than a 4-year degree).

In these cases, the mean of one or both of the variables has no real meaning, and so tests that compare means are inappropriate. Additionally, even if comparing means *might* make some sense (e.g., examining *mean* educational attainment by religious affiliation), we may not expect a clear pattern that will be detectable as a difference in means. Table 7.1 presents some hypothetical religious affiliation and educational attainment data. In these data, there are 100 Catholics and 200 Protestants. All of the Catholics have a high school diploma; there is no variability in educational attainment. None of the Protestants has a high school diploma; instead, exactly half have less schooling, while half have more.

If we were to treat educational attainment as numeric, giving the categories values of 1, 2, and 3, respectively, mean attainment would be identical for the two religion groups. As a result, the numerator of an independent samples t test would be 0, and we would reject the null hypothesis that the means do not differ between the two groups. However, this finding is clearly not the end of the story, because it ignores a clear pattern in the data.

In cases like the one presented in this example, as well as the case in which both variables of interest are measured at the nominal level, we can turn to a "nonparametric" test, the most basic of which is the chi-square test of independence. The test is called "nonparametric" for two reasons. First, unlike the z and t tests discussed in the previous chapter, the chi-square test of independence does

S.M. Lynch, *Using Statistics in Social Research: A Concise Approach*,
DOI 10.1007/978-1-4614-8573-5_7,

	Religion		
Education	Catholic	Protestant	TOTAL
Less than High School	0	100	100
High School Diploma	100	0	100
More than HS Diploma	0	100	100
TOTAL	100	200	300

Table 7.1. Crosstabulation of hypothetical religious affiliation and educational attainment data (all religions not shown have been excluded).

	Marital Status				
Region	Married	Widowed	Divorced/ Separated	Never Married	TOTAL
Northeast	91	20	29	61	201
Midwest	174	30	71	79	354
South	246	30	115	120	511
West	174	30	60	90	354
TOTAL	685	110	275	350	1420

Table 7.2. Crosstabulation of marital status and region of the country (2006 GSS data).

not depend on parameters like the mean and variance. Second, the test does not represent, describe, or reveal patterns observed in data using other types of parameters like regression analysis does, as will be discussed in later chapters.

The starting point for the chi-square test is a crosstab of two variables. Thus, you may consider this test when you have data that you would summarize using a crosstab. Table 7.2 is a crosstab of region of the country by marital status in the 2006 GSS. As the table shows, there are 91 persons (out of the 1,420 in the 2006 sample) who are both married and live in the northeast, 20 who are widowed and live in the northeast, 29 who are divorced or separated and live in the northeast, and so on. In total, there are 201 persons in the sample who live in the northeast, 354 who live in the midwest, 511 who live in the south, and 354 who live in the west. Similarly, there are 685 who are married, 110 who are widowed, 275 who are divorced or separated, and 350 who have never been married.

Notice that both of these variables are measured at the nominal level; that is, there is no way to order the categories in any meaningful way. Also notice that the raw counts themselves vary substantially from cell to cell—ranging from 20 to 246—and the marginals also vary substantially. There are far more married persons (685) than widowed persons (110), and the region distribution shows that far more live in the south (511) than in the northeast (201).

All in all, given that the raw counts vary considerably from cell to cell and across the marginals, and given that the categories of each variable cannot be meaningfully ordered, any relationship that may exist between marital status and region is difficult to discern immediately from the raw cell counts. For this reason, as we did in Chap. 4, we generally construct and include row and/or column percentages in the crosstab, depending on which variable we consider to be the independent (predictor)

	Marital Status				
Region	Married	Widowed	Divorced/ Separated	Never Married	TOTAL
Northeast	91	20	29	61	201
	45.3 %	10.0 %	14.4 %	30.3 %	100 %
Midwest	174	30	71	79	354
	49.2 %	8.5 %	20.1 %	22.3 %	100 %
South	246	30	115	120	511
	48.1 %	5.9 %	22.5 %	23.5 %	100 %
West	174	30	60	90	354
	49.2 %	8.5 %	16.9 %	25.4 %	100 %
TOTAL	685	110	275	350	1420
	48.2 %	7.7 %	19.4 %	24.6 %	100 %

Table 7.3. Crosstabulation of marital status and region of the country (2006 GSS data) with row percentages included.

vs. dependent (outcome) variable. Here, our theory might suggest that regional culture may influence marital patterns. For example, the so-called "Bible Belt" runs through the south, and so it may be more likely for individuals to become and remain married than to be separated or divorced, or at least this pattern may be stronger in the south than in other regions.

