Creating and Analyzing Contingency Tables

Defining Contingency Tables  404
Summarizing Raw Data in Tables  405
  Understanding the Mosaic Plot  407
  Understanding the Contingency Table Report  409
  Enhancing the Report  410
Creating a JMP Contingency Table from an Existing Summary Table  411
Creating Contingency Tables for Several Variables  414
  Revising Colors in Mosaic Plots  417
Performing Tests for Independence  419
  Understanding Chi Square Test Results  422
  Understanding Fisher’s Exact Test Results  423
  Enhancing the Tests Report  423
Measures of Association with Ordinal Variables  424
  Understanding the Plot  427
  Understanding the Report  428
Summaries  429
  Key Ideas  429
  JMP Steps  430
Exercises  432
Special Topic: Statistical Summary Tables  434
Do Democrats and Republicans have the same responses to a survey question asking about campaign reforms? Do rural and urban residents own different types of vehicles (sports cars, sedans, SUVs, and trucks)? Do graduate and undergraduate students differ in whether they share an apartment, have their own apartment, or live in a dorm?

These questions involve looking at two classification variables, and testing if the variables are related. Neither variable contains quantitative measurements; instead, both variables identify the respondent as belonging to a group. For example, one respondent might be “rural, truck”, and another respondent might be “urban, sedan”. The two variables can be character or numeric, and are either nominal or ordinal. This chapter discusses:

- summarizing classification data using tables and plots
- testing for independence between the classification variables
- using measures of association between classification variables

The methods in this chapter are appropriate for nominal and ordinal variables.

---

**Defining Contingency Tables**

Suppose you have nominal or ordinal variables that classify the data into groups. You want to summarize the data in a table. To make discussing summary tables easier, this section introduces some notation.

Tables that summarize two or more classification variables are called **contingency tables**. These tables are also called **crosstabulations**, **summary tables**, or **pivot tables** (in Microsoft Excel). Tables that summarize two variables are called **two-way tables**, tables that summarize three variables are called **three-way tables**, and so on. A special case of the two-way table occurs when both variables have only two levels. This special case is called a **$2 \times 2$ table**. Although this chapter shows how to create contingency tables involving several variables, the analyses are appropriate only for two-way tables. Figure 12.1 shows the parts of a contingency table.
Chapter 12: Creating and Analyzing Contingency Tables

Figure 12.1 Parts of a Contingency Table

The table consists of rows and columns. Figure 12.1 contains $r$ rows and $c$ columns, and is an $r \times c$ table. The rows and columns form cells. Each cell of a table is uniquely indexed by its row and column. For example, the cell in the second row and first column is cell$_{21}$. A contingency table usually shows the number of observations in each cell, or the cell frequency. The total number of observations in the table is $n$. The number of observations in each cell follows the same notation pattern as the cells. For example, the number of observations in cell$_{21}$ is $n_{21}$.

The contingency table in Figure 12.1 is a two-way table because it summarizes two variables. The phrase “two-way” does not refer to the number of rows or columns; it refers to the number of variables that are included in the table.

Sometimes, you have the raw data and you want to summarize the data in a table and analyze it. Other times, you already have a summary table and you want to analyze it. The next three sections discuss summarizing data in tables. The rest of the chapter discusses analyses for two-way tables.

---

Summarizing Raw Data in Tables

Suppose you have raw data that you want to summarize in a table. Table 12.1 shows data from an introductory statistics class. The instructor collected data on the gender of each student, and whether the student was majoring in Statistics.\(^1\)

---

\(^1\) Data is from Dr. Ramon Littell, University of Florida. Used with permission.
Table 12.1 Class Data

<table>
<thead>
<tr>
<th>Student</th>
<th>Gender</th>
<th>Major</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>Statistics</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>Other</td>
</tr>
<tr>
<td>3</td>
<td>Female</td>
<td>Statistics</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>Other</td>
</tr>
<tr>
<td>5</td>
<td>Female</td>
<td>Statistics</td>
</tr>
<tr>
<td>6</td>
<td>Female</td>
<td>Statistics</td>
</tr>
<tr>
<td>7</td>
<td>Male</td>
<td>Other</td>
</tr>
<tr>
<td>8</td>
<td>Male</td>
<td>Other</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>Statistics</td>
</tr>
<tr>
<td>10</td>
<td>Female</td>
<td>Statistics</td>
</tr>
<tr>
<td>11</td>
<td>Male</td>
<td>Other</td>
</tr>
<tr>
<td>12</td>
<td>Female</td>
<td>Statistics</td>
</tr>
<tr>
<td>13</td>
<td>Male</td>
<td>Statistics</td>
</tr>
<tr>
<td>14</td>
<td>Male</td>
<td>Statistics</td>
</tr>
<tr>
<td>15</td>
<td>Male</td>
<td>Other</td>
</tr>
<tr>
<td>16</td>
<td>Female</td>
<td>Statistics</td>
</tr>
<tr>
<td>17</td>
<td>Male</td>
<td>Statistics</td>
</tr>
<tr>
<td>18</td>
<td>Male</td>
<td>Other</td>
</tr>
<tr>
<td>19</td>
<td>Female</td>
<td>Other</td>
</tr>
<tr>
<td>20</td>
<td>Male</td>
<td>Statistics</td>
</tr>
</tbody>
</table>

This data is available in the Stat Majors data table in the sample data for the book.

