IBM SPSS Custom Tables 19

Note: Before using this information and the product it supports, read the general information under Notices on p. 187.

This document contains proprietary information of SPSS Inc, an IBM Company. It is provided under a license agreement and is protected by copyright law. The information contained in this publication does not include any product warranties, and any statements provided in this manual should not be interpreted as such.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Preface

IBM® SPSS® Statistics is a comprehensive system for analyzing data. The Custom Tables optional add-on module provides the additional analytic techniques described in this manual. The Custom Tables add-on module must be used with the SPSS Statistics Core system and is completely integrated into that system.

About SPSS Inc., an IBM Company

SPSS Inc., an IBM Company, is a leading global provider of predictive analytic software and solutions. The company's complete portfolio of products — data collection, statistics, modeling and deployment — captures people's attitudes and opinions, predicts outcomes of future customer interactions, and then acts on these insights by embedding analytics into business processes. SPSS Inc. solutions address interconnected business objectives across an entire organization by focusing on the convergence of analytics, IT architecture, and business processes. Commercial, government, and academic customers worldwide rely on SPSS Inc. technology as a competitive advantage in attracting, retaining, and growing customers, while reducing fraud and mitigating risk. SPSS Inc. was acquired by IBM in October 2009. For more information, visit *http://www.spss.com*.

Technical support

Technical support is available to maintenance customers. Customers may contact Technical Support for assistance in using SPSS Inc. products or for installation help for one of the supported hardware environments. To reach Technical Support, see the SPSS Inc. web site at *http://support.spss.com* or find your local office via the web site at *http://support.spss.com/default.asp?refpage=contactus.asp*. Be prepared to identify yourself, your organization, and your support agreement when requesting assistance.

Customer Service

If you have any questions concerning your shipment or account, contact your local office, listed on the Web site at *http://www.spss.com/worldwide*. Please have your serial number ready for identification.

Training Seminars

SPSS Inc. provides both public and onsite training seminars. All seminars feature hands-on workshops. Seminars will be offered in major cities on a regular basis. For more information on these seminars, contact your local office, listed on the Web site at *http://www.spss.com/worldwide*.

iii

Additional Publications

The *SPSS Statistics: Guide to Data Analysis*, *SPSS Statistics: Statistical Procedures Companion*, and *SPSS Statistics: Advanced Statistical Procedures Companion*, written by Marija Norušis and published by Prentice Hall, are available as suggested supplemental material. These publications cover statistical procedures in the SPSS Statistics Base module, Advanced Statistics module and Regression module. Whether you are just getting starting in data analysis or are ready for advanced applications, these books will help you make best use of the capabilities found within the IBM® SPSS® Statistics offering. For additional information including publication contents and sample chapters, please see the author's website: *http://www.norusis.com*

Contents

7   *Tables for Variables with Shared Categories*   *98*

8   *Summary Statistics*   *107*

9   *Summarizing Scale Variables*   *122*

10   *Test Statistics*   *132*

11 Multiple Response Sets   150

12 Missing Values   163

13 Formatting and Customizing Tables   168

Appendices

Getting Started with Custom Tables

Many procedures produce results in the form of tables. The Custom Tables add-on module, however, offers special features designed to support a wide variety of customized reporting capabilities. Many of the custom features are particularly useful for survey analysis and marketing research.

This guide assumes that you already know the basics of using IBM® SPSS® Statistics. If you are unfamiliar with basic operation, see the introductory tutorial provided with the software. From the menu bar in any open SPSS Statistics window, choose:

Help > Tutorial

Table Structure and Terminology

The Custom Tables add-on module can produce a wide variety of customized tables. While you can discover a great deal of its capabilities simply by experimenting with the table builder interface, it may be helpful to know something about basic table structure and the terms we use to describe different structural elements that you can use in a table.

Pivot Tables

Tables produced by the Custom Tables module are displayed as **pivot tables** in the Viewer window. Pivot tables provide a great deal of flexibility over the formatting and presentation of tables.

For detailed information about working with pivot tables, use the Help system.

▶ From the menus in any open window, choose:

Help > Topics

▶ In the Contents pane, double-click Core System.

▶ Then double-click Pivot Tables in the expanded contents list.

Variables and Level of Measurement

To a certain extent, what you can do with a variable in a table is limited by its defined level of measurement. The Custom Tables procedure makes a distinction between two basic types of variables, based on level of measurement:

Categorical. Data with a limited number of distinct values or categories (for example, gender or religion). Also referred to as qualitative data. Categorical variables can be string (alphanumeric) data or numeric variables that use numeric codes to represent categories (for example, 0 = *Female* and 1 = *Male*). Categorical variables can be further divided into:

- **Nominal.** A variable can be treated as nominal when its values represent categories with no intrinsic ranking (for example, the department of the company in which an employee works). Examples of nominal variables include region, zip code, and religious affiliation.
- **Ordinal.** A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (for example, levels of service satisfaction from highly dissatisfied to highly satisfied). Examples of ordinal variables include attitude scores representing degree of satisfaction or confidence and preference rating scores.

Variables defined as nominal or ordinal in the Data Editor are treated as categorical variables in the Custom Tables procedure.

Scale. A variable can be treated as scale (continuous) when its values represent ordered categories with a meaningful metric, so that distance comparisons between values are appropriate. Examples of scale variables include age in years and income in thousands of dollars. Also referred to as quantitative, or continuous, data. Variables defined as scale in the Data Editor are treated as scale variables in the Custom Tables procedure.

Value Labels

For categorical variables, the preview displayed on the canvas pane in the table builder relies on defined **value labels**. The categories displayed in the table are, in fact, the defined value labels for that variable. If there are no defined value labels for the variable, the preview displays two generic categories. The actual number of categories that will be displayed in the final table is determined by the number of distinct values that occur in the data. The preview simply assumes that there will be at least two categories.

Additionally, some custom table-building features are not available for categorical variables that have no defined value labels.

Rows, Columns, and Cells

Each dimension of a table is defined by a single variable or a combination of variables. Variables that appear down the left side of a table are called **row variables**. They define the rows in a table. Variables that appear across the top of a table are called **column variables**. They define the columns in a table. The body of a table is made up of **cells**, which contain the basic information

conveyed by the table—counts, sums, means, percentages, and so on. A cell is formed by the intersection of a row and column of a table.

Stacking

Stacking can be thought of as taking separate tables and pasting them together into the same display. For example, you could display information on *Gender* and *Age category* in separate sections of the same table.

Figure 1-1
Stacked variables

		Count
Gender	Male	1232
	Female	1600
Age category	Less than 25	242
	25 to 34	627
	35 to 44	679
	45 to 54	481
	55 to 64	320
	65 or older	479

Although the term "stacking" typically denotes a vertical display, you can also stack variables horizontally.

Figure 1-2
Horizontal stacking

Gender		Age category					
Male	Female	Less than 25	25 to 34	35 to 44	45 to 54	55 to 64	65 or older
1232	1600	242	627	679	481	320	479

Crosstabulation

Crosstabulation is a basic technique for examining the relationship between two categorical variables. For example, using *Age category* as a row variable and *Gender* as a column variable, you can create a two-dimensional crosstabulation that shows the number of males and females in each age category.

Figure 1-3
Simple two-dimensional crosstabulation

		Gender	
		Male	Female
Age category	Less than 25	108	134
	25 to 34	276	351
	35 to 44	309	370
	45 to 54	221	260
	55 to 64	136	184
	65 or older	178	301

Nesting

Nesting, like crosstabulation, can show the relationship between two categorical variables, except one variable is nested within the other in the same dimension. For example, you could nest *Gender* within *Age category* in the row dimension, showing the number of males and females in each age category.

In this example, the nested table displays essentially the same information as a crosstabulation of the same two variables.

Figure 1-4
Nested variables

Age category	Less than 25	Male	108
		Female	134
	25 to 34	Male	276
		Female	351
	35 to 44	Male	309
		Female	370
	45 to 54	Male	221
		Female	260
	55 to 64	Male	136
		Female	184
	65 or older	Male	178
		Female	301

(Count)

Layers

You can use layers to add a dimension of depth to your tables, creating three-dimensional "cubes." Layers are, in fact, quite similar to nesting; the primary difference is that only one layer category is visible at a time. For example, using *Age category* as the row variable and *Gender* as a layer variable produces a table in which information for males and females is displayed in different layers of the table.

Figure 1-5
Layered variables

Gender Female

Age category	Less than 25	134
	25 to 34	351
	35 to 44	370

Gender Male

Age category	Less than 25	108
	25 to 34	276
	35 to 44	309
	45 to 54	221
	55 to 64	136
	65 or older	178

Tables for Variables with Shared Categories

Surveys often contain many questions with a common set of possible responses. For example, our sample survey contains a number of variables concerning confidence in various public and private institutions and services, all with the same set of response categories: 1 = *A great deal*, 2 = *Only some*, and 3 = *Hardly any*. You can use stacking to display these related variables in the same table—and you can display the shared response categories in the columns of the table.

Figure 1-6
Stacked variables with shared response categories in columns

	A great deal	Only some	Hardly any
Confidence in banks & financial institutions	490	1068	306
Confidence in education	511	1055	315
Confidence in major companies	500	1078	243
Confidence in medicine	844	864	167
Confidence in press	176	878	808
Confidence in television	196	936	744

Multiple Response Sets

Multiple response sets use multiple variables to record responses to questions for which the respondent can give more than one answer. For example, our sample survey asks the question, "Which of the following sources do you rely on for news?" Respondents can select any combination of five possible choices: *Internet*, *television*, *radio*, *newspapers*, and *news magazines*. Each of these choices is stored as a separate variable in the data file, and together they make a multiple response set. With the Custom Tables module, you can define a multiple response set based on these variables and use that multiple response set in the tables you create.

Figure 1-7
Multiple response set displayed in a table

		Count	Column N %
News sources	Get news from internet	867	41.7%
	Get news from radio	551	26.5%
	Get news from television	1077	51.8%
	Get news from news magazines	294	14.1%
	Get news from newspapers	805	38.7%

You may notice in this example that the percentages total to more than 100%. Because each respondent may choose more than one answer, the total number of responses can be greater than the total number of respondents.

