Chapter 7
Crosstabulation:
Understanding Bivariate Relationships Between Categorical Variables

7.1 Introduction to Crosstabs

So far we have discussed distributions of scores on a single variable. We now turn our attention to the topic of **bivariate distributions** - ones which characterize *relationships* between *two variables* simultaneously.

7.1a Bivariate frequency tables

These types of analyses examine variability in two or more distributions of scores to determine whether or not there is any pattern of covariation, or commonality, between the variables. In Chapter 8 we will discuss bivariate relationships between *quantitative* (continuous) variables. In the present chapter we consider relationships between *qualitative* (categorical) variables.

Much research involves data that are categorical rather than continuous. In fact, the variable transformations we described in Chapter 5 involved converting continuous variables (e.g. **masctot**, on which employees could score anywhere between 5 and 35) into dichotomous categorical variables (e.g., **masc**, in which employees were placed into just two groups - high vs. low masculinity).

We can examine relationships between categorical variables via an extension of frequency analysis. We could use the **Frequencies** procedure to construct a frequency table of the number of EZ employees who fall into the low vs. the high masculinity categories. However, a more interesting (and potentially more important) question we could ask is whether the frequency distribution of low vs. high masculinity is similar or different for male vs. female employees.

We could generate separate frequency tables of this variable for each sex to address this question, but that would be unsystematic. Further, SPSS includes a procedure that is specifically designed to generate frequency distributions for two variables simultaneously. This procedure is called **Crosstabs**, and it produces tables referred to as *crosstabulations*, because they tabulate or count the frequencies of values across two variables simultaneously.

Thus, **Crosstabs** allows us to answer questions such as whether there is a *relationship* between masculinity level and gender. A reasonable hypothesis, for example, might be that male employees are more likely than female employees to be in the high masculinity category, and female employees are more likely than male employees to be in the low masculinity category.

We could test this hypothesis by examining the frequencies obtained in a crosstabulation of the variables **masc** and **gender**. Our printout would contain a table that indicates:

- the number of men who are in the low-masculinity category
- the number of men who are in the high-masculinity category
- the number of women in the low-masculinity category
- the number of women in the high-masculinity

This will allow us to compare the number of men and women in the low/high masculinity categories, but it will also allow us to examine frequencies **within** gender. For example, the table would also permit us to examine whether women are more likely to be in the low-masculinity category than in the high-masculinity category, and vice-versa for men. Further, **Crosstabs** can calculate a statistic that allows us to determine whether any observed relationship between these variables is a statistically significant one or due to chance.

There are numerous other questions relevant to our project that could be addressed using **Crosstabs** to generate bivariate frequency distributions (for example, whether the high-masculinity category is most frequently associated with the high-task skill category of leader style). You're encouraged to explore these relationships (recall that we have several categorical variables in the **ezdata.sav** file: gender, masc, fem, task, soc). We will use the**mes** and **gender** example to illustrate the **Crosstabs** procedure in this chapter, and you will be asked to do the same analysis for **fem** and **gender** in the exercise at the end of the chapter.

7.1b The logic of the Chi Square statistic

The preceding discussion suggests that the two variables, **gender** and **masc**, are not independent, i.e., there is a *relationship* between one's gender and one's masculinity level. We will test this hypothesis by examining the frequencies obtained in a crosstabulation of the variables.

To illustrate how we could determine whether or not two variables are related using a crosstabulation table, we will simplify things for the moment and assume that there are only 50 men and 50 women in our sample who have been classified as either low or high in masculinity (in our ezdata file, there are actually 110 men and 118 women).

The crosstabulation table of this example would have four cells that comprise a matrix of the four possible combinations of the two levels of the two variables. The data in the cells would be the frequencies (i.e., the number) of men and women who were classified as either low or high in masculinity.

The table would present these four cells as a **2 x 2 contingency matrix**. A contingency matrix classifies individuals into a given cell contingent upon their exhibiting a

particular *combination* of one level of the first variable combined with one level of the second variable (e.g., being *both* a man and in the high masculinity category).

If there is *no* relationship between **gender** and **masc** (i.e., if they are independent), then the printout would show just as many low-masculine men as high-masculine men, and the same would hold for female employees. Further, among high-masculinity employees, there would be an equal frequency of men and women, and the same would be true for low-masculinity employees. In other words, the 100 men and women would distribute themselves evenly across the four cells. Thus, the crosstabulation of frequencies would show equal frequencies in all four cells of this contingency table (25 per cell) as shown in Table 10.1.