To summarize the data table in JMP:

1. Open the Stat Majors data table.
2. In the JMP Starter window, click Basic→Contingency.
3. Click Major→Y, Response Category. This variable forms the columns of the summary table.
4. Click Gender→X, Grouping Category. This variable forms the rows of the summary table.
5. Click OK.
Figure 12.2 shows the results, with the Tests report hidden. (“Performing Tests for Independence” in this chapter discusses this hidden report.) JMP displays both a graph and a text report. The title of the report identifies the column variable (Major) and the row variable (Gender). The next two topics discuss the graphs and reports in Figure 12.2.

**Understanding the Mosaic Plot**

Figure 12.2 shows a two-way mosaic plot. JMP documentation calls this a contingency analysis mosaic plot or a mosaic plot.

The left axis of the plot identifies proportions. The right axis shows color coding for the column variable (Major). The right axis also shows the relative proportions for the column variable, combined across the levels of the row variable. In Figure 12.2, the bar at the right axis shows the relative proportions of Statistics and Other majors, combined across the Male and Female students.

For variables with two levels, JMP uses red and blue color coding. For variables with more levels, JMP selects a color scheme.

JMP uses the data to scale the width of the columns in the mosaic plot. For the Stat Majors data, 8 of the 20 students are female, and JMP scales the Female column accordingly.

Look at the left side of the plot. The left side shows that over 75% of the Female students are majoring in Statistics (the blue rectangle). The right side of the two-way mosaic plot shows that half of the Male students are majoring in Statistics. This difference gives a visual hint that the pattern of majoring in Statistics or Other differs for Male and Female students. Later, “Performing Tests for Independence” shows how to check this visual hint to see whether it leads to a statistical conclusion.

The two-way mosaic plot provides a visual summary of the two variables. If the pattern for the row variable is the same across the values of the column variable, then the color patterns—and the relative proportion for each color—is the same from left to right across the mosaic plot. If the pattern differs, then the relative proportion for each color differs also.
Figure 12.2 Contingency Results for Stat Majors Data

Contingency Analysis of Major by Gender

Mosaic Plot

Contingency Table

<table>
<thead>
<tr>
<th>Gender</th>
<th>Major</th>
<th>Count</th>
<th>Total %</th>
<th>Col %</th>
<th>Row %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>Other</td>
<td>1</td>
<td>5.00</td>
<td>30.00</td>
<td>50.00</td>
</tr>
<tr>
<td></td>
<td>Statistics</td>
<td>6</td>
<td>35.00</td>
<td>55.00</td>
<td>85.71</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>7</td>
<td>35.00</td>
<td>30.00</td>
<td>65.00</td>
</tr>
<tr>
<td></td>
<td>Statistics</td>
<td>6</td>
<td>30.00</td>
<td>50.00</td>
<td>87.50</td>
</tr>
<tr>
<td>Male</td>
<td>Other</td>
<td>8</td>
<td>40.00</td>
<td>50.00</td>
<td>53.85</td>
</tr>
<tr>
<td></td>
<td>Statistics</td>
<td>12</td>
<td></td>
<td></td>
<td>87.50</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>20</td>
<td></td>
<td></td>
<td>53.85</td>
</tr>
</tbody>
</table>
Understanding the Contingency Table Report

The Contingency Table report summarizes the data table. The top row shows the values for the column variable, and the leftmost column shows the values for the row variable.

The outside edges of the table give row totals and column totals. Looking at the row totals, 7 students are Female, and 13 are Male. Looking at the column totals, 12 students are majoring in Statistics, and 8 are majoring in something else (Other). The outside edges also give the row percentages and column percentages. Looking at the row percentages, 35% (calculated from 7/20) of the students are Female. Looking at the column percentages, 60% (calculated from 12/20) of the students are majoring in Statistics.

The lower-right corner cell gives the overall total for the table, which is 20. Because there are no missing values, this number matches the number of observations in the data table. However, if the data table contained the Gender and not the Major for a student, the summary table would contain only 19 observations.

The top-left corner cell gives a key to understanding the main body of the table. The list gives details:

- **Count**: Number of observations in each cell. This class has 1 female student who is majoring in Other.
- **Total %**: Percentage of the total number of observations represented by the cell count. The single Female student majoring in Other represents 5% of the class. The class has 20 students, so 1/20=5%.
- **Col %**: Percentage of observations in the column represented by the cell frequency. The single Female Other student represents 12.5% of the Other majors in the class. The class has 8 students majoring in Other, so 1/8=12.5%.
- **Row %**: Percentage of observations in the row represented by the cell frequency. The single Female Other student represents 14.29% of the females in the class. The class has 7 female students, so 1/7=14.29%.

The Col % values sum to 100% for each column.

The Row % values sum to 100% for each row.
Enhancing the Report

The Contingency Table report has a hot spot that provides options for hiding the summary statistics that JMP automatically displays, and for adding more statistics. In JMP, click the hot spot for Contingency Table. Select a checked item to hide it. Select an unchecked item to display it. JMP uses the three unchecked items (Expected, Deviation, and Cell Chi Square) when performing statistical tests.

The Contingency Table report differs from most other reports in JMP. You cannot double-click on the cells in the summary table to change their appearance. JMP displays percentages to two decimal places.

The Contingency Analysis of Major By Gender report title (at the top of the window) has a hot spot. It provides options for closing the Mosaic Plot, the Contingency Table report, and the Tests report that JMP automatically displays. This hot spot also has options for adding more analyses. See the JMP documentation for more details.

