Chapter 6

Summarizing Multivariate Data: Association Between Categorical Variables

Chapter 5 describes how to use SPSS to summarize associations between numerical variables. In this chapter, we describe how to summarize two or more categorical variables. These analyses can be used to answer questions such as:

- Is there a relationship between political party and gender? For example, do women tend to vote for Democrats and men for Republicans? Or vice versa?
- Is there a relationship between region of the country and attitude toward capital punishment? For instance, do people in the South tend to favor the death penalty whereas those in the Northeast tend to oppose it?
- Were there more female than male survivors on the Titanic?

6.1 TWO-BY-TWO FREQUENCY TABLES

The relationship between two or more categorical variables is summarized using frequency tables, or cross-classifications of observations. Two-by-two tables are

created when you have two variables, each with two possible outcomes. For example, we may have a sample of people categorized by race (minority, nonminority) and by gender. Frequencies for either variable separately are obtained using the Frequencies procedure. However, it cannot be determined from these individual frequency distributions how many minority males are in the sample, for example. This is accomplished using the Cross-tabulation procedure, which examines the counts of simultaneous occurrences of several values.

Open the data in the football data file ("football.sav"), which has record of favored team (home or away team) and winning team (home or away team) for 250 NFL games. To create a two-way frequency table of favored team by winning team, 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 shown in Figure 6.1.
4. Highlight the "favored" variable by clicking on it, and then move it to the Row(s) box by clicking on the **top right arrow button**.
5. Highlight the "winner" variable by clicking on it, and then move it to the Column(s) box by clicking on the **middle right arrow button**.
6. Click on **OK**.

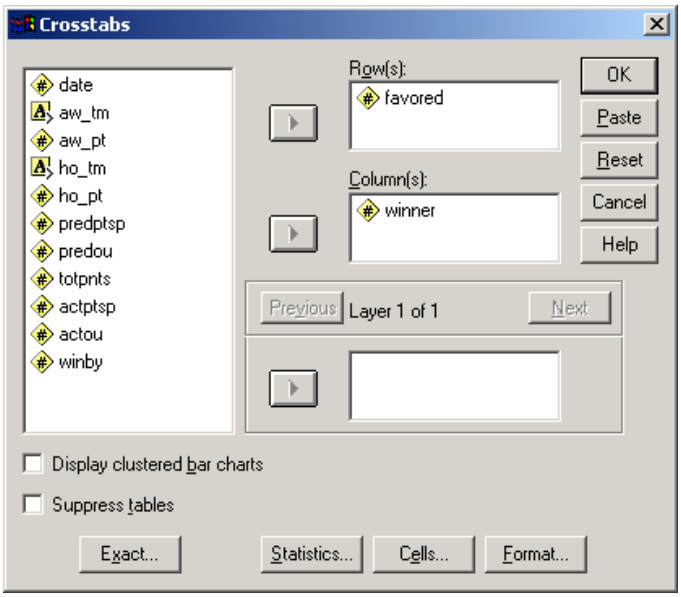

Figure 6.1 Crosstabs Dialog Box

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
favored * winner	250	100.0%	0	.0%	250	100.0%

favored * winner Crosstabulation

Count

		winner		
		home	away	Total
favored	home	125	56	181
	away	31	38	69
Total		156	94	250

Figure 6.2 Cross-tabulation of Favored by Winner

This will produce a cross-tabulation of the row ("favored") by column ("winner") variables as shown in Figure 6.2. The number of cases for each combination of values of the row by column variables is displayed in the cells of the table.

The Case Processing Summary table indicates that there are no missing values for either variable. Examining the favored * winner Cross-tabulation table, we see that favored has two values (home and away) and winner has two values (home and away). A two-by-two table such as this has four "cells."

The number that appears in each of the cells is called the Count (or frequency); this is the number of cases in that cell. In this sample, there are 125 games in which the home team was the favored team and the winning team. In 56 games, the home team was favored and the away team won.

The rows and columns labeled "Total" are the marginal totals and represent the counts for the values of favored and winning team separately (the counts that would be obtained from a simple frequency distribution). For example, in the favored (row) marginals, we see that there were 181 games in which the home team was favored and 69 in which the away team was favored. Likewise, the 156 and 94 winner (column) totals indicate the number of games the home team and away team won, respectively. The number 250 represents the total number of games in the data file.

Calculation of Percentages

In addition to the count for each cell, SPSS will also calculate row, column, and total percentages. The row percentage is the percentage of cases in a row that fall into a particular cell. The column percentage is the percent of cases in a

column that fall into a particular cell. The total percentage is the number of cases in a cell expressed as a percentage of the total number of cases in the table.