Table 7.1

		Gender		
		Male	Female	**Total**
Masculinity	Low-masculine	25	25	50
	High-masculine	25	25	50
	Total	50	50	100

In looking at this hypothetical table, we can see that, indeed, there is no pattern of frequencies beyond what would be expected by chance, indicating that the two variables are independent of each other. Specifically, looking down the *columns* of Table 10.1, we see that among the total of 50 men, 25 are low-masculine and 25 are high-masculine. The same is true for women - 25 are low-masculine and 25 are high-masculine. Thus, from this hypothetical crosstabulation, we would conclude that men are equally-likely to be low-masculine or high-masculine, and so are women.

Further, looking across the *rows*, among the total of 50 low-masculine employees, 25 are men and 25 are women. And for the 50 high-masculine employees, 25 are men and 25 are women. Thus, we would conclude that men are equally likely as women to be low in masculinity, and that women are equally likely as men to be high in masculinity. Again, this is the pattern of frequencies that would be expected by chance alone if the two variables are unrelated, as is the case in this table.

However, if the hypothesized relationship really exists, then there would be a pattern of frequencies that is different from chance expectations. A hypothetical example of a pattern reflecting a real relationship between **gender** and **masc** is illustrated in Table 10.2.

Table 7.2

		Gender	

		Male	Female	**Total**
Masculinity	Low-masculine	10	40	50
	High-masculine	40	10	50
	Total	50	50	100

It can be seen from the hypothetical data in Table 10.2 that there is a higher frequency of men who are high-masculine (40 out of 50) than men who are low-masculine (only 10 out of 50). Further, there is a higher frequency of women who are low-masculine (40 out of 50) than women who are high-masculine (only 10 out of 50). It can also be seen that among the 50 high-masculine employees, there is a higher frequency of men (40) than women (10). Last, among the 50 low-masculine employees, there is a higher frequency of women (40) than men (10).

Thus, this table shows that there is a clear relationship between an employee's gender and his/her masculinity level. Masculinity level varies systematically across gender, with men being more likely than women to be high-masculine and women being more likely than men to be low-masculine.

Of course, the above examples were made to be very clear cut. Bivariate frequency distributions often do not lend themselves to such an easy visual determination of whether the frequencies indicate that the two variables are related or not. Only by running the crosstabulation procedure and computing the appropriate statistical test could we answer the question about this hypothesized relationship between gender and masculinity. The statistical test of interest here is called ***Chi square***.

As we will see, SPSS will compute a ***Pearson chi square*** value that will answer the question of whether the actual data in our ezdata file demonstrate a **statistically significant** relationship (i.e., a *real* one) between gender and masculinity, or if they will show that the two variables are statistically independent (i.e., any apparent pattern in frequencies is not real, but due to *random chance*).

Chi square is computed based on a comparison of actual frequencies observed in our sample to that which would be expected to occur by chance alone. If there is a large difference between the observed vs. the expected frequencies, a large value for ***Chi square*** will be obtained. More importantly, the ***probability*** associated with this ***Chi* square** value is computed. This value determines whether the *chi square* value is statistically significant. The general convention used by researchers is that if this probability is **.05** or lower, then we reject random chance as an explanation, and conclude that this is a real (statistically significant) relationship.

Thus, ***Chi square*** is an inferential statistic - it allows us to make inferences from our sample to the population regarding the hypothesized relationship. This process begins

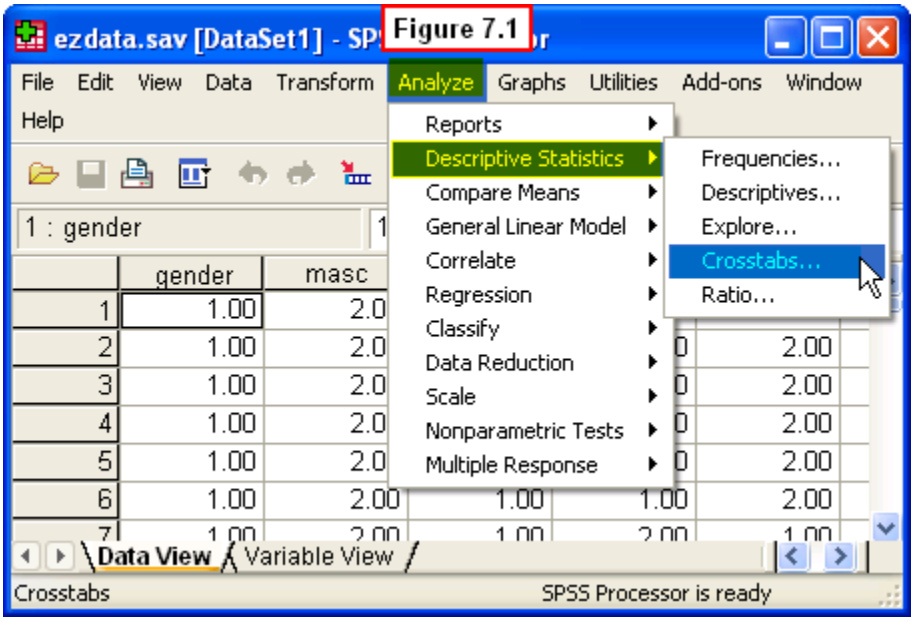
with the assumption that the relationship is due to chance - this is called the **null hypothesis** (**Ho**). Based on the obtained probability, we either retain or reject the null hypothesis using the following decision rule:

- If the probability is < or = .05, reject **Ho** and conclude the relationship is significant
- If the probability is > .05, retain **Ho** and conclude the relationship is due to chance.

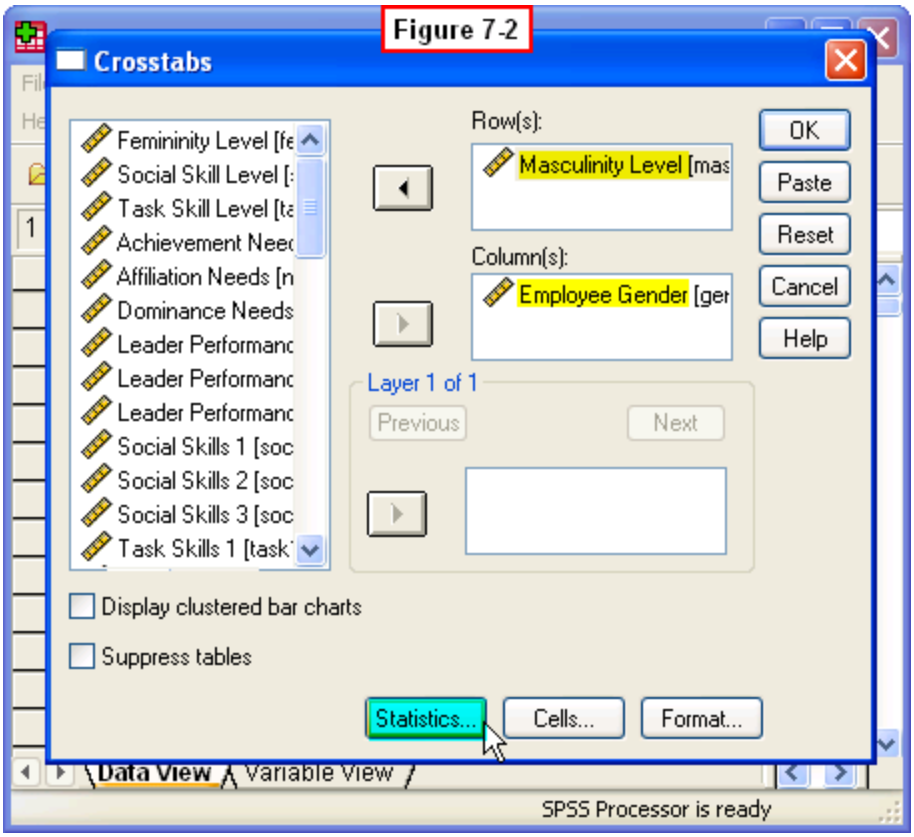
Thus, if the obtained probability is less than or equal to .05, we would conclude that the pattern of frequencies discussed in the crosstabs table is a real one (not due to chance), indicating that masculinity level is, indeed, significantly related to gender.