The Contingency Analysis of Major By Gender hot spot provides a feature for changing the display. Suppose you want to rotate the mosaic plot so that the row variable displays as rows in the plot. Some people find this easier to view, because the rows of the mosaic plot are the same as the rows in the Contingency Table report. In JMP:

1. Click the hot spot for the Contingency Analysis of Major By Gender report.
2. Select Display Options → Horizontal Mosaic.

Figure 12.3 shows the results.
Creating a JMP Contingency Table from an Existing Summary Table

Sometimes, you already have a summary table of the data. You can create a JMP data table that matches the summary table, and then analyze it. It is not necessary to expand the summary table into a data table that has one row for every observation in the summary table. Table 12.2 shows frequency counts for several court cases. The columns show the defendant’s race, and the rows identify whether the death penalty was imposed after the defendant was convicted of homicide.²

Table 12.2  Death Penalty Data

<table>
<thead>
<tr>
<th>Decision</th>
<th>Black</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>149</td>
<td>141</td>
</tr>
<tr>
<td>Yes</td>
<td>17</td>
<td>19</td>
</tr>
</tbody>
</table>

This data is available in the Penalty data table in the sample data for the book.

To create the data table in JMP, follow the steps in Chapter 3. Your data table will have four rows, one for each cell in the table. It will have three columns, one for decision, one for race, and one for count.

The data table itself summarizes the data because it displays a spreadsheet-like view. However, the results from Contingency provide more features, including the two-way mosaic plot, percentages, and statistical tests. In JMP:

1. Open the Penalty data table.
2. In the JMP Starter window, click BasicÆContingency.
3. Click Defendant RaceÆY, Response Category.
4. Click DecisionÆX, Grouping Category.
5. Click CountÆFreq.
6. Click OK.

As with the raw data table, the y variable forms the columns of the summary table, and the x variable forms the rows of the summary table. The Freq variable tells JMP the frequency count for each cell. Without a Freq variable, JMP shows a single observation for each cell.

Figure 12.4 shows the results, with the Tests report hidden.
Figure 12.4 Contingency Results with a Freq Variable

Contingency Analysis of Defendant Race by Decision

Mosaic Plot

<table>
<thead>
<tr>
<th>Decision</th>
<th>Defendant Race</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>White</td>
</tr>
<tr>
<td></td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>43.25</td>
</tr>
<tr>
<td></td>
<td>88.96</td>
</tr>
<tr>
<td>Yes</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>5.83</td>
</tr>
<tr>
<td></td>
<td>36</td>
</tr>
</tbody>
</table>

Contingency Table

<table>
<thead>
<tr>
<th>Decision</th>
<th>Defendant Race</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Black</td>
</tr>
<tr>
<td>No</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>45.71</td>
</tr>
<tr>
<td></td>
<td>88.96</td>
</tr>
<tr>
<td>Yes</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>5.21</td>
</tr>
<tr>
<td></td>
<td>36</td>
</tr>
</tbody>
</table>

Tests
JMP produces the same plots and reports with a `Freq` variable as it does when creating a summary table from raw data. Look at Figure 12.4, between the Mosaic Plot and the Contingency Table report, and you see the only difference. JMP identifies the `Freq` variable with the text `Freq: Count`.

Compare the mosaic plots in Figures 12.2 and 12.4. For the `Penalty` data, the proportion of red and blue blocks (black and white defendants) is roughly the same for both the `Yes` and `No` decisions. Figure 12.4 might lead you to initially conclude that the two variables are not related. “Performing Tests for Independence” uses a statistical test to check this initial conclusion. Figure 12.4 scales the width of the columns according to the values of the `X, Grouping Category` variable. The `Yes` column is much narrower than the `No` column, reflecting the small percentage of `Yes` decisions.

When using a `Freq` variable, you can change the statistics displayed in the Contingency Table report just as you can for summary tables that are created from raw data.

---

**Creating Contingency Tables for Several Variables**

To create contingency tables for several variables, start with the approach for creating contingency tables for two variables. Select a row variable and a column variable, and add `By` variables to create multiple contingency tables. Chapter 8 shows how to use a `By` variable in Distribution. The same approach works in Contingency. For an example, use the `Cars 1993` data. In JMP:

1. Click Help→Sample Data Directory.
2. Click the disclosure diamond for Exploratory Modeling.
3. Select **Cars 1993** and JMP opens the data table.

4. Click the X in the upper-right corner of the Sample Data Directory window to close the window.

5. In the JMP Starter window, click **Basic → Contingency**.

6. Click **Passenger Capacity → Y, Response Category**.

7. Click **Vehicle Category → X, Grouping Category**.

8. Click **Domestic Manufacturer → By**.

9. Click **OK**.

   JMP displays a warning message, informing you that **Passenger Capacity** is not a nominal or ordinal variable. The message asks you to consider canceling the action. In this case, **Passenger Capacity** is defined as a continuous variable in the data table, but it is actually an ordinal variable. As a result, the warning message can be ignored. In contrast, the multiple **Cost** and **Mileage** variables are correctly defined as continuous variables, and using these variables in a summary table is inappropriate.