To assist us in providing a descriptive depiction of this expected pattern, given the structure of our marginals (marital categories along the horizontal and regions along the vertical), we may incorporate row percentages in the table. Table 7.3 expands Table 7.2 to include these percentages. The table shows that our initial hypothesis may not be supported. Of those who live in the south, 48.1 % are married, while the percentages who are married of those who live in other regions are generally higher. For example, 49.2 % of those who live in either the midwest or the west are married. Only those who live in the northeast are less likely to be married (45.3 %).

Of course, one could argue that the age distribution of each region varies. So, the facts (1) that marital status transitions (e.g., going from being never married, to married, to divorced or separated or widowed) are age-patterned, and (2) that southerners may be either younger than persons in other regions (due to higher fertility levels) or older (due to migration to Florida post-retirement), may explain the lower percentage married. In that case, we might expect that southerners have a higher percentage who are either never married (the younger folks) or widowed (the older retirees). However, the comparisons of the percentages of persons in other marital status categories fail to support this view.

In terms of divorce or separation, the south has the highest percentage: 22.5 % of those who live in the south fall in this marital status category, while the proportions in this category range from 14.4 to 20.1 % in the other three regions. Furthermore, the south has the smallest percentage widowed and the second smallest percentage never married. Thus, the results suggest the age distribution argument is not a reasonable explanation. Instead, the results suggest that southerners divorce or separate at a high rate once they marry.

7.1 The Chi-Square (χ^2) Test of Independence

The region-by-marital status crosstab is informative, but there are limits to what can be determined from it. Most importantly, the data are sample data, and so we can't immediately conclude that patterns observed in the data exactly match those in the population. That is, we cannot immediately know whether the observed sample patterns are due to random fluctuation from sample to sample that could be drawn from a population in which there is no pattern, or whether they are real patterns. Less importantly, but also important, row and/or column percentages do not necessarily show us where the strongest patterns are in the data, especially once random variation is considered. For example, some 30 % of persons in the Northeast are in the "never married" category, while 22 % of those in the Midwest are in that category. At the same time, some 10 % of those in the Northeast are widowed, while 5.9 % of those in the South are widowed. Both of these differences in percentages seem to be substantively significant contributors to the overall pattern of regional differences in marital status. Yet, which difference is more important to the overall pattern, if one exists after considering sampling variability? The chi square test of independence can help us answer these questions.

The chi square test of independence begins with the assumption that the two variables of interest are independent in the population and involves computing the probability of observing the sample data under this assumption. As with the hypothesis testing approach outlined in the last chapter, if the probability of observing the sample data is small under the assumption of independence, we reject this "null" hypothesis and conclude that the variables are probably not independent in the population.

How do we compute the probability that the variables of interest are independent in the population? Recall from Chap. 5 that, if two events are independent, then their joint probability ($p(A, B)$) is the product of their respective marginal probabilities. In the context of the current example, one set of events is the region in which the respondent lives, and the other set is the marital status the respondent holds. Thus, "living in the Northeast" is an event, as is "being married," and so being married and living in the Northeast is a joint event. If these events are independent, then the probability that a randomly selected individual is both living in the Northeast and is married is simply the product of the probability that s/he lives in the Northeast with the probability that s/he is married.

The probability of living in the Northeast, given the data, can be computed from the row marginal: 201 of the 1,420 respondents live in the Northeast; thus, the probability of living in the Northeast is $201/1{,}420 = .142$. The probability of being married, given the data, can be computed from the column marginal: 685 of the 1,420 respondents are married; thus, the probability of being married is $685/1{,}420 = .482$. Therefore, the joint probability of living in the Northeast and being married, if region and marital status are independent, is $.142 \times .482 = .0684$.