Totals and Subtotals

You have a great deal of control over the display of totals and subtotals, including:

- Overall row and column totals
- Group totals for nested, stacked, and layered tables
- Subgroup totals

Figure 1-8
Subtotals, group totals, and table totals

		Count	Percent
Male	Less than 25	108	8.8%
	25 to 34	276	22.5%
	35 to 44	309	25.2%
	Subtotal < 45	693	56.4%
	45 to 54	221	18.0%
	55 to 64	136	11.1%
	65 or older	178	14.5%
	Subtotal 45+	535	43.6%
	Total	1228	100.0%
Female	Less than 25	134	8.4%
	25 to 34	351	21.9%
	35 to 44	370	23.1%
	Subtotal < 45	855	53.4%
	45 to 54	260	16.3%
	55 to 64	184	11.5%
	65 or older	301	18.8%
	Subtotal 45+	745	46.6%
	Total	1600	100.0%
Total	Less than 25	242	8.6%
	25 to 34	627	22.2%
	35 to 44	679	24.0%
	Subtotal < 45	1548	54.7%
	45 to 54	481	17.0%
	55 to 64	320	11.3%
	65 or older	479	16.9%
	Subtotal 45+	1280	45.3%
	Total	2828	100.0%

Custom Summary Statistics for Totals

For tables that contain totals or subtotals, you can have different summary statistics than the summaries displayed for each category. For example, you could display counts for an ordinal categorical row variable and display the mean for the "total" statistic.

Figure 1-9
Categorical variable and summary statistics in the same dimension

Confidence in television	1 A great deal	Count	196
	2 Only some	Count	936
	3 Hardly any	Count	744
	Total	Count	1876
		Mean	2.29

Sample Data File

Most of the examples presented here use the data file *survey_sample.sav*. For more information, see the topic Sample Files in Appendix A on p. 178. This data file is a fictitious survey of several thousand people, containing basic demographic information and responses to a variety of questions, ranging from political views to television viewing habits.

Building a Table

Before you can build a table, you need some data to use in the table.

▶ From the menus, choose:
File > Open > Data...

Figure 1-10
File menu, Open
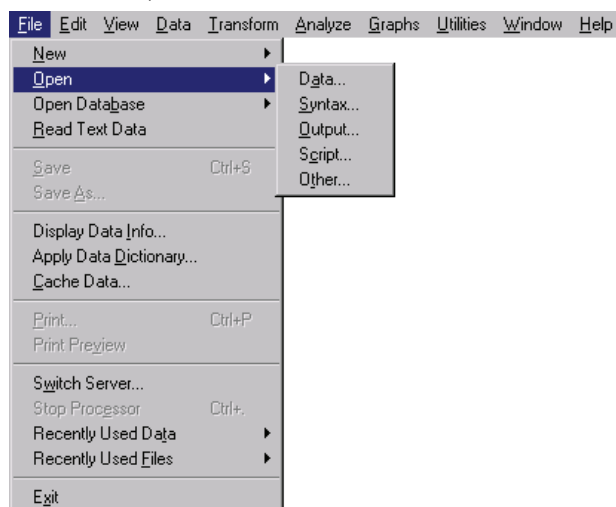
Alternatively, you can use the Open File button on the toolbar.

Figure 1-11
Open File toolbar button

▶ To use the data file in this example, see Sample Files on p. 178 for more information on data file locations.

▶ Open *survey_sample.sav*.

Opening the Custom Table Builder

▶ To open the custom table builder, from the menus, choose:
Analyze > Tables > Custom Tables...

Figure 1-12
Analyze menu, Tables
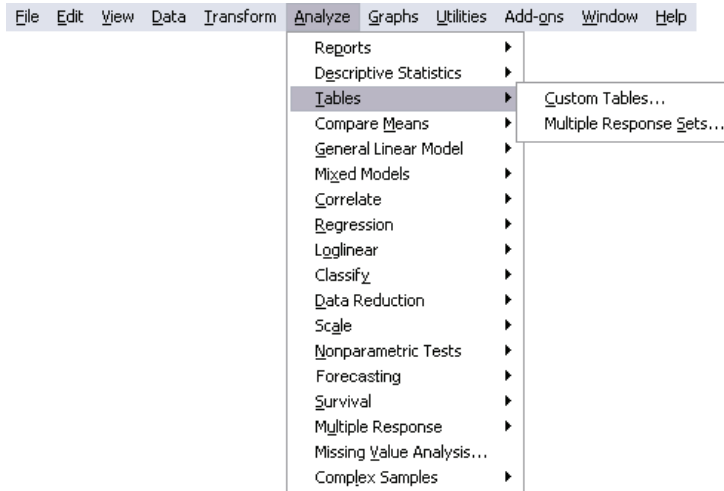
This opens the custom table builder.

Figure 1-13
Custom table builder
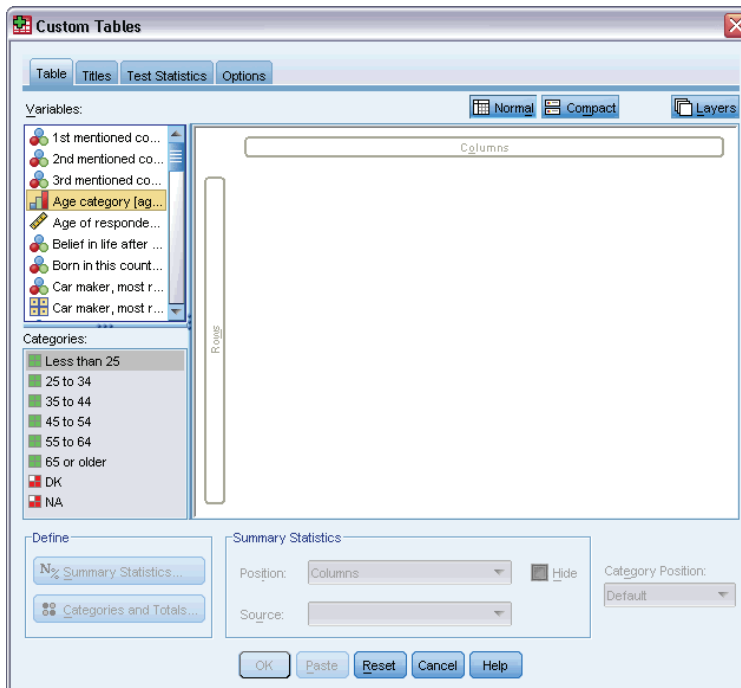
Selecting Row and Column Variables

To create a table, you simply drag and drop variables where you want them to appear in the table.

▶ Select (click) *Age category* in the variable list and drag and drop it into the Rows area on the canvas pane.

Figure 1-14
Selecting a row variable
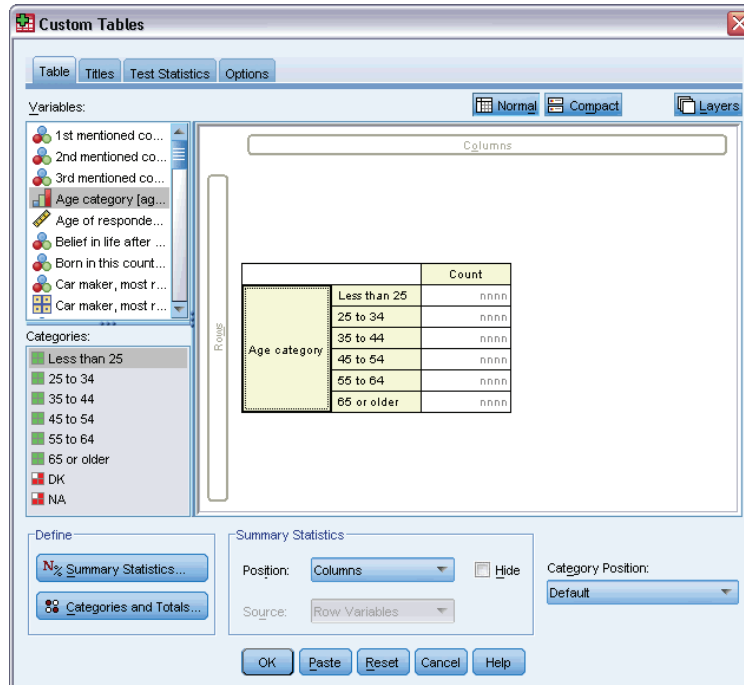
The canvas pane displays the table that would be created using this single row variable.

    The preview does not display the actual values that would be displayed in the table; it displays only the basic layout of the table.

▶ Select *Gender* in the variable list and drag and drop it into the Columns area on the canvas pane (you may have to scroll down the variable list to find this variable).

Figure 1-15
Selecting a column variable
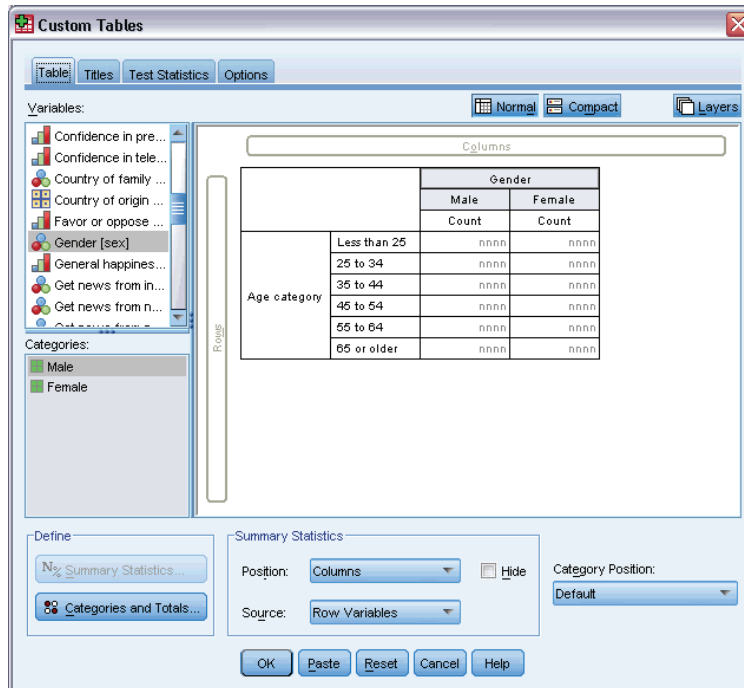
The canvas pane now displays a two-way crosstabulation of *Age category* by *Gender*.

By default, counts are displayed in the cells for categorical variables. You can also display row, column, and/or total percentages.

▶ Right-click on *Age category* on the canvas pane and select Summary Statistics from the pop-up context menu.

