Chapter 12

Inference on Categorical Data

Significance tests of categorical variables involve the comparison of a set of observed frequencies with frequencies specified by a hypothetical distribution. We may ask, for instance:

- Do people have the same probability of dying in the month in which they were born as in any other month?
- Is there a relationship between race/ethnicity and political affiliation?
- Is choice of occupation related to one's sex?

This chapter details how to use SPSS to calculate goodness of fit with equal and unequal probabilities, to perform a chi-square test of independence, and to calculate measures of association between categorical variables, for example, the phi coefficient, coefficient lambda, and coefficient gamma.

12.1 TESTS OF GOODNESS OF FIT

When data consist of one categorical variable, it is often informative to ask whether the proportions of responses in the categories conform to a particular pattern. The procedure for addressing such questions is called a *goodness of fit test*.

Equal Probabilities

In this illustration, we will use the data in the "death.sav" data file to test the hypothesis that there are equal probabilities of death occurring in one's birth month or any other month of the year. The null hypothesis is H_0: $p_1 = p_2 = ... = p_{12} = 1/12$. Each entry in the data file is a number indicating the individual's month of death relative to the month of birth; for example, –6 indicates that the month of death is 6 months prior to the month of birth, 0 indicates that both months are the same, and so on. We will test this with an α level of .01.

By default, SPSS calculates the chi-square statistic to test the hypothesis of equal proportions. After you have opened the data file:

1. Click on **Analyze** from the menu bar.
2. Click on **Nonparametric tests** from the pull-down menu.
3. Click on **Chi-Square** to open the Chi-Square Test dialog box (see Fig. 12.1).
4. Click on the variable name ("month") and the **right arrow button** to move it into the Test Variable List box.
5. Click on **OK**.

The output should appear as shown in Figure 12.2.

Because we hypothesized that the chance of dying in any month is equal in proportion, we see that the expected number of individuals who died during each of the 12 months is 348/12 = 29. The test statistic is 22.07 with 11 degrees of freedom. The *P* value of .024 leads us to accept H_0 at the 1% level and conclude that people have an equally likely chance of dying in the month in which they were born as in any other month.

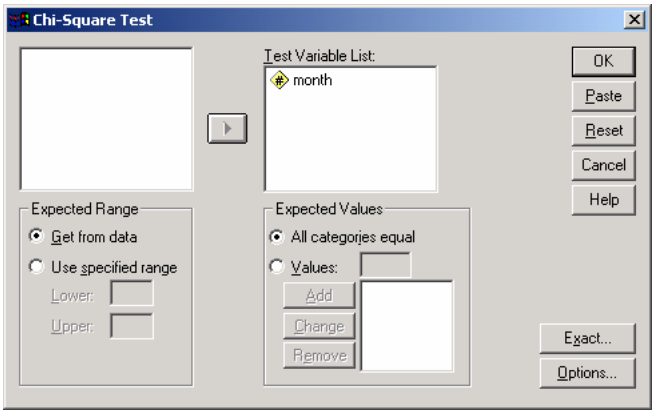

Figure 12.1 Chi-Square Test Dialog Box

MONTH

	Observed N	Expected N	Residual
-6	24	29.0	-5.0
-5	31	29.0	2.0
-4	20	29.0	-9.0
-3	23	29.0	-6.0
-2	34	29.0	5.0
-1	16	29.0	-13.0
0	26	29.0	-3.0
1	36	29.0	7.0
2	37	29.0	8.0
3	41	29.0	12.0
4	26	29.0	-3.0
5	34	29.0	5.0
Total	348		

Test Statistics

	MONTH
Chi-Square[a]	22.069
df	11
Asymp. Sig.	.024

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 29.0.

Figure 12.2 Chi-Square Test of Goodness of Fit

Probabilities Not Equal

It is also possible to conduct goodness-of-fit tests when the proportions in the categories are hypothesized to be unequal. For example, car manufacturers and dealerships keep records on popularity of colors for automobiles. These records indicate that the color distribution is: blue 9.3%, sliver/gray 26.0%, red 13.4%, black 10.2%, green 10.4%, white 14.7%, brown/gold/other 16.0%. Suppose we want to test whether taste in colors on college campuses is the same as the national average. The "cars.sav" datafile contains the results of a survey of college parking lots.

The null hypothesis for this test is H_0: $p_{blue} = .093$, $p_{silver} = .260$, $p_{red} = .134$, $p_{black} = .102$, $p_{green} = .104$, $p_{white} = .147$, $p_{brown/other} = .160$. To test this hypothesis at the .05 level of significance, open the "cars.sav" data file and follow steps 1–4 above, clicking on the variable "color." Then:

1. Click on **Values** in the Expected Values box.
2. Enter the value (proportion) that you hypothesize for the first category of your variable. In this example, blue is coded "1" so enter .093 in the Value box and click on **Add**.
3. Enter the value for the next category of your variable. Silver is coded "2," so enter .260 and click on **Add**.
4. After entering all of the expected values, click on **OK**.