Of a sample of $n = 1{,}420$ persons, this result means the number of persons that we *should* see in this category, under the assumption of independence, would be

	$X = 0$	$X = 1$	Total (Marginals for Y)
$Y = 0$	a	b	$a + b$
$Y = 1$	c	d	$c + d$
Total (Marginals for X)	$a + c$	$b + d$	$a + b + c + d = n$

Table 7.4. Generic cross-tab of variables X and Y

$E = 1{,}420 \times .0684 = 97.1$. However, the data show that 91 persons hold this joint status. The difference between the observed ("O") count and the expected count ("E") is a measure of the degree of discrepancy between reality and what reality would be expected to look like if region and marital status were independent.

Each cell in the crosstab provides us with an observed count. We can compute an expected count for each cell as we did for the first cell, and we should find that our expected counts sum to the overall sample size. Thus, a simple cell-by-cell sum of the difference between the observed and expected counts will be 0.

Computing the chi-square test statistic involves finding the discrepancies between observed and expected counts by cell, squaring them, dividing them by the expected count, and summing these transformed discrepancies across all cells. Mathematically, the chi-square statistic is computed as:

$$\chi^2 = \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{(O_{rc} - E_{rc})^2}{E_{rc}}, \tag{7.1}$$

where O_{rc} is the observed cell count in row r and column c, E_{rc} is the expected cell count under the assumption of independence, and the summation is across all cells in the R by C table.

The generic computation of the expected cell counts can be simplified from what I presented above as:

$$E_{rc} = \frac{(O_r)(O_c)}{n}. \tag{7.2}$$

Why? Consider the generic crosstab in Table 7.4. If X and Y are independent, then $p(x, y) = p(x) \times p(y)$. So, the probability of falling in the $(0, 0)$ cell is:

$$P(x = 0, y = 0) = p(x = 0) \times p(y = 0) = \frac{a + c}{n} \times \frac{a + b}{n}. \tag{7.3}$$

Thus, $p(x = 0, y = 0) = \frac{(a+c)(a+b)}{n^2}$. The expected cell *count* , then is:

$$n \times p(x = 0, y = 0) = n \times \frac{(a + c)(a + b)}{n^2} = \frac{(O_{x=0})(O_{y=0})}{n}. \tag{7.4}$$

Once we have computed the expected cell counts under the assumption of independence, we need to determine how far from this assumption the observed

cell counts are. This is captured by the test statistic formula shown above: $\sum_{r=1}^{R}\sum_{c=1}^{C}\frac{(O_{rc}-E_{rc})^2}{E_{rc}}$. This formula measures the sum of the squared differences between the observed and expected counts, after these squared differences have been adjusted for the magnitude of the expected counts. That is, if the expected count of observations is large, then we would probably expect the squared deviation between observed and expected to be large also, even if the variables are, in fact, independent. So, these squared deviations are adjusted somewhat for the magnitude of the expected cell count.

If we were to take repeated samples from a population in which two variables were independent, we would expect the sum of the squared deviations as computed above to vary across samples. Some samples would have greater sums of squared deviations than others. Overall, the sums across repeated samples would follow a chi-square distribution, just as sample means in repeated samples follow a normal distribution. So, we can use the chi-square distribution to assess the probability of observing our data under the independence assumption.

The chi square distribution is a probability distribution like the binomial, normal, and t distributions we have discussed before (see the tables at the end of the book for more discussion). It is a strictly non-negative distribution (sums of squares cannot be negative), and so it is often skewed to the right. As with the t distribution, the chi square distribution has a degree of freedom parameter that changes the shape of the distribution. For the chi square test of independence, the degrees of freedom are computed as: $(R-1)(C-1)$, where R and C are the numbers of rows and columns (respectively) in the crosstab table. This calculation simply represents the number of cells whose counts need to be known (in addition to the marginals) before the entire crosstab is determined. For example, in a two-by-two crosstab, the count in only one cell must be known before the counts in all remaining cells are determined.