Figure 1-16
Context menu for categorical variables on canvas pane
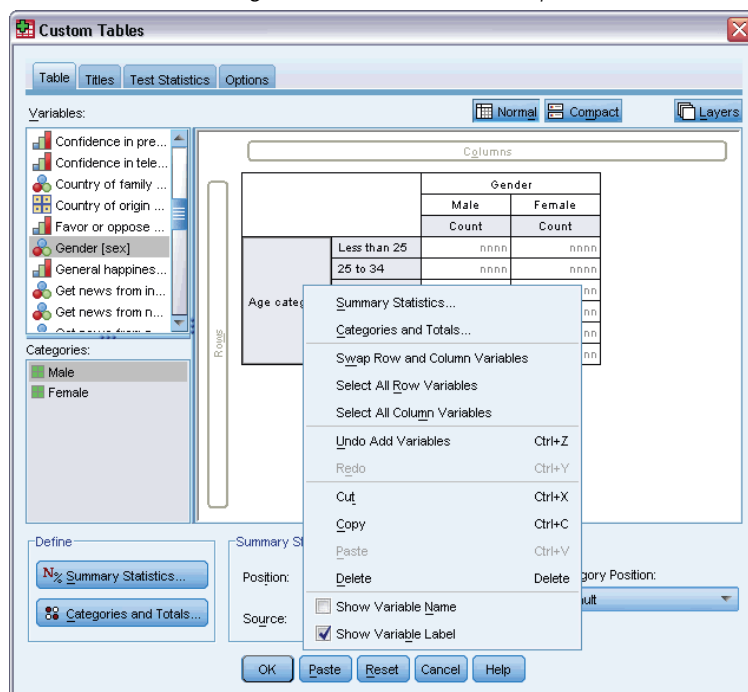
▶ In the Summary Statistics dialog box, select Row N % in the Statistics list and click the arrow button to add it to the Display list.

Now both the counts and row percentages will be displayed in the table.

Figure 1-17
Summary Statistics dialog box for categorical variables
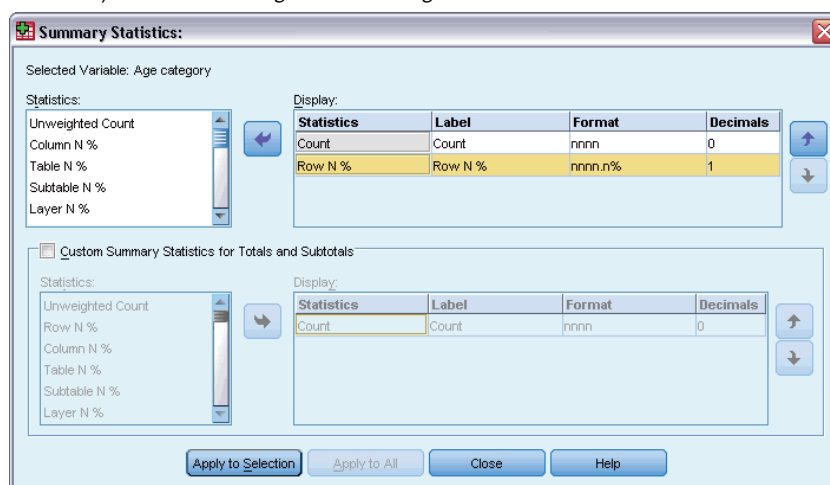
▶ Click Apply to Selection to save these settings and return to the table builder.

The canvas pane reflects the changes you have made, displaying columns for both counts and row percentages.

Figure 1-18
Counts and row percentages displayed on canvas pane
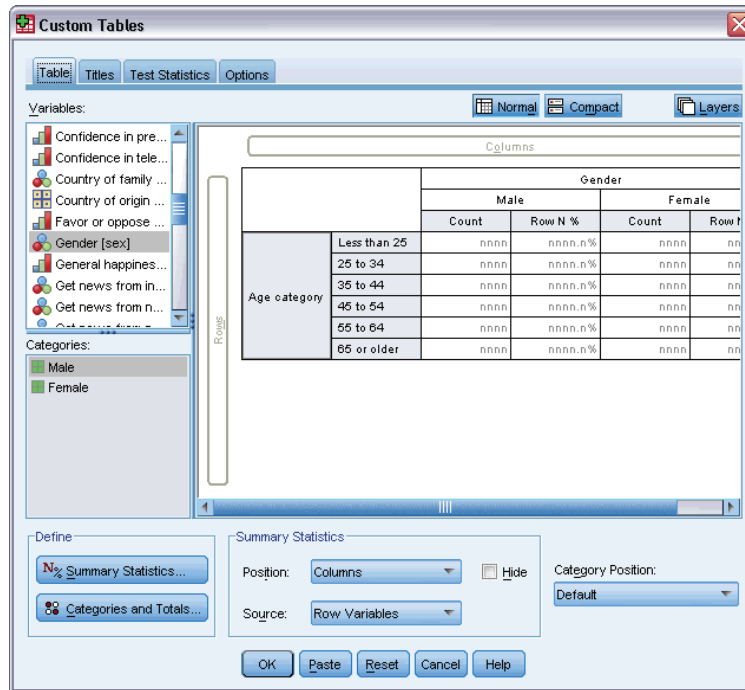
Inserting Totals and Subtotals

Totals are not displayed by default in custom tables, but it is easy to add both totals and subtotals to a table.

▶ Right-click on *Age category* on the canvas pane and select Categories and Totals from the pop-up context menu.

▶ In the Categories and Totals dialog box, select (click) 3.00 in the Value(s) list.

▶ Click Add Subtotal.

▶ In the Define Subtotal dialog, enter Subtotal <45 and then click Continue.

Figure 1-19
Define Subtotal dialog
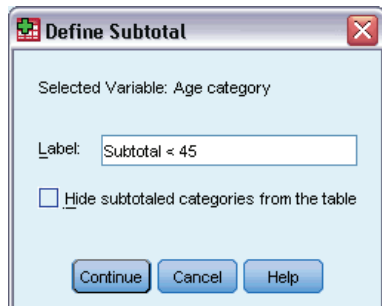
This inserts a row containing the subtotal for the first three age categories.

▶ Select (click) 6.00 in the Value(s) list.

▶ Click Add Subtotal.

▶ In the Define Subtotal dialog, enter Subtotal 45+ and then click Continue.

This inserts a row containing the subtotal for the last three age categories.

▶ To include an overall total, select the Total check box in the Show group.

Figure 1-20
Inserting totals and subtotals
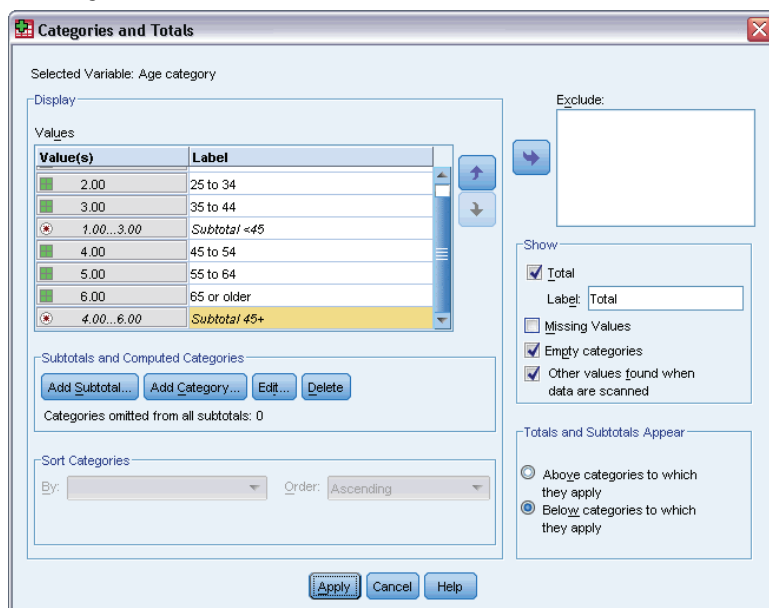
▶ Then click Apply.

The canvas pane preview now includes rows for the two subtotals and the overall total.

Figure 1-21
Total and subtotals on canvas pane
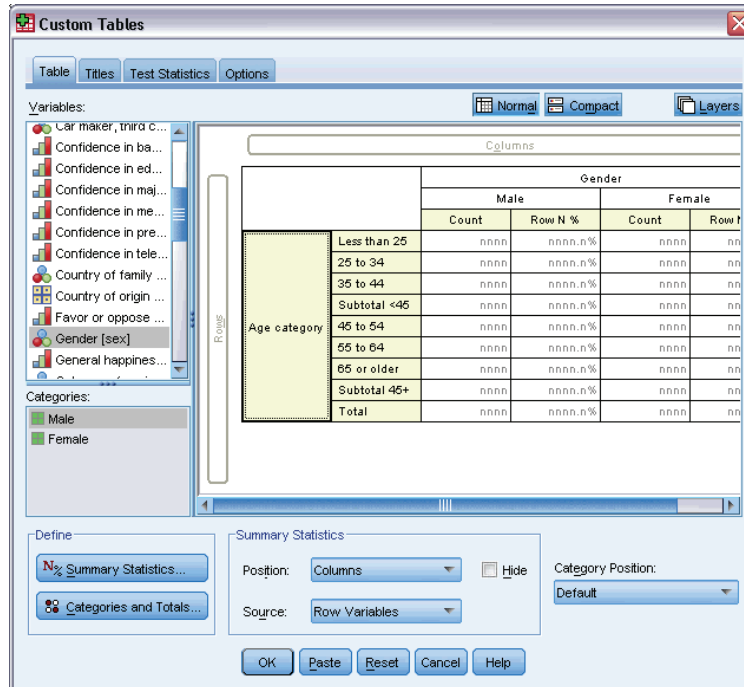
▶ Click OK to produce this table.

The table is displayed in the Viewer.

Figure 1-22
Crosstabulation with totals and subtotals

		Gender			
		Male		Female	
		Count	Row N %	Count	Row N %
Age category	Less than 25	108	44.6%	134	55.4%
	25 to 34	276	44.0%	351	56.0%
	35 to 44	309	45.5%	370	54.5%
	Subtotal < 45	693	44.8%	855	55.2%
	45 to 54	221	45.9%	260	54.1%
	55 to 64	136	42.5%	184	57.5%
	65 or older	178	37.2%	301	62.8%
	Subtotal 45+	535	41.8%	745	58.2%
	Total	1228	43.4%	1600	56.6%

Summarizing Scale Variables

A simple crosstabulation of two categorical variables displays counts or percentages in the cells of the table, but you can also display summaries of scale variables in the cells of the table.

▶ To open the custom table builder again, from the menus, choose:
   Analyze > Tables > Custom Tables...

▶ Click Reset to clear any previous selections.

▶ Select (click) *Age category* in the variable list and drag and drop it into the Rows area on the canvas pane.

Figure 1-23
Selecting a row variable

▶ Select *Hours per day watching TV* in the variable list and drag and drop it to the right of *Age category* in the row dimension of the table.