7.2 Running the Crosstabs Procedure

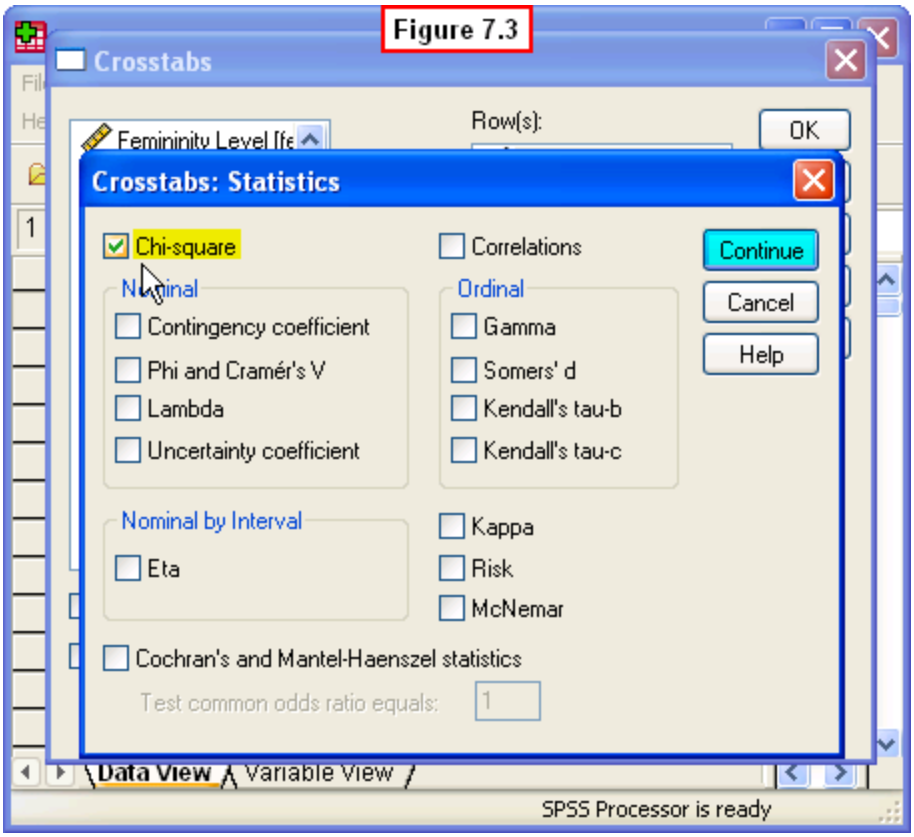
To begin, open **ezdata.sav** and select **Analyze, Descriptive Statistics, Crosstabs...** from the Data Editor menu (Figure 7.2).



Select **Employee Gender** In the left pane of the **Crosstabs** dialog window (Figure 7.2), then click the right-arrow between panes to move this variable into the **Column(s)** pane on the right. Select **Masculinity Level** in the left pane and move this variable to the **Row(s)** pane on the right. Click the **Statistics...** button at the bottom of this window.



Click the **Chi-square** checkbox in the **Crosstabs: Statistics** dialog window (Figure 7.3), then click the **Continue** button on the right.



Show Me Video!

7.3 Interpreting the Output

The first table of the resulting output file simply lists the number of cases processed (228 total employees). The next table in the output presents the 2 x 2 contingency matrix, or the crosstabulation of frequencies listing the number of male/female employees in the low/high masculinity categories (Figure 7.4). The frequencies highlighted in yellow constitute the four cells of this matrix, and we will need to interpret these if the **Chi square** test reveals that the variables of **gender** and **masculinity level** are significantly related.

For now, looking in the column totals, we see that of the 228 employees 110 are male and 118 female. Looking in the row totals we see that of the 228 employees, 88 are in the low-masculinity category and 140 are in the high-masculinity category. Thus, there is a fairly equal number of male and female employees, and in general, employees are much more likely to be high-masculine than low-masculine.

Figure 7.4

Masculinity Level * Employee Gender Crosstabulation

Count

		Employee Gender		
		Male	Female	Total
Masculinity Level	Low-Masculinity	29	59	88
	High-Masculinity	81	59	140
Total		110	118	228

Examination of the *pattern* of highlighted yellow frequencies in the *cells* of the matrix involves:

- comparing the number of men vs. women in within masculinity levels (the horizontal red double-arrows), then
- comparing the number of low- vs. high-masculine employees within gender (the vertical blue double-arrows).

Recall that if these two variables are independent (i.e., not significantly related), then the frequencies would be relatively evenly distributed across masculinity level and gender. That is, there would be relatively equal numbers of men and women in the low-masculinity category as well as in the high masculinity category. Further, there would be an equal number of low-masculine men as high-masculine men, and an equal number of low-masculine vs. high-masculine women.

Scanning these four frequency counts, it appears that the above is generally not true. Rather, there is a pattern of frequencies across the four categories that indicates unequal frequencies. Before we can interpret this apparent pattern, however, we need to examine the **Chi square** statistic to determine whether or not that this pattern reflects a statistically significant relationship. The **Pearson Chi-square** value computed by SPSS in shown in the last table of the output (Figure 7.5).

Figure 7.5

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	13.420[b]	1	.000		
Continuity Correction[a]	12.441	1	.000		
Likelihood Ratio	13.626	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	13.361	1	.000		
N of Valid Cases	228				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 42.46.