10. Click **Continue**.

    JMP creates a summary table for each level (0 and 1) of **Domestic Manufacturer**. This variable follows a coding convention where 0 means ‘no’ and 1 means ‘yes’. When a row in the data table has **Domestic Manufacturer=0**, that row corresponds to a foreign car. Figure 12.5 displays results, with the **Contingency Table** and **Tests** reports hidden.
Figure 12.5 Contingency Results Using a By Variable

Contingency Analysis of Passenger Capacity By Vehicle Category Domestic Manufacturer=0

Mosaic Plot

Contingency Analysis of Passenger Capacity By Vehicle Category Domestic Manufacturer=1

Mosaic Plot
JMP displays the summaries for the two levels of Domestic Manufacturer in separate reports. Figure 12.5 displays the mosaic plots for each level of Domestic Manufacturer. The mosaic plots show that the domestic cars (1) differ from foreign cars (0). As one example of a difference, there are no Large foreign cars.

JMP uses multiple colors in the mosaic plot because the variables have more than two levels. Using a By variable has no effect on the colors in the mosaic plot. See “Revising Colors in Mosaic Plots” for a way to enhance the mosaic plot.

You can enhance reports with a By variable using the same JMP features discussed earlier.

At the end of this chapter, see “Special Topic: Statistical Summary Tables” for a way to summarize several variables in a single table. See “Further Reading” at the end of the book or refer to JMP documentation for references on analyses involving more than two classification variables.

**JMP Hint:**

As you become familiar with JMP, you might want to select Analyze ➔ Fit Y by X instead of selecting Contingency from the JMP Starter window. Both choices launch the same JMP platform.

One advantage of Contingency is that JMP provides a warning window when you select a continuous variable. With Fit Y by X, JMP assumes that you want to perform a different analysis and does not create the contingency table.

---

**Revising Colors in Mosaic Plots**

JMP uses a red and blue color scheme for variables with two levels in mosaic plots. For variables with multiple levels, JMP uses multiple colors. When printing these mosaic plots on a gray scale printer, the colors can be difficult to distinguish. You can revise the colors to use shades of gray. In JMP:

1. Right-click on the mosaic plot and select Set Colors. If the Set Colors option does not display, carefully place your mouse pointer so that it is on the mosaic plot.
2. JMP shows the old and new colors in the Select Colors for Values window. When the window first appears, the two colors are the same because you have not changed the colors yet, as shown in the following window.

![JMP: Select Colors for Values](image)

3. Click the right colored oval for 2 and select the lightest shade of gray from the color palette, as shown in the following window.

![JMP: Select Color](image)

4. Click the right colored oval for 7 and select black from the color palette.

5. Click the Macros button and select Gradient between ends, as shown in the following window.
6. When you are finished changing colors, click **OK**.

JMP updates the mosaic plot to use the new colors. When changing colors and using a **By** variable, JMP changes the colors for one mosaic plot at a time. To change the colors for all of the mosaic plots, change them for each level of the **By** variable.

**JMP Hints:**

If you want to return to the automatic color choices, click the **Macros** button and select **Revert to old colors**.

If you find that selecting the colored oval is difficult, select the row and select the color button at the far-left side of the window. This displays a color palette and you can change colors from it.

---

### Performing Tests for Independence

When you collect classification data, you want to know whether the variables are related in some way. For the **Penalty** data, is the defendant’s race related to the verdict? Does knowing the defendant’s race tell you anything about the likelihood that the defendant will receive the death penalty?
In statistical terms, the null hypothesis is that the row and column variables are independent. The alternative hypothesis is that the row and column variables are not independent. To test for independence, you compare the observed cell frequencies with the cell frequencies that would occur in the situation where the null hypothesis is true. (See “Technical Details: Expected Cell Frequencies” at the end of this section.)

One commonly used test is a Chi Square test, which tests the hypothesis of independence. A test statistic is calculated and compared with a critical value from a Chi-Square distribution. Suppose you want to test the hypothesis of independence at the 10% significance level for the Penalty data. (For more information, see the general discussion about building hypothesis tests in Chapter 6.)

The steps for analysis are the same as the steps for comparing groups:

1. Create a JMP data table.
2. Check the data table for errors.
3. Choose the significance level for the test.
4. Check the assumptions for the test.
5. Perform the test.
6. Make conclusions from the test results.

To check the data table for errors, use Distribution and the tools from Chapter 4.

To test for independence with two classification variables, the assumptions are:

- Data are counts. Practically, this requires that the variables are nominal or ordinal. JMP performs these tests in the Contingency platform for nominal and ordinal variables.

- Observations are independent. The values for one observation are not related to the values for another observation. To check the assumption of independent observations, think about your data and whether this assumption is reasonable. This is not a step where using JMP will answer this question.
Observations are a random sample from the population. You want to make conclusions about a larger population, not about just the sample. For the statistics class, you want to make conclusions about statistics classes in general, not just this single class. To check this assumption, think about your data and whether this assumption is reasonable. This is not a step where using JMP will answer this question.

Sample size is large enough for the tests. As a general rule, the sample size should be large enough to expect five responses in each cell of the summary table. JMP prints warning messages when this assumption is not met.

JMP automatically performs appropriate tests when creating results from Contingency. Figure 12.6 shows the Tests report for the Penalty data. Figure 12.4 shows the other results, with the Tests report hidden.