Figure 1-24
Dragging and dropping a scale variable into the row dimension
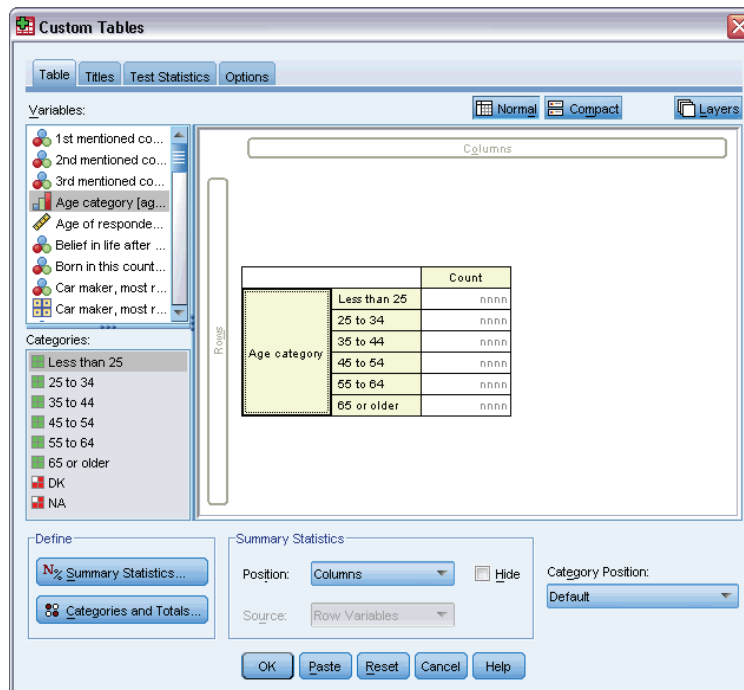
Now, instead of category counts, the table will display the mean (average) number of hours of television watched for each age category.

Figure 1-25
Scale variable summarized in table cells
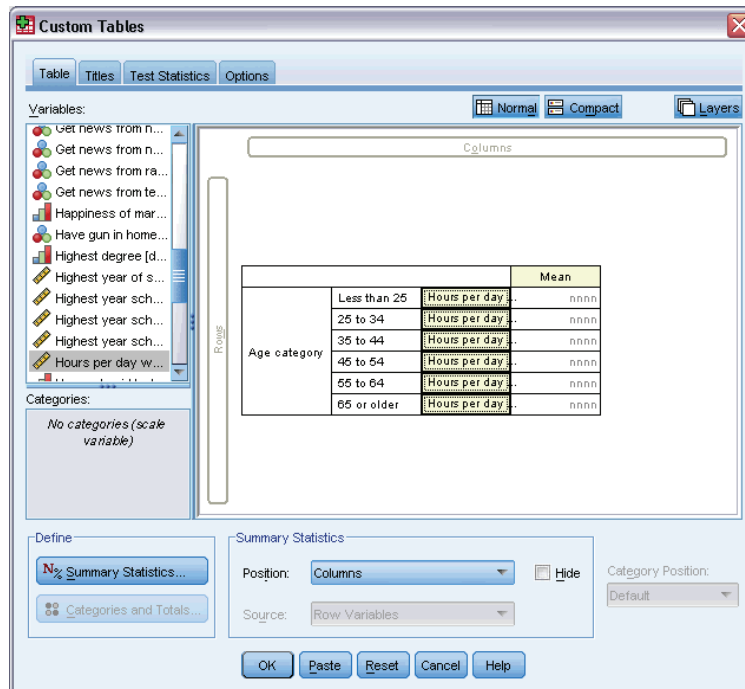
The mean is the default summary statistic for scale variables. You can add or change the summary statistics displayed in the table.

▶ Right-click the scale variable on the canvas pane, and select Summary Statistics from the pop-up context menu.

Figure 1-26
Context menu for scale variables in table preview
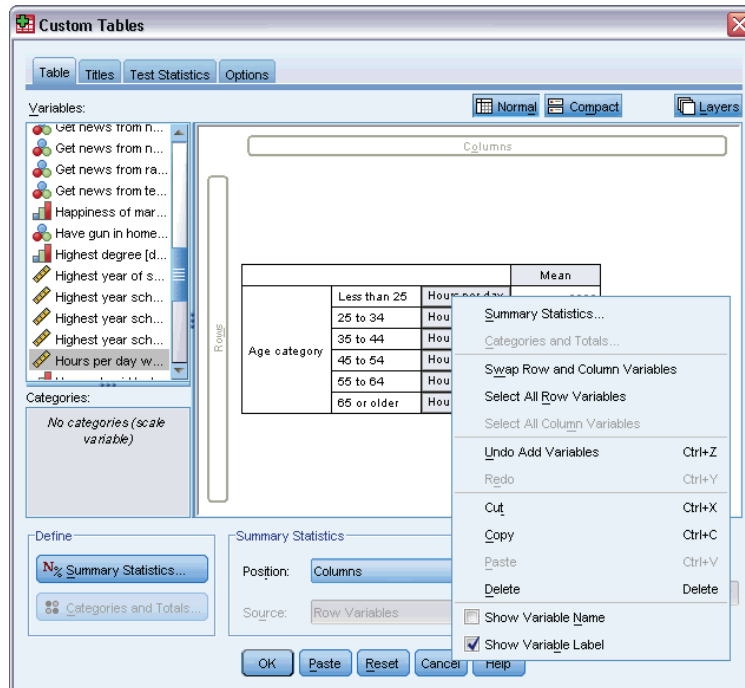
▶ In the Summary Statistics dialog box, select Median in the Statistics list and click the arrow button to add it to the Display list.

Now both the mean and the median will be displayed in the table.

Figure 1-27
Summary Statistics dialog box for scale variables
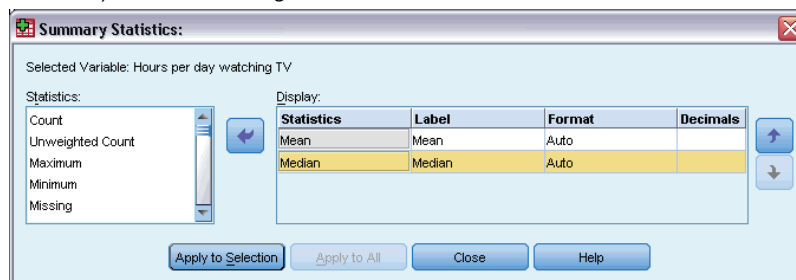
▶ Click Apply to Selection to save these settings and return to the table builder.

The canvas pane now shows that both the mean and median will be displayed in the table.

Figure 1-28
Mean and median scale summaries displayed on canvas pane
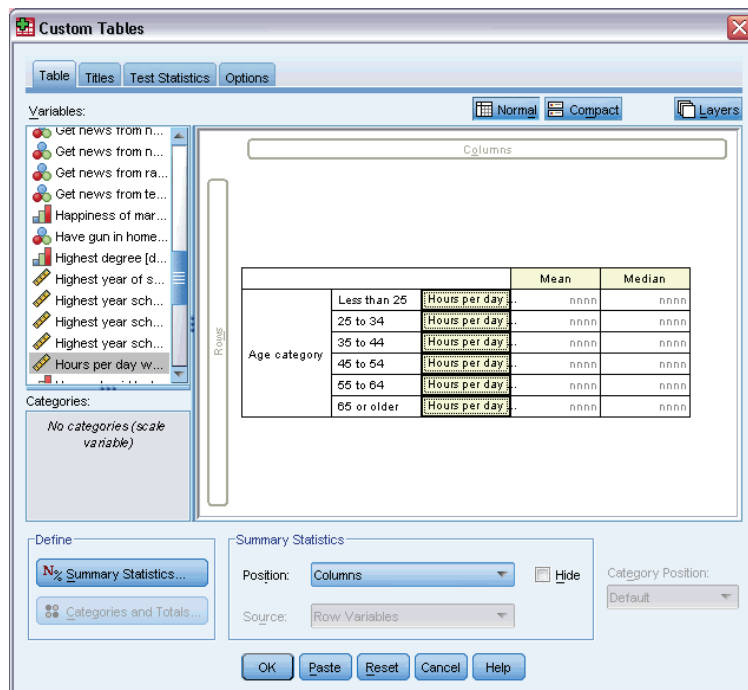
Before creating this table, let's clean it up a bit.

▶ Right-click on *Hours per day...* on the canvas pane and deselect (uncheck) Show Variable Label on the pop-up context menu.

Figure 1-29
Suppressing the display of variable labels
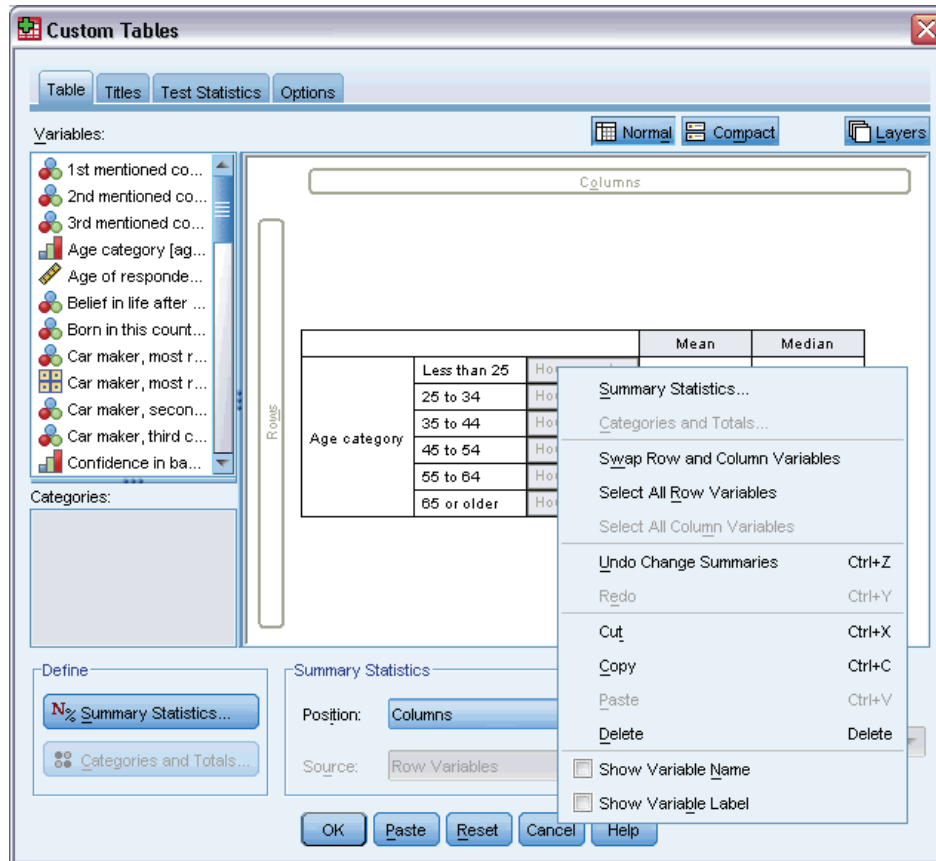
The column is still displayed in the table preview (with the variable label text grayed out), but this column will not be displayed in the final table.

▶ Click the Titles tab in the table builder.

▶ Enter a descriptive title for the table, such as Average Daily Number of Hours of Television Watched by Age Category.

