

Chapter 13

Chi-Square

This section covers the steps for running and interpreting chi-square analyses using the SPSS **Crosstabs** and **Nonparametric Tests**. Specifically, we demonstrate procedures for running two separate types of nonparametric chi-squares: The Goodness-of-Fit chi-square and Pearson's chi-square (Also called the Test of Independence). As discussed in earlier chapters, every statistical test is designed for a specific type of data (i.e., nominal, ordinal, interval, or ratio) and both chi-square procedures are most commonly used with nominal or group data. The goodness-of-fit chi-square is used when you have one nominal variable and you want to know whether the frequency of occurrence (e.g., the number of people in each group) differs from what we would expect by chance alone. The Pearson's chi-square is used to ask questions about two nominal variables, and it can be used to determine whether two nominal variables are associated in some manner.

For the following examples, we have recreated the data set based on Cartoon 13.1 (a visit to McDonald's), which is found in Table 13.1. Both variables in this example are nominal variables representing discrete groups. The independent variable is whether the adolescent bovines demonstrate

Cartoon 13.1 RUBES by Leigh Rubin



"That's it! If you kids don't start behaving, I'm taking you both to McDonald's!"

Table 13.1 Visited McDonald's by Disruptive Behavior of Adolescent Bovines

		<i>Disruptive Behavior</i>	
		<i>Yes</i>	<i>No</i>
Visited McDonald's	Yes	38	15
	No	7	40

disruptive behavior or not

and the dependent

variable is whether their

mother takes them to

McDonald's or not.

Setting Up the Data

Figure 13.1 presents the variable view of the SPSS data editor where we have defined two discrete variables. The first variable represents the two groups of bovine adolescents; those that demonstrate disruptive behavior and those that do not. We have given it the variable name **behavior** and given it the variable label “Disruptive Behavior of Adolescent Bovines (yes/no).” The second variable also represents two group of bovine adolescents; those that get taken to McDonald’s and those that to not. We have given it the variable name **mcdonlds** and the variable label “Visited McDonald’s (yes/no).”

Since the both variables represent discrete groups, for each variable we need to assign numerical values to each group and provide value labels for each of the groups. Again, this is done by clicking on the cell in the variable row desired in the **Values** column and then clicking on the little gray box that appears inside the cell. In the **Value Labels** dialogue box you can pair numerical values with labels for each group. For the **behavior** variable we have paired 1 with the label “yes,” representing the bovine adolescents who have been disruptive; and 0 with the label “no,” representing the bovine adolescents who did not demonstrate disruptive behavior. For the **mcdonlds** variable we have paired 1 with the label “yes,” representing the bovine adolescents who’s mothers took them to McDonald’s; and 0 with the label “no,” representing the bovine adolescents who’s mothers did not take them to McDonald’s.