Use the file "titanic.sav" to illustrate these percentages. This data file contains sex and survival information for the 2201 passengers on the Titanic. We will run the Crosstabs procedure with "sex" as the row variable and "survived" as the column variable. This will address the question, was it really a case of "women first" — that is, were women evacuated to lifeboats ahead of men?

To calculate row, column, and total percentages, follow steps 1–5 in the preceding section, and then:

1. Click on **Cells** to open the Crosstabs: Cell Display dialog box.
2. Click on each box in the Percentages box to indicate that you wish you calculate **Row**, **Column**, and **Total** percentages.
3. Click on **Continue**.
4. Click on **OK**.

Your output should appear as in Figure 6.3.

The top number in each cell is the count. The cell counts indicate that there were 126 female nonsurvivors, 344 female survivors, 1364 male nonsurvivors, and 367 male survivors. The marginal counts indicate that there were 470 females and 1731 males, and that there were 1490 nonsurvivors and 711 survivors. The total number of passengers, 2201, appears in the bottom right-hand corner of the table.

SEX * SURVIVED Crosstabulation

			SURVIVED		
			no	yes	Total
SEX	female	Count	126	344	470
		% within SEX	26.8%	73.2%	100.0%
		% within SURVIVED	8.5%	48.4%	21.4%
		% of Total	5.7%	15.6%	21.4%
	male	Count	1364	367	1731
		% within SEX	78.8%	21.2%	100.0%
		% within SURVIVED	91.5%	51.6%	78.6%
		% of Total	62.0%	16.7%	78.6%
Total		Count	1490	711	2201
		% within SEX	67.7%	32.3%	100.0%
		% within SURVIVED	100.0%	100.0%	100.0%
		% of Total	67.7%	32.3%	100.0%

Figure 6.3 Cross-tabulation of Sex by Survived with Percentages

Below the count is the "% within SEX," or the row percentage. These figures represent the percentages by values of sex (male and female). That is, 26.8% of females did not survive (which is calculated by dividing 126 by the total number of females in the sample (470) and multiplying by 100). Conversely, 78.8% of the males did not survive (1364 ÷ 1731 × 100). Thus, it seems that women were given preferential treatment and were more likely to survive.

Figure 6.3 also displays "% within SURVIVED." For example, of the 1490 people who did not survive, 8.5% females (126 ÷ 1490 × 100). The fourth line of each cell, labeled "% of Total," contains the total percentages. For example, the 344 female survivors represent 15.6% of the total 2201 passengers.

Phi Coefficient

The phi coefficient is an index of association for a two-by-two table. Like the correlation coefficient, its value is between –1 and 1. Values close to –1 or 1 indicate strong association of two variables, whereas values close to zero indicate a weak association. To compute the phi coefficient, follow steps 1–5 in Section 6.1, and then:

1. Click on **Statistics** to open the Crosstabs: Statistics dialog box.
2. Click on **Phi and Cramer's V** in the Nominal box.
3. Click on **Continue**.
4. Click on **OK**.

For example, the value of the phi coefficient for the data in Figure 6.3 is –.456. This indicates a moderate association of sex with survival; the majority of females survived, and the majority of males did not.

6.2 LARGER TWO-WAY FREQUENCY TABLES

In many studies, categorical variables have more than two values. The number of categories is not limited to two, and virtually any size row-by-column table is possible. For example, "political affiliation" may be recorded in three categories: Conservative, Liberal, and Moderate; "occupation" may have 15 categories representing 15 different job titles. The procedure for calculating counts and percentages, as well as the interpretations of the frequencies, is the same as described in Section 6.1.

As an example, open the "spit.sav" data file. This is a study on the effectiveness of two interventions to help major league baseball players stop

using spit tobacco.* The frequency table for this example is a 3-×-2 table; there are two variables, "outcome" with three levels and "intervention" with two levels.

Using this data set, follow the steps in Section 6.1 to create a frequency table containing counts, and row, column, and total percentages. Your results, omitting the case-processing summary, should look like Figure 6.4. The row variable is the length of intervention (minimum or extended), and the column variable is player outcome (quit; tried to quit; failed to quit). Looking just at the total row, we see that 5 players successfully quit, 15 tried to quit but were unsuccessful, and 34 failed to quit. So, because overall only 9.3% of the players (5 divided by 54) were able to quit, the interventions were not particularly successful.