Figure 1-30
Custom Tables dialog box, Titles tab
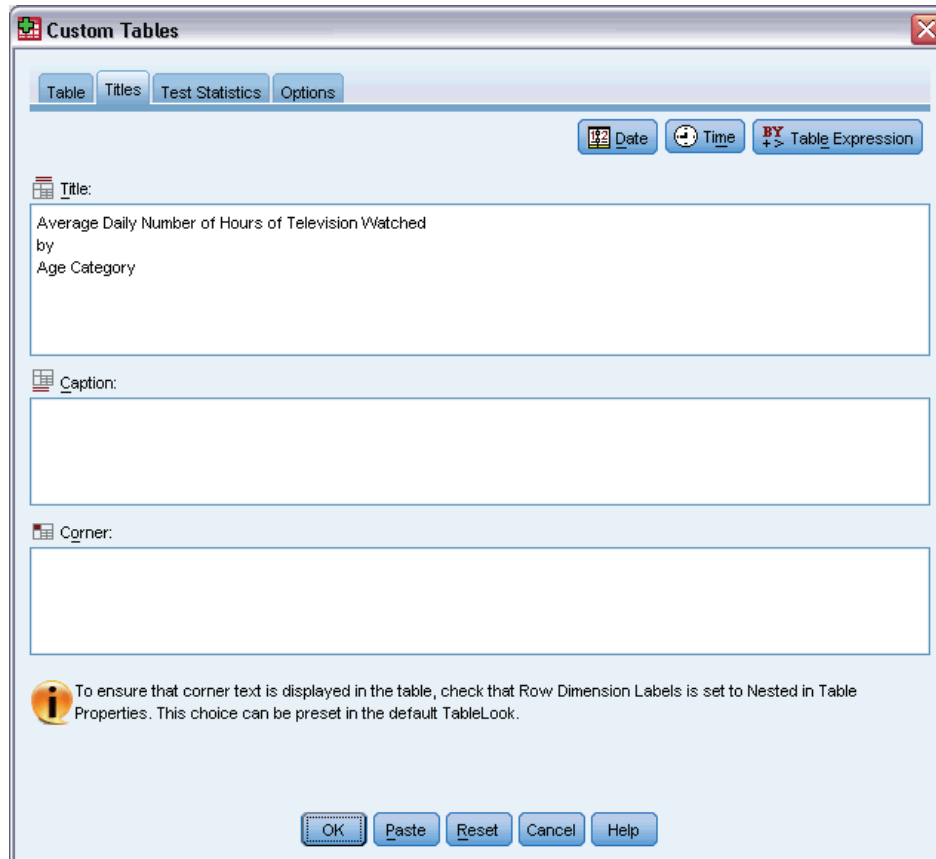
▶ Click OK to create the table.

The table is displayed in the Viewer window.

Figure 1-31
Mean and median number of TV hours by age category

**Average Daily Number of Hours of Television Watched
by
Age Category**

		Mean	Median
Age category	Less than 25	2.85	2.00
	25 to 34	2.78	2.00
	35 to 44	2.56	2.00
	45 to 54	2.58	2.00
	55 to 64	3.02	2.50
	65 or older	3.58	3.00

Table Builder Interface

Custom Tables uses a simple drag-and-drop table builder interface that allows you to preview your table as you select variables and options. It also provides a level of flexibility not found in a typical dialog box, including the ability to change the size of the window and the size of the panes within the window.

Building Tables

Figure 2-1
Custom Tables dialog box, Table tab
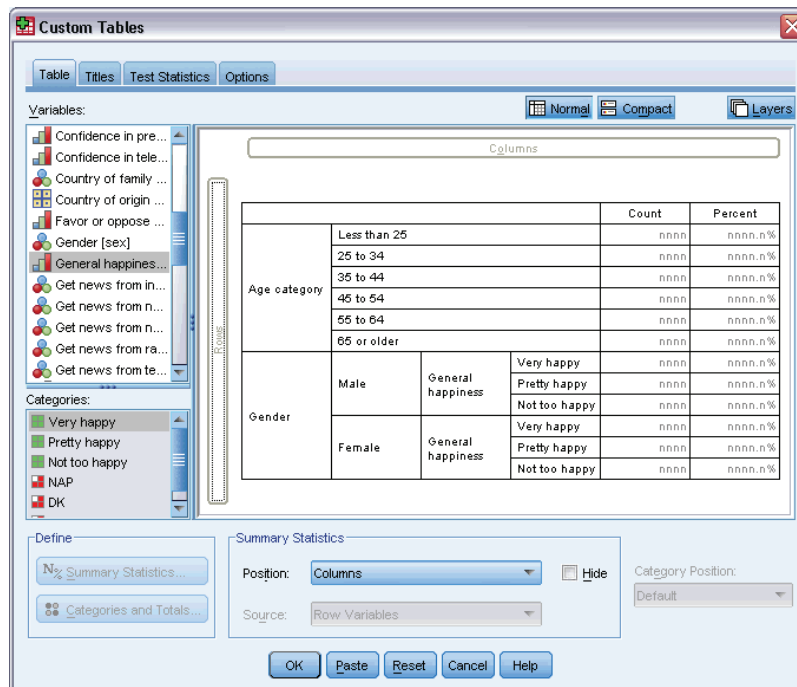
You select the variables and summary measures that will appear in your tables on the Table tab in the table builder.

Variable list. The variables in the data file are displayed in the top left pane of the window. Custom Tables distinguishes between two different measurement levels for variables and handles them differently depending on the measurement level:

Categorical. Data with a limited number of distinct values or categories (for example, gender or religion). Categorical variables can be string (alphanumeric) or numeric variables that use numeric codes to represent categories (for example, 0 = *male* and 1 = *female*). Also referred to as qualitative data. Categorical variables can be either **nominal** or **ordinal**:

- **Nominal.** A variable can be treated as nominal when its values represent categories with no intrinsic ranking (for example, the department of the company in which an employee works). Examples of nominal variables include region, zip code, and religious affiliation.

- **Ordinal.** A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (for example, levels of service satisfaction from highly dissatisfied to highly satisfied). Examples of ordinal variables include attitude scores representing degree of satisfaction or confidence and preference rating scores.

Scale. Data measured on an interval or ratio scale, where the data values indicate both the order of values and the distance between values. For example, a salary of $72,195 is higher than a salary of $52,398, and the distance between the two values is $19,797. Also referred to as quantitative or continuous data.

Categorical variables define categories (row, columns, and layers) in the table, and the default summary statistic is the count (number of cases in each category). For example, a default table of a categorical gender variable would simply display the number of males and the number of females.

Scale variables are typically summarized within categories of categorical variables, and the default summary statistic is the mean. For example, a default table of income within gender categories would display the mean income for males and the mean income for females.

You can also summarize scale variables by themselves, without using a categorical variable to define groups. This is primarily useful for **stacking** summaries of multiple scale variables.

Multiple Response Sets

Custom Tables also supports a special kind of "variable" called a **multiple response set**. Multiple response sets are not really variables in the normal sense. You cannot see them in the Data Editor, and other procedures do not recognize them. Multiple response sets use multiple variables to record responses to questions where the respondent can give more than one answer. Multiple response sets are treated like categorical variables, and most of the things you can do with categorical variables, you can also do with multiple response sets.

An icon next to each variable in the variable list identifies the variable type.

Measurement Level	Data Type			
	Numeric	String	Date	Time
Scale (Continuous)		n/a		

Ordinal				
Nominal				

Multiple response set, multiple categories	
Multiple response set, multiple dichotomies	

You can change the measurement level of a variable in the table builder by right-clicking the variable in the variable list and selecting Categorical or Scale from the pop-up context menu. You can permanently change a variable's measurement level in the Variable View of the Data Editor. Variables defined as **nominal** or **ordinal** are treated as categorical by Custom Tables.

Categories. When you select a categorical variable in the variable list, the defined categories for the variable are displayed in the Categories list. These categories will also be displayed on the canvas pane when you use the variable in a table. If the variable has no defined categories, the Categories list and the canvas pane will display two placeholder categories: *Category 1* and *Category 2*.

The defined categories displayed in the table builder are based on **value labels**, descriptive labels assigned to different data values (for example, numeric values of 0 and 1, with value labels of *male* and *female*). You can define value labels in Variable View of the Data Editor or with Define Variable Properties on the Data menu in the Data Editor window.

Canvas pane. You build a table by dragging and dropping variables onto the rows and columns of the canvas pane. The canvas pane displays a preview of the table that will be created. The canvas pane does not show actual data values in the cells, but it should provide a fairly accurate view of the layout of the final table. For categorical variables, the actual table may contain more categories than the preview if the data file contains unique values for which no value labels have been defined.

- Normal view displays all of the rows and columns that will be included in the table, including rows and/or columns for summary statistics and categories of categorical variables.

- Compact view shows only the variables that will be in the table, without a preview of the rows and columns that the table will contain.

Basic Rules and Limitations for Building a Table

- For categorical variables, summary statistics are based on the innermost variable in the statistics source dimension.

- The default statistics source dimension (row or column) for categorical variables is based on the order in which you drag and drop variables into the canvas pane. For example, if you drag a variable to the rows tray first, the row dimension is the default statistics source dimension.

- Scale variables can be summarized only within categories of the innermost variable in either the row or column dimension. (You can position the scale variable at any level of the table, but it is summarized at the innermost level.)

- Scale variables cannot be summarized within other scale variables. You can stack summaries of multiple scale variables or summarize scale variables within categories of categorical variables. You cannot nest one scale variable within another or put one scale variable in the row dimension and another scale variable in the column dimension.

- If any variable in the active dataset contains more than 12,000 defined value labels, you cannot use the table builder to create tables. If you don't need to include variables that exceed this limitation in your tables, you can define and apply variable sets that exclude those variables. If you need to include any variables with more than 12,000 defined values labels, you can use CTABLES command syntax to generate the tables.

To Build a Table

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ Drag and drop one or more variables to the row and/or column areas of the canvas pane.

▶ Click OK to create the table.

To delete a variable from the canvas pane in the table builder:

▶ Select (click) the variable on the canvas pane.

▶ Drag the variable anywhere outside the canvas pane, or press the Delete key.

To change the measurement level of a variable:

▶ Right-click the variable in the variable list (you can do this only in the variable list, not on the canvas).

▶ Select Categorical or Scale from the pop-up context menu.

Fields with Unknown Measurement Level

The Measurement Level alert is displayed when the measurement level for one or more variables (fields) in the dataset is unknown. Since measurement level affects the computation of results for this procedure, all variables must have a defined measurement level.