**Figure 12.6  Testing for Independence with the Penalty Data**

<table>
<thead>
<tr>
<th>Tests</th>
<th>Source</th>
<th>DF</th>
<th>-LogLike</th>
<th>RSquare (U)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>1</td>
<td>0.11073</td>
<td>0.0005</td>
</tr>
<tr>
<td>Error</td>
<td>324</td>
<td></td>
<td>225.80064</td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>325</td>
<td></td>
<td>225.91076</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>325</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>ChiSquare</td>
<td>Prob&gt;ChiSq</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>0.221</td>
<td>0.6379</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson</td>
<td>0.221</td>
<td>0.6379</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fisher's Exact Test</td>
<td>Prob</td>
<td>Alternative Hypothesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left</td>
<td>0.7412</td>
<td>Prob(Defendant Race=White) is greater for Decision=No than Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td>0.3643</td>
<td>Prob(Defendant Race=White) is greater for Decision=Yes than No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Tail</td>
<td>0.7246</td>
<td>Prob(Defendant Race=White) is different across Decision</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first report in Figure 12.6 (with Source as the first column heading) is similar to an analysis of variance table for a continuous response variable. See the JMP documentation for more details. The test for independence does not use this report.
Before making decisions about the test, think about the assumptions. The data are counts, so the first assumption is reasonable. The observations are independent, because the race of and decision for a defendant is unrelated to the race of and decision for another defendant. The second assumption is reasonable. The observations are a random sample from the population of defendants convicted of homicide, so the third assumption is reasonable. In the Contingency Table report, use the hot spot to display the Expected values, and confirm that the fourth assumption is reasonable. Also, when this assumption is not met, JMP prints a warning message. Now, look at the results from the test for independence.

**Understanding Chi Square Test Results**

In the mosaic plot, the proportion of white and black defendants was roughly the same for the Yes and No decisions. The mosaic plot led to an initial conclusion that the variables were unrelated. The Chi Square test for independence quantifies the initial conclusion.

Look under the heading Test in Figure 12.6. JMP displays two Chi Square tests. The Pearson and Likelihood Ratio tests both have the same assumptions. The Pearson test uses the observed and expected cell frequencies, and the Likelihood Ratio test uses a more complex formula.

Look at the number under Prob>ChiSq. This value is 0.6379, which is greater than the significance level of 0.10. You conclude that there is not enough evidence to reject the null hypothesis of independence between the defendant’s race and the decision. (Refer to Agresti in “Further Reading” for an additional analysis of this data that considers the race of the victim.) Although the p-values for the two tests are identical in Figure 12.6, this is not always true. Typically, the p-values are similar, but are not identical.

In general, to interpret JMP results, look at the p-value under Prob>ChiSq, in the line for Pearson. If the p-value is less than the significance level, reject the null hypothesis that the two variables are independent. If the p-value is greater, you fail to reject the null hypothesis.
Understanding Fisher’s Exact Test Results

Fisher’s exact test was developed for the special case of a 2×2 table. This test is very useful when the assumptions for a Chi Square test are not reasonable, and is especially useful for tables with small cell frequencies. JMP automatically performs this test for 2×2 tables, but it isn’t available for larger tables.

Look under the heading Fisher’s Exact Test in Figure 12.6. JMP displays results for both one-sided tests and the two-sided test. Look at the line labeled 2-Tail. The p-value is 0.7246, so you fail to reject the null hypothesis that the two variables are independent.

In general, to interpret JMP results, look at the 2-Tail p-value, which tests for independence between the two variables. The one-sided Left and Right p-values might be useful in specific situations. Interpret the p-value the same way you do for the Chi Square test.

Enhancing the Tests Report

You can double-click on a column in the Tests report and change the format of the numbers that are displayed. As with other reports in JMP, you can change the report titles.

Position the mouse pointer near a p-value and move it around in a very small circle. JMP displays a pop-up window, which includes text explaining the p-value. To close the window, move your mouse pointer again. JMP highlights p-values less than 0.05 with an asterisk.

To hide the Tests report, click the disclosure diamond next to the report title.

To remove the Tests report, click the hot spot at the top of the Contingency report. Click the Tests selection so that it is deselected, and JMP removes the report and its title.
**Technical Details: Expected Cell Frequencies**

The Chi Square test compares the observed cell frequencies with the expected cell frequencies, assuming the null hypothesis that the variables are independent. Click the hot spot for the Contingency Table report to display the Expected frequencies. To calculate the expected cell frequencies, multiply the row total and column total, and divide by the total number of observations. For the Black-No cell, the formula is:

\[
\frac{[(\text{row total for No}) \times (\text{column total for Black})]}{\text{total } N}
\]

\[
= \frac{(290 \times 166)}{326}
\]

\[
= 147.67
\]

The Chi Square test is always valid if there are no empty cells (no cells with a cell frequency of 0), and if the expected cell frequency for all cells is 5 or greater.

Because these conditions are true for the Penalty data, the Chi Square test is a valid test. If these two conditions are not true, JMP prints a message warning that the Chi Square test might not be valid.

There is some disagreement among statisticians about exactly when the test should not be used, and what to do when the test is not valid. Two practical recommendations are to collect more data, or to combine low-frequency categories.

Combine low-frequency categories only when it makes sense. For example, consider a survey that asked people how often they called a Help Desk in the past month. Suppose the categories are: 0 (no calls), 1-2, 3-5, 6-8, 9-11, 12-15, 16-20, and “over 20”. Now, suppose the expected cell counts for the last four categories are less than 5. It makes sense to combine these last four categories into a new category of “9 or more”.