The output should appear as shown in Figure 12.3.

The Expected *N* column indicates the expected number of cars of each color if the hypothesized distribution were true. For example, 9.3% of the sample of 64 cars is 6.0. The actual number of blue cars in the sample was 7 (see the Observed *N* column). The test statistic is a measure of the magnitude of the differences between observed and expected *N*'s. In this example, $\chi^2 = 2.060$, $P = .914$. Thus, using a 5% significance level, we do not reject the null hypothesis of independence and conclude that choice of car colors on college campuses does not depart from the national figures.

car color

	Observed N	Expected N	Residual
blue	7	6.0	1.0
grey	15	16.6	-1.6
red	7	8.6	-1.6
black	9	6.5	2.5
green	8	6.7	1.3
white	8	9.4	-1.4
brown	10	10.2	-.2
Total	64		

Test Statistics

	car color
Chi-Square[a]	2.060
df	6
Asymp. Sig.	.914

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 6.0.

Figure 12.3 Goodness-of-Fit Test for Unequal Proportions

12.2 CHI-SQUARE TESTS OF INDEPENDENCE

When a study concerns the relationship of two categorical variables, SPSS can be used to test whether the variables are independent in the population. The pattern of observed frequencies in the sample is compared to the pattern that would be expected if the variables were independent. If the two patterns are quite different, we conclude that the variables are related in the population.

To illustrate, we shall use the data in the "popular.sav" data file that contains the results of a survey administered to 478 middles school children. One of the survey items directed the children to pick their most important goal: make good grades, be popular, or be good in sports. Consider the hypothesis that goal is independent of gender among children, using a 5% error rate.

To test this hypothesis, open the "popular.sav" data file and follow these steps:

1. Click on **Analyze** from the menu bar.
2. Click on **Descriptive Statistics** from the pull-down menu.
3. Click on **Crosstabs** to open the Crosstabs dialog box.
4. Click on the name of the row variable ("gender") and the **top right arrow button**.
5. Click on the name of the column variable ("goals") and the **middle right arrow button**.
6. Click on the **Cells** button to open the Crosstabs: Cell Display dialog box.
7. Click on **Row** in the Percentages box to indicate that you want percentages by gender (the row variable).
8. Click on **Continue**.
9. Click on the **Statistics** button to open the Crosstabs: Statistics dialog box (see Fig. 12.4).
10. Click on **Chi-Square.**
11. Click on **Continue.**
12. Click on **OK.**

Relevant output is displayed in Figure 12.5.

The test statistic we require is the one labeled Pearson under the Chi-Square heading. For these data the test statistic is $\chi^2 = 21.455$ with 2 degrees of freedom. The *P* value (Asymp. Sig. (2-sided) column) is less than .0005 (and is rounded to .000), leading us to conclude that choice of goal is not independent of gender. Looking at the percentages in the Cross-tabulation table, we see that girls are more likely than boys to choose being popular as a goal (in the sample, 36.3% compared to 22.0%), while boys have a larger tendency than girls to select being good in sports as a goal.

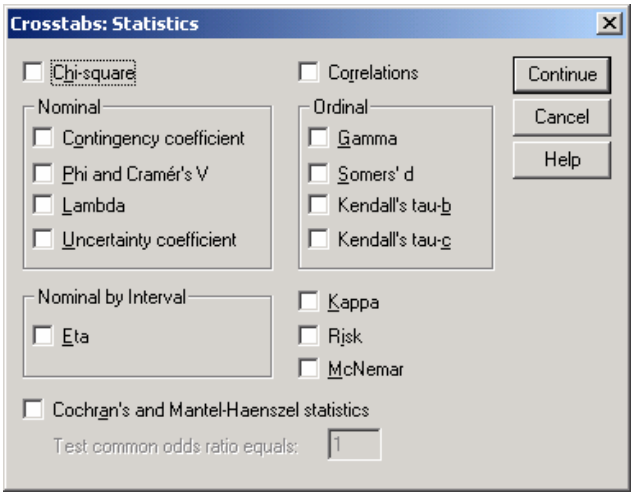

Figure 12.4 Crosstabs: Statistics Dialog Box

GENDER * GOALS Crosstabulation

			GOALS			
			make good grades	be popular	be good in sports	Total
GENDER	girl	Count	130	91	30	251
		% within GENDER	51.8%	36.3%	12.0%	100.0%
	boy	Count	117	50	60	227
		% within GENDER	51.5%	22.0%	26.4%	100.0%
Total		Count	247	141	90	478
		% within GENDER	51.7%	29.5%	18.8%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	21.455[a]	2	.000
Likelihood Ratio	21.769	2	.000
Linear-by-Linear Association	4.322	1	.038
N of Valid Cases	478		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 42.74.