Figure 2-2
Measurement level alert
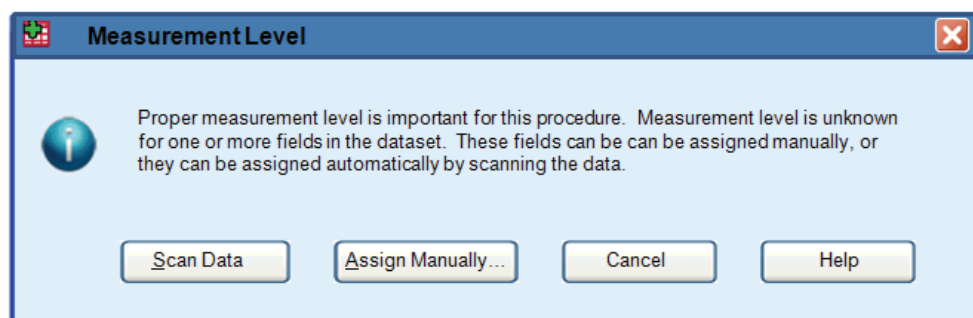
■ **Scan Data.** Reads the data in the active dataset and assigns default measurement level to any fields with a currently unknown measurement level. If the dataset is large, that may take some time.

■ **Assign Manually.** Opens a dialog that lists all fields with an unknown measurement level. You can use this dialog to assign measurement level to those fields. You can also assign measurement level in Variable View of the Data Editor.

Since measurement level is important for this procedure, you cannot access the dialog to run this procedure until all fields have a defined measurement level.

Stacking Variables

Stacking can be thought of as taking separate tables and pasting them together into the same display. For example, you could display information on *Gender* and *Age category* in separate sections of the same table.

To Stack Variables

▶ In the variable list, select all of the variables you want to stack, then drag and drop them together into the rows or columns of the canvas pane.

or

▶ Drag and drop variables separately, dropping each variable either above or below existing variables in the rows or to the right or left of existing variables in the columns.

Figure 2-3
Stacked variables

		Count
Gender	Male	1232
	Female	1600
Age category	Less than 25	242
	25 to 34	627
	35 to 44	679
	45 to 54	481
	55 to 64	320
	65 or older	479

Nesting Variables

Nesting, like crosstabulation, can show the relationship between two categorical variables, except that one variable is nested within the other in the same dimension. For example, you could nest *Gender* within *Age category* in the row dimension, showing the number of males and females in each age category.

You can also nest a scale variable within a categorical variable. For example, you could nest *Income* within *Gender*, showing separate mean (or median or other summary measure) income values for males and females.

To Nest Variables

▶ Drag and drop a categorical variable into the row or column area of the canvas pane.

▶ Drag and drop a categorical or scale variable to the left or right of the categorical row variable or above or below the categorical column variable.

Figure 2-4
Nested categorical variables

			Count
Age category	Less than 25	Male	108
		Female	134
	25 to 34	Male	276
		Female	351
	35 to 44	Male	309
		Female	370
	45 to 54	Male	221
		Female	260
	55 to 64	Male	136
		Female	184
	65 or older	Male	178
		Female	301

Figure 2-5
Scale variable nested within a categorical variable

Average Daily Number of Hours of Television Watched
by
Age Category

		Mean	Median
Age category	Less than 25	2.85	2.00
	25 to 34	2.78	2.00
	35 to 44	2.56	2.00
	45 to 54	2.58	2.00
	55 to 64	3.02	2.50
	65 or older	3.58	3.00

Note: Technically, the preceding table is an example of a categorical variable nested within a scale variable, but the resulting information conveyed in the table is essentially the same as nesting the scale variable within the categorical variable, without redundant labels for the scale variable. (Try it the other way around, and you will understand.)

Note: Custom Tables do not honor layered split file processing. To achieve the same result as layered split files, place the split file variables in the outermost nesting layers of the table.

Layers

You can use layers to add a dimension of depth to your tables, creating three-dimensional "cubes." Layers are similar to nesting or stacking; the primary difference is that only one layer category is visible at a time. For example, using *Age category* as the row variable and *Gender* as a layer variable produces a table in which information for males and females is displayed in different layers of the table.

28

Chapter 2

To Create Layers

▶ Click Layers on the Table tab in the table builder to display the Layers list.

▶ Drag and drop the scale or categorical variable(s) that will define the layers into the Layers list.

Figure 2-6
Layered variables

You cannot mix scale and categorical variables in the Layers list. All variables must be of the same type. Multiple response sets are treated as categorical for the Layers list. Scale variables in the layers are always stacked.

If you have multiple categorical layer variables, layers can be stacked or nested.

■ Show each category as a layer is equivalent to stacking. A separate layer will be displayed for each category of each layer variable. The total number of layers is simply the *sum* of the number of categories for each layer variable. For example, if you have three layer variables, each with three categories, the table will have nine layers.

■ Show each combination of categories as a layer is equivalent to nesting or crosstabulating layers. The total number of layers is the *product* of the number of categories for each layer variable. For example, if you have three variables, each with three categories, the table will have 27 layers.

Showing and Hiding Variable Names and/or Labels

The following options are available for the display of variable names and labels:

■ Show only variable labels. For any variables without defined variable labels, the variable name is displayed. This is the default setting.

■ Show only variable names.

■ Show both variable labels and variable names.

■ Don't show variable names or variable labels. Although the column/row that contains the variable label or name will still be displayed in the table preview on the canvas pane, this column/row will not be displayed in the actual table.

To show or hide variable labels or variable names:

▶ Right-click the variable in the table preview on the canvas pane.

▶ Select Show Variable Label or Show Variable Name from the pop-up context menu to toggle the display of labels or names on or off. A check mark next to the selection indicates that it will be displayed.

Summary Statistics

The Summary Statistics dialog box allows you to:

- Add and remove summary statistics from a table.
- Change the labels for the statistics.
- Change the order of the statistics.
- Change the format of the statistics, including the number of decimal positions.

Figure 2-7
Summary Statistics Categorical Variables dialog box
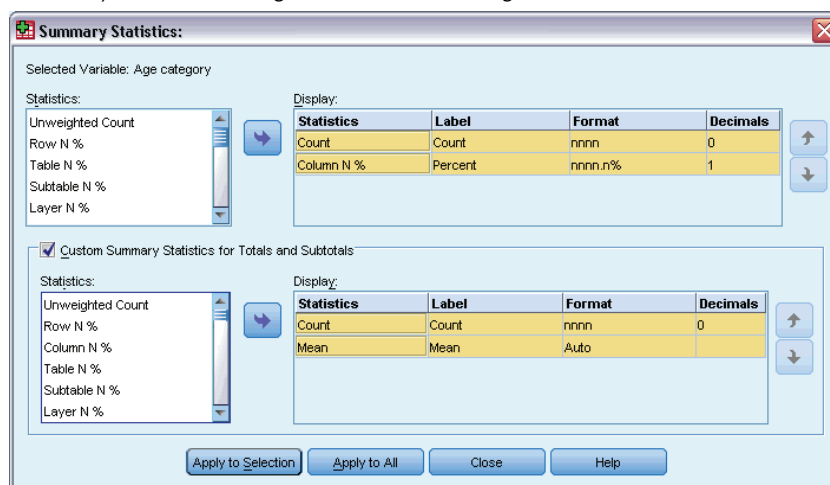
The summary statistics (and other options) available here depend on the measurement level of the summary statistics source variable, as displayed at the top of the dialog box. The source of summary statistics (the variable on which the summary statistics are based) is determined by:

- **Measurement level**. If a table (or a table section in a stacked table) contains a scale variable, summary statistics are based on the scale variable.
- **Variable selection order**. The default statistics source dimension (row or column) for categorical variables is based on the order in which you drag and drop variables onto the canvas pane. For example, if you drag a variable to the rows area first, the row dimension is the default statistics source dimension.
- **Nesting**. For categorical variables, summary statistics are based on the innermost variable in the statistics source dimension.

A stacked table may have multiple summary statistics source variables (both scale and categorical), but each table section has only one summary statistics source.

To Change the Summary Statistics Source Dimension

▶ Select the dimension (rows, columns, or layers) from the Source drop-down list in the Summary Statistics group of the Table tab.

To Control the Summary Statistics Displayed in a Table

▶ Select (click) the summary statistics source variable on the canvas pane of the Table tab.

▶ In the Define group of the Table tab, click Summary Statistics.

or

▶ Right-click the summary statistics source variable on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ Select the summary statistics you want to include in the table. You can use the arrow to move selected statistics from the Statistics list to the Display list, or you can drag and drop selected statistics from the Statistics list into the Display list.

▶ Click the up or down arrows to change the display position of the currently selected summary statistic.

▶ Select a display format from the Format drop-down list for the selected summary statistic.

▶ Enter the number of decimals to display in the Decimals cell for the selected summary statistic.

▶ Click Apply to Selection to include the selected summary statistics for the currently selected variables on the canvas pane.

▶ Click Apply to All to include the selected summary statistics for all stacked variables of the same type on the canvas pane.

Note: Apply to All differs from Apply to Selection only for stacked variables of the same type already on the canvas pane. In both cases, the selected summary statistics are automatically included for any additional stacked variables of the same type that you add to the table.

Summary Statistics for Categorical Variables

The basic statistics available for categorical variables are counts and percentages. You can also specify custom summary statistics for totals and subtotals. These custom summary statistics include measures of central tendency (such as mean and median) and dispersion (such as standard deviation) that may be suitable for some ordinal categorical variables. For more information, see the topic Custom Total Summary Statistics for Categorical Variables on p. 33.

Count. Number of cases in each cell of the table or number of responses for multiple response sets.

Unweighted Count. Unweighted number of cases in each cell of the table.

Column percentages. Percentages within each column. The percentages in each column of a subtable (for simple percentages) sum to 100%. Column percentages are typically useful only if you have a categorical *row* variable.

Row percentages. Percentages within each row. The percentages in each row of a subtable (for simple percentages) sum to 100%. Row percentages are typically useful only if you have a categorical *column* variable.

Layer Row and Layer Column percentages. Row or column percentages (for simple percentages) sum to 100% across all subtables in a nested table. If the table contains layers, row or column percentages sum to 100% across all nested subtables in each layer.

Layer percentages. Percentages within each layer. For simple percentages, cell percentages within the currently visible layer sum to 100%. If you do not have any layer variables, this is equivalent to table percentages.

Table percentages. Percentages for each cell are based on the entire table. All cell percentages are based on the same total number of cases and sum to 100% (for simple percentages) over the entire table.

Subtable percentages. Percentages in each cell are based on the subtable. All cell percentages in the subtable are based the same total number of cases and sum to 100% within the subtable. In nested tables, the variable that precedes the innermost nesting level defines subtables. For example, in a table of *Marital status* within *Gender* within *Age category*, *Gender* defines subtables.

Multiple response sets can have percentages based on cases, responses, or counts. For more information, see the topic Summary Statistics for Multiple Response Sets on p. 32.