Figure 13.1 SPSS: Variable View for Visiting McDonald's Data

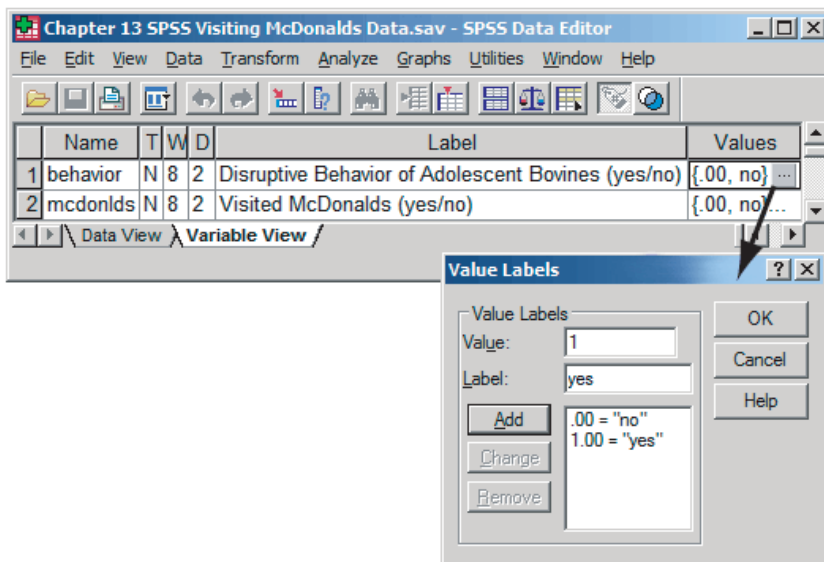


Figure 13.2 SPSS: Data View for Visiting McDonald's Data

Figure 13.2 consists of two screenshots of the SPSS Data Editor. The first screenshot, labeled 'A', shows the data view for rows 1 through 9. The second screenshot, labeled 'B', shows the data view for rows 92 through 100. Both screenshots show two columns: 'behavior' and 'mcdonlds'.

Case	behavior	mcdonlds
1	yes	yes
2	yes	yes
3	yes	yes
4	yes	yes
5	yes	yes
6	yes	yes
7	yes	yes
8	yes	yes
9	yes	yes
...
92	no	no
93	no	no
94	no	no
95	no	no
96	no	no
97	no	no
98	no	no
99	no	no
100	no	no

Figure 13.2 presents the data view of the SPSS data editor. Here, we have entered the visiting McDonald's data for the 100 adolescents in our sample. Remember that the columns represent each of the different variables and the rows represent each observation, which in this case is each bovine adolescent. For example, the first adolescent (found in part A of the figure)

behaved disruptively and its mother took him/her to McDonald's. Similarly, the 100th adolescent bovine (found in part B of the figure) did not behave disruptively and his/her mother did not take him/her to McDonald's.

The Goodness-of-Fit Chi-Square

The goodness-of-fit chi-square allows us to determine whether the observed group frequencies, for a single discrete variable, differ from what we would expect by chance alone. That is, it tells us whether observed differences in group frequencies is random or not. In many ways this chi-square is identical to the Pearson's chi-square. In fact, they both use the same chi-square formula and the significance of the statistic is tested in the same manner. However, they differ in several important ways. First, the goodness-of-fit chi-square only looks at one variable at a time, while the Pearson's chi-square evaluates the pattern of frequencies across two discrete variables. Second, because the goodness-of-fit test only evaluates one variable at a time, it does not tell us anything about the relationship between two variables. Finally, the goodness-of-fit test often assumes that the expected values for each group are equal. For example, with two groups we may expect 50% of our sample to be in each group. Alternatively, for 3 groups we may expect

33.33% of our sample to be in each group. In some cases the expected values for each group may not be equal. For example, if the variable of interest represents ethnicity of the sample (e.g., European Am., African Am., etc.), we would not find equal numbers of each group in the population, and therefore would not expect equal numbers of people in each group if the sample were randomly drawn from the population.

In our present example we will calculate two separate goodness-of-fit chi-square tests, one for each variable in our data set. For the **behavior** variable, the results of the chi square test will tell us whether the number of bovine adolescents who do and do not misbehave differs from a 50%/50% pattern or whether one occurs more frequently than the other. Similarly, for the **mcdonlds** variable, the results of the chi-square test will tell us whether the number of bovine adolescents who get taken to MacDonalds by their mother is significantly greater than or less than the number of adolescents who do not get taken to MacDonalds.