Figure 12.5 Chi-Square Test of Independence Between Gender and Goals

There is additional information available and provided by the Crosstabs procedure. For instance, by default SPSS prints the minimum expected frequency (42.74) and the number (and percent) of cells that have an expected value less than 5. If many of the cells (e.g., 20% or more) have expected values below 5, the data analyst should consider combining some of the response categories. This is accomplished using the Recode procedure (Chapter 1) prior to conducting the chi-square test.

As an option, SPSS will print the expected values for each cell. To obtain the expected frequencies, follow the steps outlined above, but remembering to select the **Expected** option in the Crosstabs: Cell Display dialog box (refer to steps 6 and 7).

12.3 MEASURES OF ASSOCIATION

As with numerical variables, when two categorical variables are not independent, they are said to be correlated or associated with one another. It is possible to calculate an index that measures the degree of association between the two variables. The type of index that is most appropriate depends upon the nature of the variables (nominal or ordinal) and whether one of the variables can be considered a predictor of the other. SPSS labels a correlation in which there is no predictor variable as symmetrical, and a correlation in which there is a predictor as directional.

The Phi coefficient (ϕ) is appropriate for two nominal, dichotomous, variables and symmetrical situations. Nominal variables and a directional situation call for a lambda coefficient (λ). The gamma coefficient (γ) is a correlation for ordinal variables in which there is no prediction (symmetrical); Somer's d is used for ordinal variables in which it is appropriate to predict one variable from the other (directional).

The procedure for computing these coefficients with SPSS is virtually identical. We shall illustrate with "titanic.sav" by computing coefficient lambda for the relationship between class (first, second, third, crew) and survival (no, yes). We consider both these variables nominal, and predict survival from class. To use SPSS to compute this measure of association, open the data file and repeat the Crosstabs procedure described in Section 12.2, with the following features: use class as the row variable and survived as the column variable; select the **Lambda** coefficient instead of the Chi-square statistic in the Crosstabs: Statistics dialog box (see Fig. 12.4).

The output will appear as shown in Figure 12.6. The table shows that the coefficient for survived as the dependent variable is 0.114, which is weak. The test of significance for the λ-coefficients are displayed as *t*-statistics (Approx. T) and *P* values (Approx. Sig.). In this example, $P < .0005$, so we conclude that the variables are related.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
CLASS * SURVIVED	2201	100.0%	0	.0%	2201	100.0%

CLASS * SURVIVED Crosstabulation

Count

		SURVIVED		
		no	yes	Total
CLASS	crew	673	212	885
	first	122	203	325
	second	167	118	285
	third	528	178	706
Total		1490	711	2201

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	.040	.009	4.514	.000
		CLASS Dependent	.000	.000	.[c]	.[c]
		SURVIVED Dependent	.114	.024	4.514	.000
	Goodman and Kruskal tau	CLASS Dependent	.025	.004		.000[d]
		SURVIVED Dependent	.087	.013		.000[d]

a. Not assuming the null hypothesis.

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

c. Cannot be computed because the asymptotic standard error equals zero.

d. Based on chi-square approximation

Figure 12.6 Measure of Association Between Class and Survival Using Lambda

Chapter Exercises

12.1 Using the "titanic.sav" datafile:

a. Perform a chi-square test to determine whether class and survival are independent. Use a significance level of .05; state the test statistic, *P* value, and your conclusion.

b. If you rejected the null hypothesis in part a, describe the nature of the relationship between the two variables.

12.2 The "popular.sav" data file contains information on students' ratings of the importance of such things as making money, looks, sports, and grades

in their lives. The students' responses are on a 4-point ordinal scale, where 1 = most important and 4 = least important.

a. Perform a chi-square test to determine whether gender and importance of money are independent. Use an α level of .05. What is your conclusion?

b. Perform a second test to determine whether gender and importance of looks are independent. Use an α level of .05. What is your conclusion?

12.3 Use the data on interventions aimed at reducing tobacco use among baseball players ("spit.sav") to answer the following:

a. What correlation coefficient is appropriate for determining the relationship of intervention and outcome? Why?

b. Compute the appropriate correlation coefficient. Are the variables related, based on a .05 level of error? If so, describe the nature of the relationship.