But what about the relationship between the two variables? That is, is one of the intervention types more successful than the other? To explore this, we look at the "percent within intervention group" figures. We note that of the players in the minimum intervention group, 0% quit successfully, compared with 19.2% in the extended intervention group. Similarly, we see that 89.3% of the minimum intervention group failed to quit, compared with 34.6% of the extended intervention group. Therefore, the extended intervention seems relatively more successful for this sample of baseball players.

intervention group * outcome of intervention Crosstabulation

			outcome of intervention			
			quit	tried	failed	Total
intervention group	minimum	Count	0	3	25	28
		% within intervention group	.0%	10.7%	89.3%	100.0%
		% within outcome of intervention	.0%	20.0%	73.5%	51.9%
	extended	Count	5	12	9	26
		% within intervention group	19.2%	46.2%	34.6%	100.0%
		% within outcome of intervention	100.0%	80.0%	26.5%	48.1%
Total		Count	5	15	34	54
		% within intervention group	9.3%	27.8%	63.0%	100.0%
		% within outcome of intervention	100.0%	100.0%	100.0%	100.0%

Figure 6.4 Cross-tabulation of Outcome by Intervention

*Data reproduced from summary tables in Greene, J.C., Walsh, M.M., & Masouredis, C. (1994). Report of a pilot study: A program to help major league baseball players quit using spit tobacco. *Journal of the American Dental Association, 125,* 559-567.

6.3 EFFECTS OF A THIRD VARIABLE

At times you may be interested in summarizing more than two categorical variables. There are two different approaches for doing so; you may look at the marginal association between each pair of variables, or the conditional association of two variables at particular values of the third. We will illustrate both approaches for three dichotomous variables. Often, the conditional approach reveals interesting patterns of results and interactions of variables that are masked in the marginal analysis.

Marginal Association of Three Dichotomous Variables

The data for this example come from a study of crowding and antisocial behavior in 75 community areas in Chicago. Three characteristics of the communities are cross-classified to examine the relationships among socioeconomic status ("SES"), population density ("pop_dens"), and delinquency rate ("delinq"). Each variable is dichotomous, and has been coded as 1 = low and 2 = high. This data file is named "delinq.sav." The cross-tabulation of these three variables is shown in Table 6.1.

The marginal approach involves examining the association of each pair of variables. With three variables, there are three two-way combinations possible: SES × pop_dens, SES × delinq, and pop_dens × delinq. The three tables appear in Figure 6.5. In order to have SPSS for Windows produce these tables, you need to create three separate 2 × 2 tables using the steps in Section 6.1.

The first table shows the relationship between socioeconomic status and population density. The second and third tables portray the relationship of socioeconomic status with delinquency and population density with delinquency, respectively. Interpreting the first table, we see that 87.5% of the low SES communities have high population density, whereas only 17.1% of the high SES communities have high population density. Thus, low SES tends to be associated with high population density, and vice versa.

Table 6.1 Crosstabulation of 75 Communities by Delinquency, Population Density, and Socioeconomic Status

	Low SES		**High SES**	
	Low population density	**High population density**	**Low population density**	**High population density**
Low delinquency	3	2	27	3
High delinquency	2	33	2	3

ses * population density Crosstabulation

			population density		
			low	high	Total
ses	low	Count	5	35	40
		% within ses	12.5%	87.5%	100.0%
		% within population density	14.7%	85.4%	53.3%
	high	Count	29	6	35
		% within ses	82.9%	17.1%	100.0%
		% within population density	85.3%	14.6%	46.7%
Total		Count	34	41	75
		% within ses	45.3%	54.7%	100.0%
		% within population density	100.0%	100.0%	100.0%

SES * rate of juvenile delinquency Crosstabulation

			rate of juvenile delinquency		
			low	high	Total
SES	low	Count	5	35	40
		% within SES	12.5%	87.5%	100.0%
		% within rate of juvenile delinquency	14.3%	87.5%	53.3%
		% of Total	6.7%	46.7%	53.3%
	high	Count	30	5	35
		% within SES	85.7%	14.3%	100.0%
		% within rate of juvenile delinquency	85.7%	12.5%	46.7%
		% of Total	40.0%	6.7%	46.7%
Total		Count	35	40	75
		% within SES	46.7%	53.3%	100.0%
		% within rate of juvenile delinquency	100.0%	100.0%	100.0%
		% of Total	46.7%	53.3%	100.0%

Figure 6.5 Separate Two-Way Cross-tabulations for Three Dichotomous Variables

The second table indicates that low SES is associated with high rate of juvenile delinquency, and high SES with low rate of juvenile delinquency. The third table indicates that low population density is associated with low juvenile delinquency rate (88.2% of low population density communities have low delinquency rate). And, high population density is associated with high juvenile delinquency rate (87.8%).