Running the Analysis

Goodness-of-Fit Chi-Square (See

Figure 13.3 SPSS: Running Goodness of Fit Chi Square

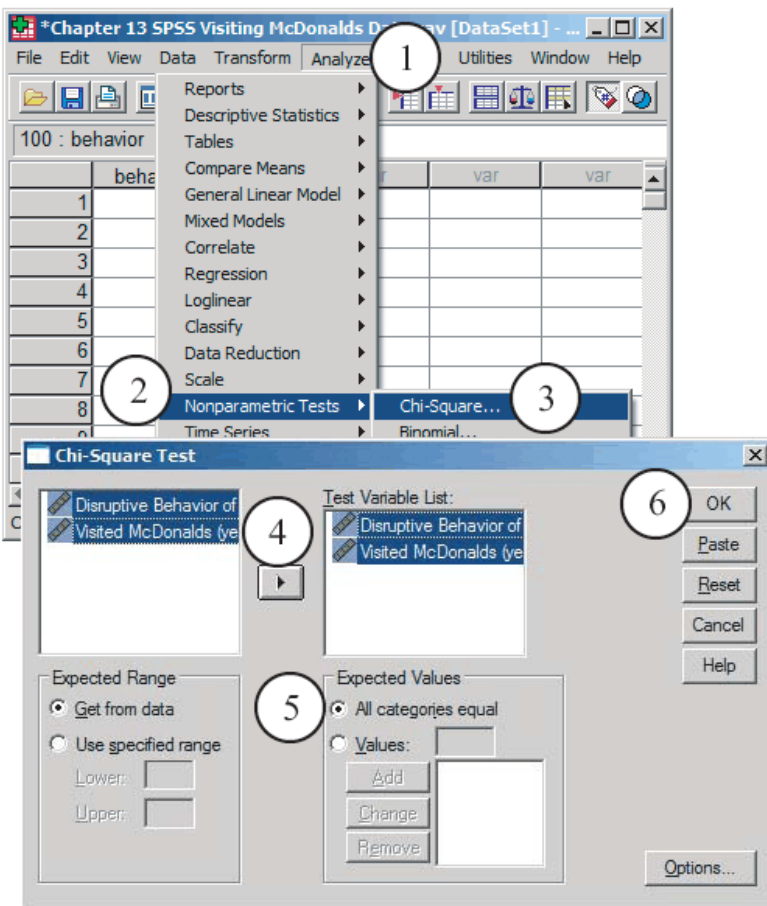


Figure 13.3): From the **Analyze** (1) pull down menu select **Nonparametric Tests** (2), then select **Chi-Square...** (3) from the side menu. In the **Chi-Square Test** dialogue box, enter the variables **behavior** and **mcdonlds** in the **Test Variable List:** field by either double left-clicking on each variable or selecting each variable and left-clicking on the boxed arrow (4) pointing to the right. Next, decide whether the expected values for each group are equal or unequal. For unequal expected values enter the number of cases expected to be found in each group in the

values field and click add. In this case, we are using equal expected frequencies for each group and have selected the **All categories equal** option (5) under the **Expected Values** options. Finally, double check your variables and options and either select **OK** (6) to run, or **Paste** to create syntax to run at a later time.

If you selected the paste option from the procedure above, you should have generated the following syntax:

NPAR TEST

/CHISQUARE=behavior mcdonlds

/EXPECTED=EQUAL

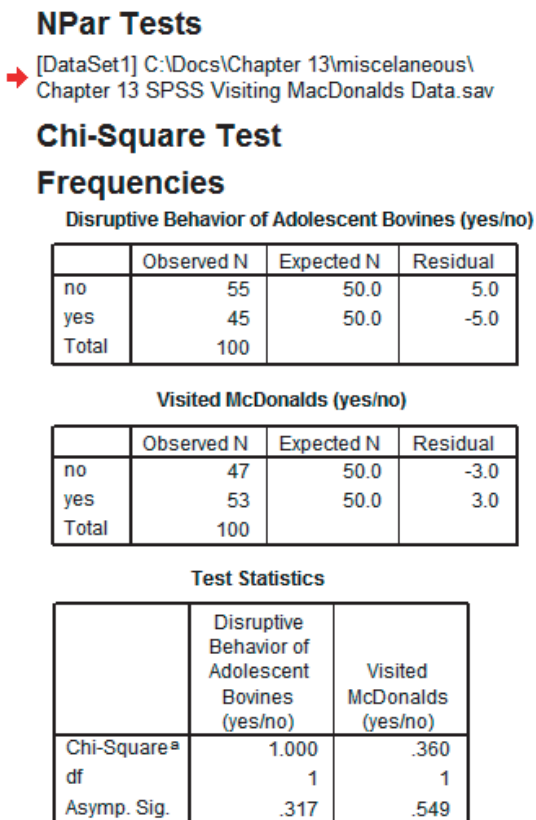
/MISSING ANALYSIS.