7.1.1 The Test Applied to the Region and Marital Status Data

Table 7.5 shows the chi square test of independence calculations applied to the region/marital status data (only first two columns show $O_r O_c/n$ computations). The test statistic for these data was $\chi^2 = 14.65$ on $(4-1)(4-1) = 9$ degrees of freedom. If we look this chi square value up in Appendix A, we find that our χ^2 value is just short of reaching the $p < .1$ level, and it is certainly too small to be statistically significant at the usual $\alpha = .05$ level. Thus, we cannot reject the null hypothesis that region and marital status are independent. Notice that, unlike the z and t tests we conducted in the previous chapter, we only use one-tail p-values in the chi square test. The reason for this is that the distribution is non-negative, and perfectly independent variables should therefore produce a chi square value of 0. There is no left-tail extreme value to consider. Instead, the question is simply whether our observed discrepancies between the observed data and the expected data under the independence assumption are large enough to make us doubt the independence assumption.

	Marital Status				
Region	Married	Widowed	Divorced/ Separated	Never Married	TOTAL
NE	91	20	29	61	201
E:	$\frac{(201)(685)}{1420} = 97.0$	$\frac{(201)(110)}{1420} = 15.6$	38.9	49.5	
$\frac{(O-E)^2}{E}$	.37	1.24	2.52	2.67	6.80
MW	174	30	71	79	354
E:	$\frac{(354)(685)}{1420} = 170.8$	$\frac{(354)(110)}{1420} = 27.4$	68.6	87.3	
$\frac{(O-E)^2}{E}$	.06	.25	.08	.79	1.18
S	246	30	115	120	511
E:	$\frac{(511)(685)}{1420} = 246.5$	$\frac{(511)(110)}{1420} = 39.6$	99.0	126.0	
$\frac{(O-E)^2}{E}$	.001	2.33	2.59	.29	5.20
W	174	30	60	90	354
E:	$\frac{(354)(685)}{1420} = 170.8$	$\frac{(354)(110)}{1420} = 27.4$	68.6	87.3	
$\frac{(O-E)^2}{E}$	.06	.25	1.08	.08	1.47
TOTAL	685	110	275	350	1420

Table 7.5. Crosstabulation of marital status and region of the country (2006 GSS data) with chi square test of independence calculations ($\chi^2 = 14.65$, $df = 9$, $p > .10$)

7.2 The Lack-of-Fit χ^2 Test

The chi square test can be modified for the case in which one has either true or hypothesized population values and observed sample values and wishes to test whether the observed sample came from the known (or hypothetical) population. The computation of the test is the same as with the test of independence, but the data are generally arranged in a list. The degrees of freedom for the lack-of-fit test is $C - 1$, where C is the number of groups in the population.

As a simple example of the lack-of-fit test, suppose I know that the population in 2000 was 80 % white, 14 % black, and 6 % other races. Of the 1,667 persons in the 2000 sample of the GSS, 1,380 were white, 173 were black, and 114 were of other races. Given the known population proportions, in a sample of 1,667 persons, we would expect to have $1{,}667 \times .8 = 1{,}333.6$ whites, $1{,}667 \times .14 = 233.4$ blacks, and $1{,}667 \times .06 = 100.0$ others. Table 7.6 shows the observed, expected, and chi-square computations for these data and population proportions. The lack-of-fit statistic was 19.2 on 2 degrees of freedom. This value is much larger than 13.82, the critical value needed for obtaining $p < .001$. If our population proportions had been hypothesized, we would therefore reject the hypothesized values: the sample would probably not arise from the hypothesized proportions.

In this case, however, the population proportions were known. Given that the observed data would probably not occur in a simple random sample, given the

Race	Sample (Observed)	Expected (Pop. $\%\times n$)	$(O-E)^2/E$
White	1380	1333.6	1.61
Black	173	233.4	15.63
Other	114	100.0	1.96
TOTAL	1667	1667	$\chi^2 = 19.2$

Table 7.6. Observed and expected counts of racial group members (2000 GSS data)

population proportions, we may conclude that the data did not come from a simple random sample. In particular, based on the contribution to the chi square statistic made by the black subpopulation (15.63), it seems clear that blacks were undersampled. As this example demonstrates, the lack-of-fit test can therefore be used, when population proportions are known for a particular variable, to determine whether the sample is a simple random subset of the population.