---

**Measures of Association with Ordinal Variables**

When you reject the null hypothesis for either the Chi Square test or Fisher’s exact test, you conclude that the two variables are not independent. But, you do not know how the two variables are related. When both variables are ordinal, measures of association give more insight into your data. As defined in Chapter 3, ordinal variables have values that provide names and an implied order.
Two measures of association are Kendall’s \textit{tau} and Spearman’s rank correlation coefficient. Both measures range from -1.0 to 1.0. Values close to 1.0 indicate a positive association, and values close to -1.0 indicate a negative association. (The Spearman’s rank correlation coefficient is similar to correlation coefficients discussed in Chapter 10. This correlation coefficient is essentially the Pearson correlation, applied to ranks instead of actual values.)

An animal epidemiologist tested dairy cows for the presence of a bacterial disease\(^3\). The disease is detected by analyzing blood samples, and the disease severity for each animal was classified as none (0), low (1), or high (2). The size of the herd that each cow belonged to was classified as small (10), medium (100), or large (1000). Table 12.3 shows the number of cows in each herd size, and in each disease severity category. Both the herd size and disease severity variables are ordinal. The disease is transmitted from cow to cow by bacteria, so the epidemiologist wanted to know whether disease severity is affected by herd size. As the herd size gets larger, is there either an increasing or a decreasing trend in disease severity?

The epidemiologist tested for independence using the Chi Square test. However, she still does not know whether there is a trend in disease severity that is related to an increase in herd size. Kendall’s \textit{tau} or Spearman’s rank correlation coefficient can answer this question.

\begin{table}[h]
\centering
\caption{Cow Disease Data}
\begin{tabular}{|l|c|c|c|}
\hline
Herd Size & None (0) & Low (1) & High (2) \\
\hline
Small (10) & 9 & 5 & 9 \\
\hline
Medium (100) & 18 & 4 & 19 \\
\hline
Large (1000) & 11 & 88 & 136 \\
\hline
\end{tabular}
\end{table}

This data is available in the \textit{Cow Disease} data table in the sample data for the book. The icons for \textit{Herd Size} and \textit{Disease} indicate that these two variables are ordinal, as shown in the following window. The variables are also numeric, which JMP requires for the \texttt{Multivariate} platform.

\footnote{\text{Data is from Dr. Ramon Littell, University of Florida. Used with permission.}}
To create the data table in JMP, follow the steps in Chapter 3. Your data table will have 9 rows, one for each cell in the table. It will have 3 columns, one for herd size, one for disease severity, and one for number.

JMP requires numeric variables to create Kendall’s tau and Spearman’s rank correlation coefficient. JMP provides these statistics in the Multivariate platform. This platform automatically creates a scatterplot matrix that is useful for investigating multiple continuous variables. To create the measures of association in JMP:

1. Open the Cow Disease data table.
2. In the JMP Starter window, click Multivariate→Multivariate.
3. Click Herd Size and Disease and then Y, Columns.
4. Click Number→Freq.
5. Click OK.

JMP displays a warning message, informing you that Herd Size and Disease are not continuous variables. The message asks you to consider canceling the action. In this case, both variables are ordinal. As a result, the warning message can be ignored. Using these two variables to create the measures of association is appropriate.

6. Click Continue. JMP creates the results.
7. Hold down the ALT key and click the hot spot for the Multivariate report.
8. Select Correlations Multivariate to close this report. It makes sense for continuous variables.
9. Select **Scatterplot Matrix** to close the plot. It makes sense for continuous variables.

10. Select **Nonparametric Correlations** ➔ **Spearman’s ρ**. (The ρ is the Greek letter rho, which is how statisticians often refer to Spearman’s rank correlation coefficient.)

11. Select **Nonparametric Correlations** ➔ **Kendall’s τ**. (The τ is the Greek letter tau.)

12. Click **OK**.

Figure 12.7 shows the results.

**Figure 12.7 Measures of Association for the Cow Disease Data**

| Variable by Variable | Spearman’s ρ | Prob>|ρ| |
|----------------------|-------------|-------|
| Disease, Herd Size   | 0.2331      | <.0001* |

| Variable by Variable | Kendall’s τ | Prob>|τ| |
|----------------------|-------------|-------|
| Disease, Herd Size   | 0.2173      | <.0001* |

Figure 12.7 highlights the **Freq** variable by showing **Freq: Number** just below the **Multivariate** report title. JMP performs exactly the same analyses with raw data. The only difference in the results is that the report does not show a **Freq** variable. The plots and reports are similar to the **Pairwise Correlations** results, discussed in Chapter 10.

**Understanding the Plot**

For both measures of association, JMP creates a bar chart scaled from -1 to 1. JMP colors from the value 0 to the bar segment that matches the value of the measure of association. Both Spearman’s rank correlation coefficient and Kendall’s tau have a value of about 0.2, so the same area is colored for each plot.
Understanding the Report

For each measure of association, JMP lists the two variables involved, the measure, and a p-value.

A positive measure of association indicates an increasing trend between the two variables. As the ordinal levels of one variable increase, so do the ordinal levels of the other variable. For the Cow Disease data, the epidemiologist can conclude that disease severity increases with increasing herd size.

A negative measure of association indicates a decreasing trend between the two variables. As the ordinal levels of one variable increase, the ordinal levels of the other variable decrease.

The closer a measure of association is to 0, the weaker the strength of the relationship. A value of -0.8 or 0.8 indicates a stronger relationship than a value of -0.1 or 0.1.

The p-value is the result of a test that the measure of association is significantly different from 0. For this test, the null hypothesis is that the measure of association is 0. The alternative hypothesis is that the measure of association is different from 0. JMP highlights p-values less than 0.05 with an asterisk.

For the Cow Disease data, you conclude that the measures of association are significantly different from 0.