To run the analyses using the syntax, while in the Syntax Editor, select **All** from the **Run** pull-down menu.

Reading the Goodness-of-Fit Chi-Square Output

The Goodness-of-Fit Chi-Square Output is presented in Figure 13.4. This output consists of two

Figure 13.4 SPSS: Goodness of Fit Chi-Square Output



a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 50.0.

major parts: Frequencies and Test Statistics. The frequencies output reports the observed, expected and residual frequencies for each group. The residual frequency represents the difference between the expected and the observed frequencies and is obtained by subtracting the expected from the observed frequencies. In our example, since we requested two separate analyses (one for each variable), we are presented with two blocks of frequency output.

The first block of the frequencies output represents the data for the **behavior** variable. In our sample 55 of the adolescent bovine did not misbehave, 50 were expected to misbehave, and the difference between the observed and the expected values is 5.00. Also, in our sample 45 of the adolescent bovine did misbehave, 50 were expected to

misbehave, and the difference between the observed and the expected values is -5.00.

The second block of frequencies output represents the data for the **mcdonalds** variable. In our sample 47 of the adolescent bovine were not taken to MacDonalds, 50 were expected to be taken to MacDonalds, and the difference between the observed and expected values is -3.00. Also, in our sample 53 of the adolescents did visit MacDonalds, 50 were expected to visit MacDonalds, and the difference between the observed and the expected values was 3.00.

The test statistics output reports the chi-square obtained, the degrees of freedom (# of groups -1), and the exact level of significance are reported in separate columns for each of the analyses we requested. For the **behavior** variable the chi-square obtained is 1.00. With 1 degree of freedom (2 group - 1 = 1) and a significance level of .317, which falls well above the .05 alpha level, the difference between the observed and expected values is not significant. Thus we can conclude that the number of bovine adolescents who misbehave does not significantly differ from the number who do misbehave.

With respect the **mcdonalds** variable, the chi-square obtained is .360. With 1 degree of freedom and a significance level of .549, which falls well above the .05 alpha level, the difference between the observed and expected values is not significant. Again, we can conclude that the number of adolescents who's mothers make them visit MacDonalds does not significantly differ from the number who's mother did not make them visit MacDonalds. Statistically, it is a 50%/50% relationship.

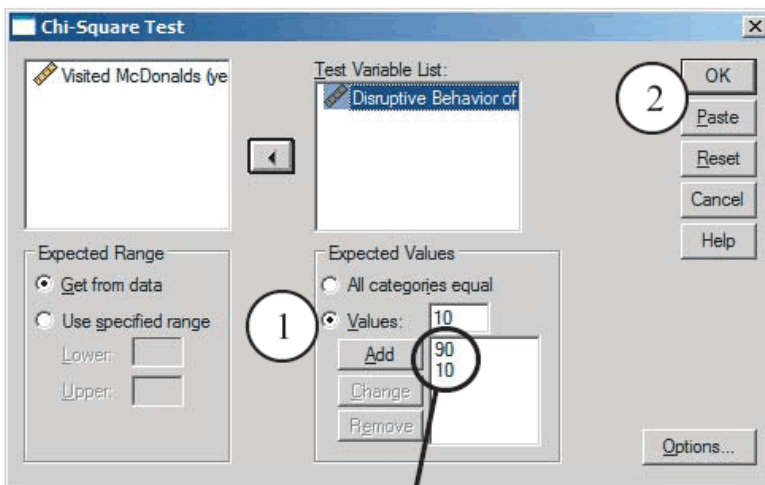
A Second Goodness-of-Fit Chi-Square Example: Unequal Expected Values