Stacked Tables

For percentage calculations, each table section defined by a stacking variable is treated as a separate table. Layer Row, Layer Column, and Table percentages sum to 100% (for simple percentages) within each stacked table section. The percentage base for different percentage calculations is based on the cases in each stacked table section.

Percentage Base

Percentages can be calculated in three different ways, determined by the treatment of missing values in the computational base:

Simple percentage. Percentages are based on the number of cases used in the table and always sum to 100%. If a category is excluded from the table, cases in that category are excluded from the base. Cases with system-missing values are always excluded from the base. Cases with user-missing values are excluded if user-missing categories are excluded from the table (the default) or included if user-missing categories are included in the table. Any percentage that does not have *Valid N* or *Total N* in its name is a simple percentage.

Total N percentage. Cases with system-missing and user-missing values are added to the Simple percentage base. Percentages may sum to less than 100%.

Valid N percentage. Cases with user-missing values are removed from the Simple percentage base even if user-missing categories are included in the table.

Note: Cases in manually excluded categories other than user-missing categories are always excluded from the base.

Summary Statistics for Multiple Response Sets

The following additional summary statistics are available for multiple response sets.

Col/Row/Layer Responses %. Percentage based on responses.

Col/Row/Layer Responses % (Base: Count). Responses are the numerator and total count is the denominator.

Col/Row/Layer Count % (Base: Responses). Count is the numerator and total responses are the denominator.

Layer Col/Row Responses %. Percentage across subtables. Percentage based on responses.

Layer Col/Row Responses % (Base: Count). Percentages across subtables. Responses are the numerator and total count is the denominator.

Layer Col/RowResponses % (Base: Responses). Percentages across subtables. Count is the numerator and total responses is the denominator.

Responses. Count of responses.

Subtable/Table Responses %. Percentage based on responses.

Subtable/Table Responses % (Base: Count). Responses are the numerator and total count is the denominator.

Subtable/Table Count % (Base: Responses). Count is the numerator and total responses are the denominator.

Summary Statistics for Scale Variables and Categorical Custom Totals

In addition to the counts and percentages available for categorical variables, the following summary statistics are available for scale variables and as custom total and subtotal summaries for categorical variables. These summary statistics are not available for multiple response sets or string (alphanumeric) variables.

Mean. Arithmetic average; the sum divided by the number of cases.

Median. Value above and below which half of the cases fall; the 50th percentile.

Mode. Most frequent value. If there is a tie, the smallest value is shown.

Minimum. Smallest (lowest) value.

Maximum. Largest (highest) value.

Missing. Count of missing values (both user- and system-missing).

Percentile. You can include the 5th, 25th, 75th, 95th, and/or 99th percentiles.

Range. Difference between maximum and minimum values.

Standard error of the mean. A measure of how much the value of the mean may vary from sample to sample taken from the same distribution. It can be used to roughly compare the observed mean to a hypothesized value (that is, you can conclude that the two values are different if the ratio of the difference to the standard error is less than –2 or greater than +2).

Standard deviation. A measure of dispersion around the mean. In a normal distribution, 68% of the cases fall within one standard deviation of the mean and 95% of the cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution (the square root of the variance).

Sum. Sum of the values.

Sum percentage. Percentages based on sums. Available for rows and columns (within subtables), entire rows and columns (across subtables), layers, subtables, and entire tables.

Total N. Count of non-missing, user-missing, and system-missing values. Does not include cases in manually excluded categories other than user-missing categories.

Valid N. Count of non-missing values. Does not include cases in manually excluded categories other than user-missing categories.

Variance. A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself (the square of the standard deviation).

Stacked Tables

Each table section defined by a stacking variable is treated as a separate table, and summary statistics are calculated accordingly.

Custom Total Summary Statistics for Categorical Variables

For tables of categorical variables that contain totals or subtotals, you can have different summary statistics than the summaries displayed for each category. For example, you could display counts and column percentages for an ordinal categorical row variable and display the median for the "total" statistic.

To create a table for a categorical variable with a custom total summary statistic:

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

The table builder will open.

▶ Drag and drop a categorical variable into the Rows or Columns area of the canvas.

▶ Right-click on the variable on the canvas and select Categories and Totals from the pop-up context menu.

▶ Click (check) the Total check box, and then click Apply.

▶ Right-click the variable again on the canvas and select Summary Statistics from the pop-up context menu.

▶ Click (check) Custom Summary Statistics for Totals and Subtotals, and then select the custom summary statistics you want.

By default, all summary statistics, including custom summaries, are displayed in the opposite dimension from the dimension containing the categorical variable. For example, if you have a categorical row variable, summary statistics define columns in the table, as in:

Figure 2-8
Default position of summary statistics

		Count	Mean
Confidence in television	1 A great deal	196	
	2 Only some	936	
	3 Hardly any	744	
	Total	1876	2.29

To display summary statistics in the same dimension as the categorical variable:

▶ On the Table tab in the table builder, in the Summary Statistics group, select the dimension from the Position drop-down list.

For example, if the categorical variable is displayed in the rows, select Rows from the drop-down list.

Figure 2-9
Categorical variable and summary statistics in the same dimension

Confidence in television	1 A great deal	Count	196
	2 Only some	Count	936
	3 Hardly any	Count	744
	Total	Count	1876
		Mean	2.29

Summary Statistics Display Formats

The following display format options are available:

nnnn. Simple numeric.

nnnn%. Percentage sign appended to end of value.

Auto. Defined variable display format, including number of decimals.

N=nnnn. Displays *N=* before the value. This can be useful for counts, valid *N*, and total *N* in tables where the summary statistics labels are not displayed.

(nnnn). All values enclosed in parentheses.

(nnnn)(neg. value). Only negative values enclosed in parentheses.

(nnnn%). All values enclosed in parentheses and a percentage sign appended to end of values.

n,nnn.n. Comma format. Comma used as grouping separator and period used as decimal indicator regardless of locale settings.

n.nnn,n. Dot format. Period used as grouping separator and comma used as decimal indicator regardless of locale settings.

$n,nnn.n. Dollar format. Dollar sign displayed in front of value; comma used as grouping separator and period used as decimal indicator regardless of locale settings.

CCA, CCB, CCC, CCD, CCE. Custom currency formats. The current defined format for each custom currency is displayed in the list. These formats are defined on the Currency tab in the Options dialog box (Edit menu, Options).

General Rules and Limitations

- With the exception of Auto, the number of decimals is determined by the Decimals column setting.

- With the exception of the comma, dollar, and dot formats, the decimal indicator used is the one defined for the current locale in your Windows Regional Options control panel.

- Although comma/dollar and dot will display either a comma or period respectively as the grouping separator, there is no display format available at creation time to display a grouping separator based on the current locale settings (defined in the Windows Regional Options control panel).

Categories and Totals

The Categories and Totals dialog box allows you to:

- Reorder and exclude categories.
- Insert subtotals and totals.
- Insert computed categories.
- Include or exclude empty categories.
- Include or exclude categories defined as containing missing values.
- Include or exclude categories that do not have defined value labels.

Figure 2-10
Categories and Totals dialog box
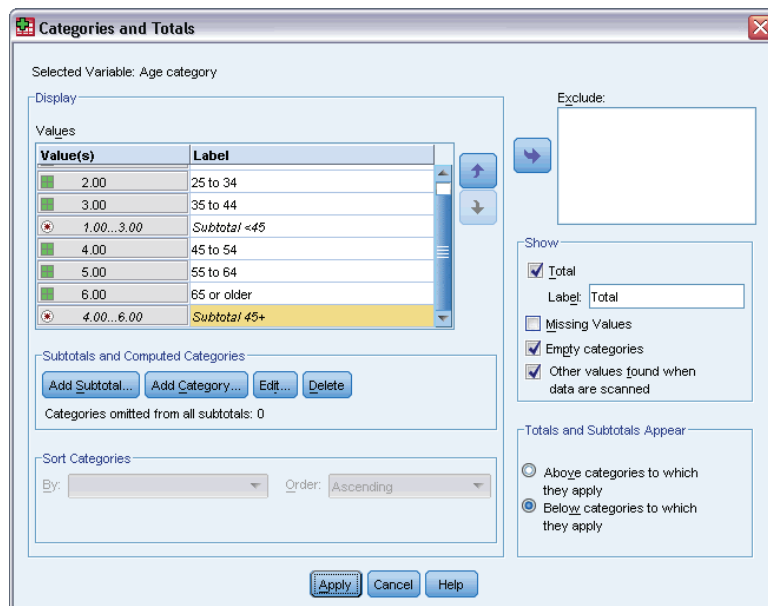
- This dialog box is available only for categorical variables and multiple response sets. It is not available for scale variables.

- For multiple selected variables with different categories, you cannot insert subtotals, insert computed categories, exclude categories, or manually reorder categories. This occurs only if you select multiple variables in the canvas preview and access this dialog box for all selected variables simultaneously. You can still perform these actions for each variable separately.

- For variables with no defined value labels, you can only sort categories and insert totals.

To Access the Categories and Totals Dialog Box

▶ Drag and drop a categorical variable or multiple response set onto the canvas pane.

▶ Right-click the variable on the canvas pane, and select Categories and Totals from the pop-up context menu.

   or

▶ Select (click) the variable on the canvas pane, and then click Categories and Totals in the Define group on the Table tab.

   You can also select multiple categorical variables in the same dimension on the canvas pane:

▶ Ctrl-click each variable on the canvas pane.

   or

▶ Click outside the table preview on the canvas pane, and then click and drag to select the area that includes the variables you want to select.

   or

▶ Right-click any variable in a dimension and select Select All [dimension] Variables to select all of the variables in that dimension.

To Reorder Categories

To manually reorder categories:

▶ Select (click) a category in the list.

▶ Click the up or down arrow to move the category up or down in the list.

   or

▶ Click in the Value(s) column for the category, and drag and drop it in a different position.

To Exclude Categories

▶ Select (click) a category in the list.

▶ Click the arrow next to the Exclude list.

or

▶ Click in the Value(s) column for the category and drag and drop it anywhere outside the list.

If you exclude any categories, any categories without defined value labels will also be excluded.

To Sort Categories

You can sort categories by data value, value label, cell count, or summary statistic in ascending or descending order.

▶ In the Sort Categories group, click the By drop-down list and select the sort criterion you want to use: value, label, count, or summary statistic (such as mean, median, or mode). The available summary statistics for sorting depends on the summary statistics you have selected to display in the table.

▶ Click the Order drop-down list to select the sort order (ascending or descending).

Sorting categories is not available if you have excluded any categories.

Subtotals

▶ Select (click) the category in the list that is the last category in the range of categories that you want to include in the subtotal.

▶ Click Add Subtotal....

▶ In the Define Subtotal dialog box, modify the subtotal label text if desired.

▶ To show only a subtotal and suppress the display of the categories that define the subtotal, select Hide subtotaled categories from the table.