7.3 Conclusions

In this chapter, we developed two statistical tests that can be applied to nominal and ordinal variables. The chi square test of independence is used to determine whether two variables are associated with one another or are independent. Given that the variables being considered may be nominal, the test is designed to determine whether any sort of patterning is present in the data. Thus, the test is often also used with ordinal data to capture nonlinear patterning.

The lack-of-fit test can be used to determine whether the proportions of sample members in each category of a nominal (or other) variable are consistent with a set of known or hypothesized proportions of population members in those categories. A discrepancy can be used to show either (1) that a set of hypothesized proportions is incorrect, or (2) that a sample is not representative of a population (i.e., it is not a simple random sample).

Although these chi square tests are extremely useful, especially when the level of measurement of variables is nominal, or when one is interested in determining whether nonlinear relationships may exist between variables, chi square tests do suffer from two important limitations. First, chi square test statistics are somewhat unstable when cell counts are small. A common rule of thumb is not to trust tests when any cells are expected to have fewer than five observations. Second, chi square tests are highly sensitive to sample sizes. Consider the calculation of the chi square: in the numerator of the calculation of each cell's contribution to the chi square statistic, the difference between the observed and expected counts is squared, but the denominator uses a single number. Thus, as the sample size increases, the numerator tends to grow at a faster rate than the denominator, making it easier to obtain large chi square statistics in large samples.

7.4 Items for Review

- Independence of events and the implications for crosstab cells
- Expected cell counts
- Chi-square test of independence
- Chi-square lack-of-fit test
- Problems with chi square tests

7.5 Homework

1. Reconstruct Table 7.3 to include column percentages rather than row percentages. Interpret.
2. I rolled a single six-sided die 1,000 times and obtained the following counts for each number 1–6 (respectively): 152, 163, 146, 173, 171, 195. Is the die weighted?
3. Is political party affiliation associated with happiness? The following data are from the 2006 GSS:

	Happiness			
Party	Unhappy	Somewhat Happy	Very Happy	Total
Democratic	69	304	119	492
Independent	67	264	128	459
Republican	41	239	189	469
Total	177	807	436	1420

4. Are health and happiness related? The following data are from the 2006 GSS:

	Happiness			
Health	Unhappy	Somewhat Happy	Very Happy	Total
Poor	23	39	10	72
Fair	59	160	52	271
Good	72	418	185	675
Excellent	23	190	189	402
Total	177	807	436	1420

5. 51 % of the population is female, and 49 % is male. In the GSS data, there are 12,038 males and 14,190 females. Is the sample consistent with these population proportions?
6. Of a sample of 50 men and 40 women, 60 % of the men voted, while 45 % of the women voted. Are sex and voting propensities independent?

7. Is marital status associated with political party affiliation?

	Party			
Marital Status	Democrat	Independent	Republican	Total
Married	192	204	289	685
Widowed	45	23	42	110
Div./Sep.	105	92	78	275
Never Married	150	140	60	350
Total	492	459	469	1420

8. In the previous problem, incorporate row or column percentages as appropriate and interpret.
9. Some argue that certain political positions are aligned within political parties. For example, Republicans tend to be both anti-abortion and pro-death penalty, while Democrats tend to be pro-choice and anti-death penalty. If this alignment is true, we would expect to see an association between how people respond to questions about these positions. Below is the data from the GSS for 2010. Based on the data, what would you conclude about the hypothesis about positions?

	Death Penalty			
Abortion	Pro	Anti	Don't Know	Total
Pro-Choice	358	156	21	535
Anti Abortion	449	203	31	683
Don't Know	17	9	7	33
Total	824	368	59	1251

10. Are births distributed equally across months, or is there some patterning? The following table shows the number of individuals born in each month in the GSS:

Month	Births
January	2651
February	2535
March	2709
April	2474
May	2482
June	2650
July	2694
August	2856
September	2781
October	2662
November	2591
December	2610
Total	31,695