population density * rate of juvenile delinquency Crosstabulation

			rate of juvenile delinquency		
			low	high	Total
population density	low	Count	30	4	34
		% within population density	88.2%	11.8%	100.0%
		% within rate of juvenile delinquency	85.7%	10.0%	45.3%
		% of Total	40.0%	5.3%	45.3%
	high	Count	5	36	41
		% within population density	12.2%	87.8%	100.0%
		% within rate of juvenile delinquency	14.3%	90.0%	54.7%
		% of Total	6.7%	48.0%	54.7%
Total		Count	35	40	75
		% within population density	46.7%	53.3%	100.0%
		% within rate of juvenile delinquency	100.0%	100.0%	100.0%
		% of Total	46.7%	53.3%	100.0%

Figure 6.5 *Continued*

Conditional Association of Three Dichotomous Variables

Examining three variables with the marginal approach is useful, but may hide some valuable information regarding how all three variables are related simultaneously. The conditional approach examines the association of two variables at each specific value ("layer") of the third. The relationship between two variables may be maintained, increased, decreased, or even reversed when a third variable is taken into account.

As an example, let's again look at the data on juvenile delinquency rates, SES, and population density. Table 6.2 displays the cell counts from the third table of Figure 6.5.

As we discussed, most of the low-delinquency communities are located in low-density areas, while most of the high-delinquency communities are in high-density areas. Now let us examine the relationship between delinquency and population when the third variable, socioeconomic status, is taken into account. To create a three-way cross-tabulation summary of the data, you need to use the "layer" option in the crosstabs dialog box (Fig. 6.1). Follow the same procedure detailed in Section 6.1, using "pop_dens" as the row variable, "delinq" as the column variable, and "SES" as the Layer 1 of 1 variable.

Your SPSS cross-tabular table should look like that in Figure 6.6.

Table 6.2 Cross-tabulation of Population Density by Delinquency

		DELINQ		
		Low	**High**	**All**
POP_DENS	Low	30	4	34
	High	5	36	41
	All	35	40	

population density * rate of juvenile delinquency * SES Crosstabulation

SES				rate of juvenile delinquency		Total
				low	high	
low	population density	low	Count	3	2	5
			% within population density	60.0%	40.0%	100.0%
			% within rate of juvenile delinquency	60.0%	5.7%	12.5%
			% of Total	7.5%	5.0%	12.5%
		high	Count	2	33	35
			% within population density	5.7%	94.3%	100.0%
			% within rate of juvenile delinquency	40.0%	94.3%	87.5%
			% of Total	5.0%	82.5%	87.5%
	Total		Count	5	35	40
			% within population density	12.5%	87.5%	100.0%
			% within rate of juvenile delinquency	100.0%	100.0%	100.0%
			% of Total	12.5%	87.5%	100.0%
high	population density	low	Count	27	2	29
			% within population density	93.1%	6.9%	100.0%
			% within rate of juvenile delinquency	90.0%	40.0%	82.9%
			% of Total	77.1%	5.7%	82.9%
		high	Count	3	3	6
			% within population density	50.0%	50.0%	100.0%
			% within rate of juvenile delinquency	10.0%	60.0%	17.1%
			% of Total	8.6%	8.6%	17.1%
	Total		Count	30	5	35
			% within population density	85.7%	14.3%	100.0%
			% within rate of juvenile delinquency	100.0%	100.0%	100.0%
			% of Total	85.7%	14.3%	100.0%

Figure 6.6 Cross-tabulation of 75 Communities by Delinquency, Population Density, and Socioeconomic Status

Figure 6.6 reveals a pattern that was not evident with the marginal associations shown in Figure 6.5 and Table 6.2. Among low SES areas, 60% of the low population density areas have low juvenile delinquency rate, and 94.3% of the high population density areas have a high delinquency rate. The story is not the same in high-SES areas, however. For instance, in high-SES areas that have high population density, the juvenile delinquency rate is split equally (50%) between low and high. Therefore, the pattern between population density and juvenile delinquency is different in different SES areas.

Chapter Exercises

6.1 Using the "titanic.sav" file, examine the relationship between "class" and "survival." Create a cross-tabulation table, with cell counts, and row, column, and total percentages.

- **a.** What percentage of first class passengers survived? Of the third class passengers?
- **b.** How many crew members were there? What percentage of the crew survived?
- **c.** Do you see a pattern regarding the relationship between survival and class?

6.2 Using the "fire.sav" data, use SPSS to do a cross-tabulation of race and sex and answer the following questions:

- **a.** What percentage of all firefighter applicants were minority females?
- **b.** Of the male applicants, were there more minority or white applicants?
- **c.** Calculate the phi-coefficient. How strong is the relationship between race and sex?

6.3 Using the "popular.sav" data, create a two-way frequency table of goals and gender

- **a.** What is the relationship, if any, between the two variables?
- **b.** Create a new crosstabulation table controlling for urbanicity. Does the location of the school affect the relationship between goals and gender? If so, how?

Part III

Probability