In general, if the p-value is less than the significance level, reject the null hypothesis and conclude that the measure of association is significantly different from 0. If the p-value is greater than the significance level, fail to reject the null hypothesis. Conclude that there is not enough evidence to say that the measure of association is significantly different from 0.

JMP Hint:

As you become familiar with JMP, you might want to select Analyze ➔ Multivariate Methods ➔ Multivariate instead of selecting Multivariate ➔ Multivariate from the JMP Starter window. Both choices launch the same JMP platform.

One advantage of selecting Multivariate in the JMP Starter window is that JMP provides a warning window for an ordinal variable. Click Continue and create reports with measures of association for ordinal variables. When you select Analyze from the menu, JMP requires continuous variables in the Multivariate platform.
Chapter 12: Creating and Analyzing Contingency Tables

Summaries

Key Ideas

- Contingency tables are tables that summarize two or more classification variables. The rows and columns of a contingency table form cells, and the number of observations in a cell is the cell frequency for that cell.

- The Basic Contingency choice in the JMP Starter window creates contingency tables and a two-way mosaic plot, and performs statistical tests on the two variables. Use nominal or ordinal variables in the Contingency platform. As you become familiar with JMP, you might want to select Analyze Fit Y by X from the menu instead of selecting Contingency in the JMP Starter window.

- Use By variables to summarize more than two variables. JMP creates separate reports for each level of the By variable. With multiple By variables, JMP creates a separate report for each combination of levels of the multiple By variables.

- Both the Chi Square test and Fisher’s exact test are used to test for independence between two classification variables. Generally, the Chi Square test should not be used if there are empty cells or if cells have expected cell frequencies of less than 5. One option is to collect more data, which increases the expected cell frequency. Another option is to combine levels that have only a few observations.

- Regardless of the test, the steps for analysis are:
  1. Create a JMP data table.
  2. Check the data table for errors.
  3. Choose the significance level for the test.
  4. Check the assumptions for the test.
  5. Perform the test.
  6. Make conclusions from the test results.
Regardless of the test, to make conclusions, compare the \( p \)-value for the test with your significance level.

- If the \( p \)-value is less than the significance level, reject the null hypothesis that the two variables are independent. JMP displays an asterisk next to \( p \)-values that are less than 0.05.
- If the \( p \)-value is greater than the significance level, fail to reject the null hypothesis.

Kendall’s \( \tau \) and Spearman’s rank correlation coefficient are both measures of association that provide information about how strongly related ordinal variables are. Use these statistics to decide whether there is an increasing or a decreasing trend in the two variables, or if there is no trend at all. JMP tests whether the measure of association is significantly different from 0, and reports the \( p \)-value.

The **Multivariate** choice in the JMP Starter window creates measures of association for the two variables. Use numeric ordinal variables in the **Multivariate** platform. As you become familiar with JMP, you might want to select **Analyze**\( \rightarrow \)Multivariate Methods from the menu instead of selecting **Multivariate** in the JMP Starter window, but remember that it requires numeric continuous variables.

---

**JMP Steps**

To summarize two nominal or ordinal variables in a contingency table:

1. In the JMP Starter window, click **Basic**\( \rightarrow \)Contingency.
2. Click the column variable and then **Y, Response Category**.
3. Click the row variable and then **X, Grouping Category**.
4. If your data already form a summary table, click **Count**\( \rightarrow \)Freq.
5. Click **OK**.

To summarize three or more variables in a contingency table:

1. Complete steps 1 through 4 above.
2. Before step 5, click the additional variables and then **By**. JMP creates a contingency table for each combination of levels of the additional variables.
3. Click **OK**.
To rotate the mosaic plot:

1. Click the hot spot for the Contingency Analysis report.
2. Click Display Options → Horizontal Mosaic.

To revise colors to gray scale in the mosaic plot:

1. Right-click on the mosaic plot and select Set Colors. If the Set Colors option does not display, carefully place your mouse pointer so that it is on the mosaic plot.
2. JMP shows the old and new colors in the Select Colors for Values window. When the window first appears, the two colors are the same because you have not changed colors yet.
3. Click the right colored oval for the top color in the list and select the lightest shade of gray.
4. Click the right colored oval for the bottom color in the list and select black from the color palette.
5. Click the Macros button and select Gradient between ends.
6. Click OK.

To perform tests for independence:

Follow the steps above for creating a contingency table. JMP automatically creates the Tests report, which contains results for the Chi Square test. For 2 × 2 tables, JMP includes results from Fisher’s exact test. However, see the steps for analysis in “Key Ideas.” Although JMP automatically performs the tests, you still need to check assumptions and think about your data.

To generate measures of association:

The measures of association discussed in this chapter require two ordinal variables. JMP requires numeric variables.

1. In the JMP Starter window, click Multivariate → Multivariate.
2. Click the row variable and column variable and then Y, Columns.
3. If the data already form a summary table, click the Count → Freq.
4. Click OK.

If your variables are ordinal, JMP displays a warning message, informing you that the variables are not continuous variables. The message asks you to
consider canceling the action. When the variables are numeric ordinal variables, creating measures of association is appropriate.