In our last example of the Goodness-of-Fit Chi-Square, the expected values for each group were equal because of the question we were asking. In that case we wanted to know whether the group frequencies differed from each other, and therefore differed from a 50%/50% pattern. However, if we change our question of interest, then we will need to change the expected values accordingly. Using the adolescent bovine disruptive behavior variable as an example, assume that we wanted to know whether the frequency of disruptive behavior in the sample differed from the frequencies found in the general population

of adolescent bovines. In this case, the expected values we use will depend on the frequency found in the general population. For example, assume we find that only 10% of general population of adolescent bovines demonstrate disruptive behavior. If our sample is representative of the population, then we should expect 10 bovines in our sample to show disruptive behavior and 90 bovines in our sample not show disruptive behavior. Thus, our expected values are 10 for the “yes” group and 90 for the “no” group.

Figure 13.5 presents the **Chi-Square Test** dialogue box and the resulting output for a goodness-of-fit chi-square using unequal expected values for the **behavior** variable. In the **Expected Values** options of the **Chi-Square Test** dialogue box we have selected the unequal expected values option by left-clicking on

Figure 13.5 SPSS: Goodness of Fit Chi-Square with Unequal Expected Values



the **Values:** option (1). In the **Values:** field, we first entered the expected value for the “no” groups (no disruptive behavior), which was 90, and then left clicked the **Add** button. Next, we entered the expected value for the “yes” group, which was 10. Finally, we ran the analysis by clicking **OK**.

Selecting the paste option from the procedure above generates the following syntax:

```

NPAR TEST
  /CHISQUARE=behavior
  /EXPECTED=90 10
  /MISSING ANALYSIS.

```

To run the analyses using the syntax, while in the Syntax Editor, select **All** from the **Run** pull-down menu.

NPar Tests

→ [DataSet1] C:\Docs\Chapter 13\miscellaneous\Chapter 13 SPSS Visiting MacDonalDs\Data.sav

Chi-Square Test

Frequencies

Disruptive Behavior of Adolescent Bovines (yes/no)

	Observed N	Expected N	Residual
no	55	90.0	-35.0
yes	45	10.0	35.0
Total	100		

Test Statistics

	Disruptive Behavior of Adolescent Bovines (yes/no)
Chi-Square ^a	136.111
df	1
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 10.0.

Again, the Chi-Square Test output is split into two parts: Frequencies and Test Statistics. The frequencies output reports the obtained, expected, and residual (obtained - expected) values for each of our groups. As we requested in the **Chi-Square Test** dialogue box, the expected values for the “no” and “yes” groups are 90 and 10, respectively. These expected values are substantially different from the observed values and result in quite large residuals (-35 and 35) for the “no” and “yes” groups.

The test statistics output reports the chi-square obtained, the degrees of freedom (# of groups -1), and the exact level of significance for the **behavior** variable. Compared to the analysis we ran where the expected values were equal, the substantially larger residual values resulted in a substantially larger chi-square obtained, 136.111. Thus, difference between the observed frequencies found within our sample and the frequencies expected in the general population is significant at least at the .001 level, with one degree of freedom. This indicates to us that there is less than a 1 in 1000 chance that behavior of our sample of adolescent bovines is actually representative of the population of adolescent bovines from which they were drawn. As you can see, changing the research question that we ask substantially changes the results that we obtain using the same data and same statistical procedures.