The **Chi square** value is 13.42. Recall that to determine whether or not this value indicates a significant relationship, we need to examine the **probability** that this distribution of frequencies occurred by chance alone. Recall that the conventional probability level used to answer this question is **.05** and the following decision rule is employed:

- if the probability is **greater than .05**, then the variables are **not** significantly related
- if the probability is **less than or equal to .05**, then the variables **are** significantly related.

Instead of using the term, **probability**, SPSS uses the term, **significance**, and abbreviates it as **Sig.** More specifically, in Figure 7.5, the **Assym. Sig.** (2-sided) column lists the probability of interest. The probability is **.000** in this column for the **Pearson Chi-square** statistic.

Note that this means the probability is actually something lower than .0005 (SPSS rounds off to three decimal places). Thus, while the actual probability may not be *exactly* equal to zero, it is certainly less than the cut-off value of .05. So since the probability is less than .05, we can reject the null hypothesis of chance as an explanation and conclude that there *is* a statistically significant relationship between **gender** and **masculinity**. This justifies our interpretation of the pattern of frequencies in the cells of Figure 7.4. Examination of this matrix in Figure 7.4 reveals that the pattern is more similar to that shown in Table 7.2 than that in Table 7.1.

Starting with the comparisons between men and women (the horizontal red double-arrows), we see that of the 88 low-masculine employees, 59 are women, while only 29 are men. Thus, there were more female EZ employees than male employees in the low masculine category. Further, of the 140 high-masculine employees, 81 are men compared to 59 women. Thus, there were more male employees than female employees in the high-masculine category.

The above pattern follows social stereotypes about sex role identity. That is, given the emphasis on masculinity in the socialization process, it is perhaps not surprising that our sample would have more low-masculine women than men, and more high-masculine men than women.

However, comparisons within genders (the vertical blue double-arrows) reveal a different and interesting pattern. On the one hand, social stereotypes can still be seen for men: of the 110 men, 81 were in the high-masculinity category, compared to only 29 in the low-masculinity category. Thus, the majority of male employees were high-masculine.

On the other hand, the comparison within female employees disconfirms social stereotypes: of the 118 women, half were low-masculine (59) and half were high-masculine (59). Thus, at least at EZ Manufacturing, there were just as many high-masculine as low-masculine women. This illustrates the truism that research and data analyses sometimes confirm hypotheses and expectations, but sometimes they yield surprises - this one reason research and data analysis can be so interesting!

There are a variety of possible explanations of the large number of high-masculine female employees, some of which we will discuss in later chapters. For now, one possible explanation is that EZ is a high-tech manufacturing firm, and technological-orientation is stereotypically associated with masculinity in our culture. Thus, perhaps female EZ employees either adapted their personalities to fit the masculinity stereotype, or perhaps high-masculine women self-selected to work at EZ rather than in more traditionally feminine work roles.

We will gain more insights into the relationships among variables in from our project in Chapter 8, where we introduce a commonly used analytical approach known as correlation. This type of analysis also concerns bivariate distributions, but rather than examining frequencies to assess relationships (as in **Crosstabs**), correlation directly assesses covariation in scores on two variables.

7.4 Chapter Review Video

Review Me!

7.5 Try It! Exercises

1. Examining the relationship between Gender & Femininity Level using Crosstabs

In the chapter example we saw there was a relationship between gender and masculinity level among EZ employees. Is there a relationship between gender and femininity level? Use the **Crosstabs** procedure described in Section 7.2 to create a crosstabulation of frequencies on the variables, **Gender** and **Femininity Level**.

- In the **Crosstabs** dialog window, move **Employee Gender** to the **Column(s)** box, and move **Femininity Level** to the **Row(s)** box. Then click the **Statistics** button.
- Check the **Pearson Chi-square** box in the **Crosstabs: Statistics** dialog window, then click the **Continue** button.
- Click the **OK** button in the **Crosstabs** window. The crosstabulation and **Chi square** tables will appear in an Output Viewer window.
- **Print** your output file to submit to your instructor.
- **Write an interpretation of the results** to submit to your instructor. Follow the example in Section 7.3, stating your conclusions:
    - Interpret the **Pearson Chi-square** value; is **Gender** significantly related to **Femininity Level**, or are these variables independent of each other?
    - State the relative number of male vs. female employees in the low-femininity category
    - State the relative number of male vs. female employees in the high-femininity category
    - State the number of males who are in the low- vs. the high-masculinity category
    - State the number of females who are in the low- vs. the high-masculinity category
- **Write a statement explaining** these patterns of frequencies. That is, how might these patterns be explained in terms of social stereotypes about gender and femininity level?