5. Click **Continue**. JMP creates the results.

6. Hold down the ALT key and click the hot spot for the **Multivariate** report.

7. Select **Correlations Multivariate** to close this report. It makes sense for continuous variables.

8. Select **Scatterplot Matrix** to close the plot. It makes sense for continuous variables.

9. Select **Nonparametric Correlations→Spearman’s ρ**.

10. Select **Nonparametric Correlations→Kendall’s τ**.

11. Click **OK**.

---

### Exercises

1. Perform a test for independence for the **Stat Majors** data. Are the assumptions for the test reasonable? Define the null and alternative hypotheses. Use a 10% alpha level. Discuss your conclusions.

2. Test for independence with the **Cow Disease** data. Are the assumptions for the test reasonable? Define the null and alternative hypotheses. Use a 5% alpha level. Discuss your conclusions. Does it make a difference in the reports and graphs if you make the variables ordinal? Do the results of the statistical test change?

3. Create a new data table for the **Cow Disease** data. Use 10, 20, and 30 for the **Small, Medium**, and **Large** values of **Herd Size**. Use 0, 4, and 8 for the **None, Low**, and **High** values of **Disease Severity**. Repeat the analysis in the chapter, creating Kendall’s τ and Spearman’s rank correlation coefficient. Do the results differ? Do the values used for a variable have an effect on these statistics?

4. From the **Sample Data Directory**, click the disclosure diamond for **Categorical Models** and open the **Alcohol** data table. Create a summary table of **Relapsed** by **Alcohol Consumption**. Rotate the mosaic plot.
5. From the Sample Data Directory, click the disclosure diamond for Business and Demographic and open the Movies data table. Create a summary table of Type by Rating. Are the assumptions for the test for independence reasonable? If so, define the null and alternative hypotheses and use a 5% alpha level. Discuss your conclusions. Does it make sense to create measures of association for this data? If so, generate them and discuss your conclusions.

6. From the Sample Data Directory, click the disclosure diamond for Business and Demographic and open the Titanic data table. Create a summary table of Survived by Class. Are the assumptions for the test for independence reasonable? If so, define the null and alternative hypotheses and use a 5% alpha level. Discuss your conclusions. Does it make sense to create measures of association for this data? If so, generate them and discuss your conclusions.

7. From the Sample Data Directory, click the disclosure diamond for Exploratory Modeling and open the Cars 1993 data table. Create a summary table of Vehicle Category by Passenger Capacity. Are the assumptions for the test for independence reasonable? If so, define the null and alternative hypotheses and use a 5% alpha level. If not, explain whether it makes sense to combine categories and discuss which categories you would combine.

8. With the Cars 1993 data table from exercise 7, create a summary table of Vehicle Category by Domestic Manufacturer. Are the assumptions for the test for independence reasonable? If so, define the null and alternative hypotheses and use a 5% alpha level. Discuss your conclusions. Does it make sense to create measures of association for this data? If so, generate them and discuss your conclusions.

9. From the sample data for the book, open the Cereal Revised data table. Create a summary table of Enriched by Fiber Gr. Are the assumptions for the test for independence reasonable? If so, define the null and alternative hypotheses and use a 5% alpha level. Perform the test. Create measures of association for these variables. Discuss your conclusions.

10. From the Sample Data Directory, click the disclosure diamond for Exploratory Modeling and open the Mushroom data table. Create a summary table of Odor by Habitat for each level of Edibility. (Use Edibility as a By variable.) Are the assumptions for the test for independence reasonable? If so, define the null and alternative hypotheses and use a 5% alpha level. Discuss your conclusions, and whether results differ for edible and poisonous mushrooms.
Special Topic: Statistical Summary Tables

The Contingency platform with By variables summarizes three or more classification variables. JMP provides two other tools that summarize multiple variables in a single table. This section gives a brief introduction to the Tables > Summary platform. JMP also provides the Tables > Tabulate platform, which has a drag-and-drop interface for creating tables. The Tabulate platform has some similarities with creating pivot tables in Microsoft Excel.

This example uses the Cars 1993 data and creates a summary table. The table condenses the two tables from the Contingency platform (used with a By variable earlier in the chapter) into a single table. In JMP:

1. Click Help > Sample Data Directory.
2. Click the disclosure diamond for Exploratory Modeling.
3. Click Cars 1993.jmp. JMP opens the data table.
4. Click the X in the upper-right corner of the Sample Data Directory window to close it.
5. Click Tables > Summary.
6. Click Domestic Manufacturer > Group.
7. Click Vehicle Category > Group.
8. Click Passenger Capacity > Group. Compare your window with Figure 11S.1 to confirm your choices.
9. Click OK.
Figure 11S.1  Window for Summary Table

Figure 11S.2 shows the results.

The N Rows column gives the frequency count for the combination of levels for the three variables in that row. For example, the first row in the data table reports a single car that is from a foreign manufacturer (Domestic Manufacturer equal to 0), is Compact, and can carry 4 passengers.
**Figure 11S.2** Summary Table for Cars 1993 Data

![Summary Table for Cars 1993 Data](image)

**JMP Hint:**

The order of variables is important in the Tables→Summary platform. JMP uses the order to create the columns in the new data table. If you hold down the CTRL key and click to select variables in the Select Columns area, JMP uses the columns in the order they appear in the data table. This might not be what you want.

One advantage of using Tables→Summary is that the results are in a data table. You can create bar charts, or use the data in any appropriate JMP platform.

This special topic is an extremely brief introduction to summarizing multiple variables in tables. The Summary platform can create very complex tables, with means, standard deviations, and other statistics in columns. The Tabulate platform can create similarly complex tables, with the distinction that these tables are in a report window, rather than in a data table. Discussing all of the features of these two platforms would require another book!