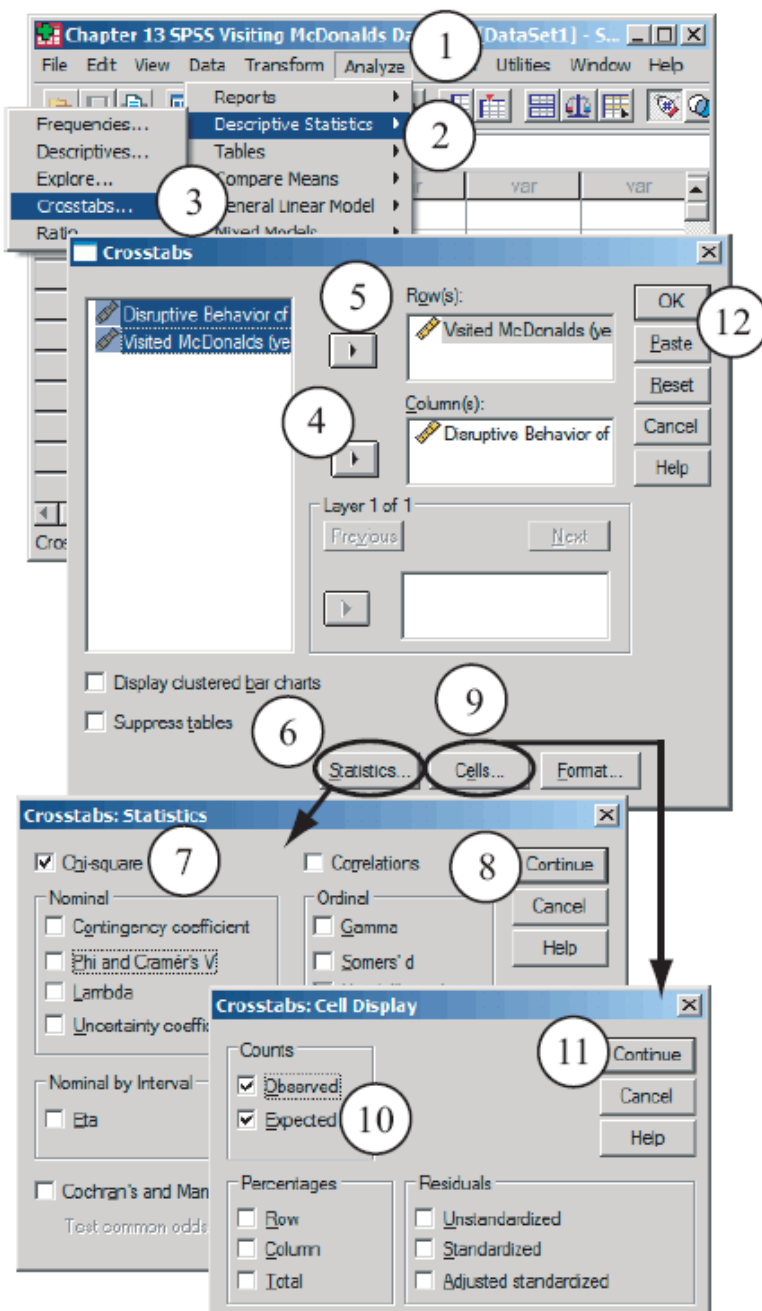
Pearson's Chi-Square

Like the goodness-of-fit chi-square, the Pearson's Chi-square tells us whether the differences found between group frequencies is likely due to chance alone. That is, it tells us whether the pattern of frequencies is a random pattern or not. The Pearson's chi-square differs from the goodness-of-fit in that the former evaluates the pattern of two variables at the same time. Because the Pearson's chi-square evaluates two variables, a significant chi-square not only tells us that the pattern of frequencies is significantly different from a random pattern, but it also tells us that the two variables are associated with one another.

Running Pearson's Chi-Square

Goodness-of-Fit Chi-Square (See Figure 13.6): From the **Analyze** (1) pull down menu select **Descriptive Statistics** (2), then select **Crosstabs...** (3) from the side menu. In the **Crosstabs** dialogue box, enter the variable **behavior** in the **Column(s):** field by left-clicking the **behavior** variable and then clicking on the boxed arrow (4) pointing to the **Column(s):** field. Next, enter the variable **mcdonlds** in the **Row(s):** field by left-clicking on the variable and left-clicking on the boxed arrow (5) pointing to the **Row(s):** field.

Figure 13.6 SPSS: Running Pearson's Chi-Square



[We should note that the decision regarding which variable goes in the rows and which goes in the column is rather arbitrary.

However, in cases where one variable has more groups than the other (e.g., a 2 x 3 design), it is preferable to put the variable with the fewest groups in the columns as it makes the printout of the output easier to read (unless you use the landscape print option then the opposite is true).] Next, to

select the Pearson's chi-square statistic left-click the **Statistics...** button (6). In the **Crosstabs: Statistics** dialogue box check the **chi-square** option (7) and then click **Continue** (8) to return to the **Crosstabs** dialogue box. Next, to select the type of information that will be displayed for each cell of the chi-square matrix in the output click the **Cells...** button (9). In the **Crosstabs: Cell Display** dialogue box check the **Observed** (already selected for you) and **Expected** (10) options under the **Counts** options and then click **Continue** (11) to return to the **Crosstabs** dialogue box. Finally, double check your variables and options and either select **OK** (12) to run, or **Paste** to create syntax to run at a later time.

If you selected the paste option from the procedure above, you should have generated the following syntax:

```
CROSSTABS
```

```
  /TABLES=mcdonlds BY behavior
```

```
  /FORMAT= AVALUE TABLES
```

```
  /STATISTIC=CHISQ
```

```
  /CELLS= COUNT EXPECTED .
```

To run the analyses using the syntax, while in the Syntax Editor, select **All** from the **Run** pull-down menu.

Reading the Pearson's Chi-Square Crosstabs Output

The Crosstabs Chi-Square Output is presented in Figure 13.7. This output consists of three major parts: Case Processing Summary, the Crosstabulation matrix, and Chi-Square Tests. The case processing summary output reports the number (N) and percentage for the valid cases (cases for which there is no missing data), missing cases (cases for which there is not reported data), and the total sample (regardless of their valid or missing status). In this case we are not missing any participant data for either of our variables so the valid and total numbers are the same, while the missing cases have N's and percentages of 0.

The crosstabulation matrix output presents tabulations (counts) for each group of one variable separated **across** the groups of the second variable. Thus the term, "cross"-tabulation". Here, the disruptive

Figure 13.7 SPSS: Pearson's Chi-Square Output

Crosstabs[DataSet1] C:\Docs\Chapter 13\miscellaneous\
Chapter 13 SPSS Visiting MacDonalds Data.sav**Case Processing Summary**

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Visited MacDonalds (yes/no) * Disruptive Behavior of Adolescent Bovines (yes/no)	100	100.0%	0	.0%	100	100.0%

**Visited MacDonalds (yes/no) * Disruptive Behavior of Adolescent Bovines (yes/no)
Crosstabulation**

			Disruptive Behavior of Adolescent Bovines (yes/no)		Total
			no	yes	
Visited MacDonalds (yes/no)	no	Count	40	7	47
		Expected Count	25.9	21.2	47.0
	yes	Count	15	38	53
		Expected Count	29.2	23.9	53.0
Total		Count	55	45	100
		Expected Count	55.0	45.0	100.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	32.476 ^b	1	.000		
Continuity Correction ^a	30.222	1	.000		
Likelihood Ratio	34.914	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	32.151	1	.000		
N of Valid Cases	100				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 21.

behavior groups are presented in the columns and the visit to MacDonalds groups are presented in the rows.

Within each cell the observed values are listed first (labeled as "Count" in the row heading) and the expected values are listed below the observed (labeled as "Expected Count" in the row heading). The row and column totals are also presented.

In the chi-square tests output we are given an almost overwhelming number of statistics to interpret. For simplicity's sake we will only concern

ourselves with the first row of statistics labeled Pearson Chi-Square. In this row we are given the chi-square obtained, the degrees of freedom [$df = (R-1)(C-1)$], and the exact level of significance. In this case we have a chi-square of 32.476, with 1 degree of freedom [$df = (2-1)(2-1) = 1$], which is significant at least at the .001 alpha level. Thus we can conclude that there is a significant relationship between bovine adolescents misbehaving and being taken to MacDonalds. By eyeballing the observed frequencies in the crosstabulation matrix, it appears that adolescents who misbehave tend to get turned into hamburger and those who do not misbehave do not get turned into hamburger.