▶ Click Continue to add the subtotal.

Totals

▶ Click the Total check box. You can also modify the total label text.

If the selected variable is nested within another variable, totals will be inserted for each subtable.

Display Position for Totals and Subtotals

Totals and subtotals can be displayed above or below the categories included in each total.

■ If Below is selected in the Totals and Subtotals Appear group, totals appear above each subtable, and all categories above and including the selected category (but below any preceding subtotals) are included in each subtotal.

■ If Above is selected in the Totals and Subtotals Appear group, totals appear below each subtable, and all categories below and including the selected category (but above any preceding subtotals) are included in each subtotal.

Important: You should select the display position for subtotals before defining any subtotals. Changing the display position affects all subtotals (not just the currently selected subtotal), and it also *changes the categories included in the subtotals*.

Computed Categories

You can display categories computed from summary statistics, totals, subtotals, and/or constants. For more information, see the topic Computed Categories on p. 38.

Custom Total and Subtotal Summary Statistics

You can display statistics other than "totals" in the Totals and Subtotals areas of the table using the Summary Statistics dialog box. For more information, see the topic Summary Statistics for Categorical Variables on p. 30.

Note: If you select multiple custom total statistics that are also in the body of the table and you hide the statistics labels, then the totals are resorted into the same order as in the body of the table—and since the labels aren't displayed, you may not know what each total statistic actually represents. In general, selecting multiple statistics and hiding the statistics labels is probably not a good idea.

Totals, Subtotals, and Excluded Categories

Cases from excluded categories are not included in the calculation of totals.

Missing Values, Empty Categories, and Values without Value Labels

Missing values. This controls the display of **user-missing** values, or values defined as containing missing values (for example, a code of 99 to represent "not applicable" for pregnancy in males). By default, user-missing values are excluded. Select (check) this option to include user-missing categories in tables. Although the variable may contain more than one missing value category, the table preview on the canvas will display only one generic missing value category. All defined user-missing categories will be included in the table. **System-missing values**(empty cells for numeric variables in the Data Editor) are always excluded.

Empty categories. Empty categories are categories with defined value labels but no cases in that category for a particular table or subtable. By default, empty categories are included in tables. Deselect (uncheck) this option to exclude missing categories from the table.

Other values found when data are scanned. By default, category values in the data file that do not have defined value labels are automatically included in tables. Deselect (uncheck) this option to exclude values without defined value labels from the table. If you exclude any categories with defined value labels, categories without defined value labels are also excluded.

Computed Categories

In addition to displaying the aggregated results of summary statistics, a table can display one or more categories computed from these aggregated results, from constant values, from subtotals and totals, or a combination of them. The results are known as computed categories or postcomputes.

A computed category acts like a category in a single variable with the following similarities and differences:

- A computed category is positioned like the other categories.
- A computed category operates on the same statistics as the other categories.
- Computed categories do not affect subtotals, totals, or significance tests.
- By default, the values of computed categories use the same formatting for summary statistics as the other categories. You can override the format when defining the computed category.

Because computed categories can be used to total aggregated results, they can be similar to subtotals. However, computed categories have the following advantages over subtotals:

- Computed categories can be calculated from the results of other subtotals.
- Computed categories can overlap with each other, operating on the same (or some of the same) categories.
- Computed categories do not have to include values from all other categories above or below the computed category. That is, computed categories are not exhaustive.
- Computed categories can include values from categories that are not adjacent.

Unlike totals and subtotals, computed categories are calculated from the aggregated data rather than the original data. Therefore, the values of computed categories may not match the results of totals and subtotals. Also, because you have the option to hide source categories when defining the computed category, it may be difficult to interpret subtotals in the resulting table. If you use computed categories, it is recommended that you specify custom labels for subtotals.

To Define a Computed Category

Computed categories are added from the Categories and Totals dialog box. For information about accessing that dialog box, see the topic Categories and Totals on p. 35.

▶ In the Categories and Totals dialog box, click Add Category...

Figure 2-11
Define Computed Category dialog box
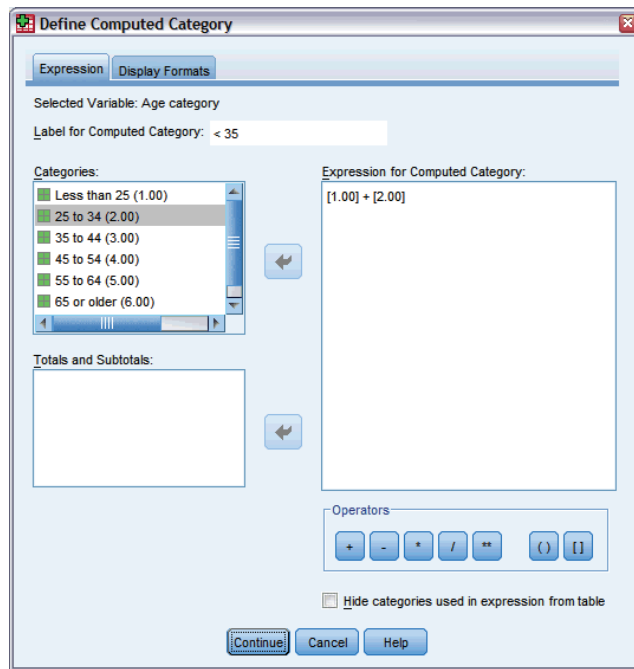
▶ In Label for Computed Category, specify a label for the computed category. You can drag categories from the Categories list to include labels for those categories.

▶ Build an expression by selecting categories and/or totals and subtotals and using operators to define the computed categories. You can also type constant values (e.g., 500) to include in the expression.

▶ To show only a computed category and suppress the display of the categories that define the computed category, select Hide categories used in expression from table.

▶ Click the Display Formats tab to change the display format and number of decimal places for the computed category. For more information, see the topic Display Formats for Computed Categories on p. 40.

▶ Click Continue to add the computed category.

Display Formats for Computed Categories

By default, a computed category uses the same display format and number of decimal places as the other categories in the variable. You can override these on the Display Formats tab in the Computed Category dialog box. The Display Formats tab lists the current summary statistics on which the computed category operates in addition to the display formats and number of decimal places for those statistics.

For each summary statistic, you can:

▶ Select a display format from the Format drop-down list for the summary statistic. For a full list of display formats, see the topic Summary Statistics Display Formats on p. 34.

▶ Enter the number of decimals to display in the Decimals cell for the selected summary statistic.

Tables of Variables with Shared Categories (Comperimeter Tables)

Surveys often contain many questions with a common set of possible responses. You can use stacking to display these related variables in the same table, and you can display the shared response categories in the columns of the table.

To Create a Table for Multiple Variables with Shared Categories

▶ Drag and drop the categorical variables from the variable list into the Rows area of the canvas. The variables should be **stacked**. For more information, see the topic Stacking Variables on p. 26.

▶ From the Category Position drop-down list, select Row labels in columns.

Figure 2-12
Stacked variables with shared response categories in columns

	A great deal	Only some	Hardly any
Confidence in banks & financial institutions	490	1068	306
Confidence in education	511	1055	315
Confidence in major companies	500	1078	243
Confidence in medicine	844	864	167
Confidence in press	176	878	808
Confidence in television	196	936	744

For more information, see the topic Tables for Variables with Shared Categories in Chapter 7 on p. 98.

Customizing the Table Builder

Unlike standard dialog boxes, you can change the size of the table builder in the same way that you can change the size of any standard window:

▶ Click and drag the top, bottom, either side, or any corner of the table builder to decrease or increase its size.

On the Table tab, you can also change the size of the variable list, the Categories list, and the canvas pane.

▶ Click and drag the horizontal bar between the variable list and the Categories list to make the lists longer or shorter. Moving it down makes the variable list longer and the Categories list shorter. Moving it up does the reverse.

▶ Click and drag the vertical bar between the variable list and Categories list from the canvas pane to make the lists wider or narrower. The canvas automatically resizes to fit the remaining space.

Custom Tables: Options Tab

The Options tab allows you to:

■ Specify what is displayed in empty cells and cells for which statistics cannot be computed.

■ Control how missing values are handled in the computation of scale variable statistics.

■ Set minimum and/or maximum data column widths.

■ Control the treatment of duplicate responses in multiple category sets.

Figure 2-13
Custom Tables dialog box, Options tab
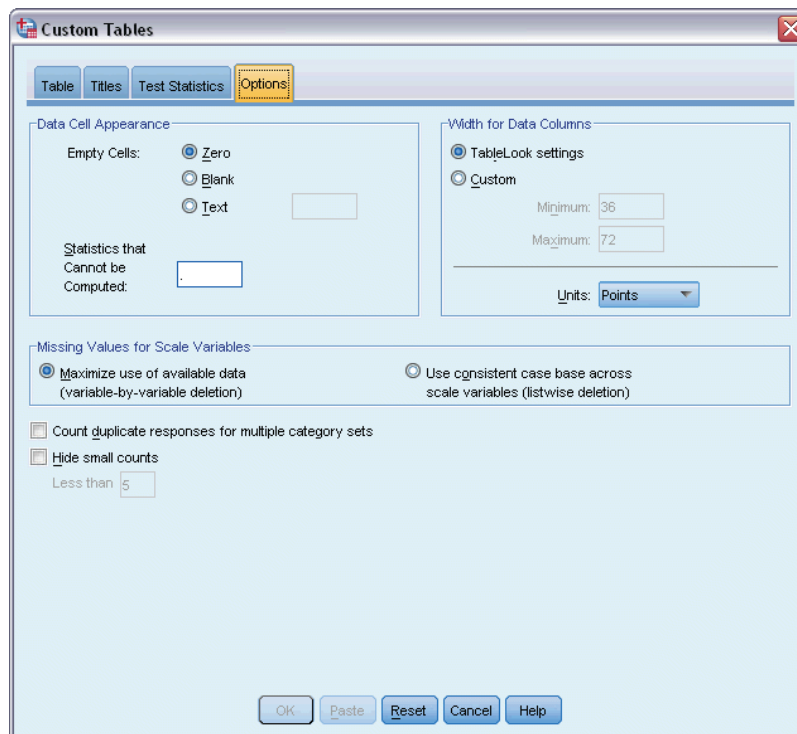
Data Cell Appearance. Controls what is displayed in empty cells and cells for which statistics cannot be computed.

■ **Empty cells.** For table cells that contain no cases (cell count of 0), you can select one of three display options: zero, blank, or a text value that you specify. The text value can be up to 255 characters long.

■ **Statistics that cannot be computed.** Text displayed if a statistic cannot be computed (for example, the mean for a category with no cases). The text value can be up to 255 characters long. The default value is a period (.).

Width for Data Columns. Controls minimum and maximum column width for data columns. This setting does not affect columns widths for row labels.

- **TableLook settings.** Uses the data column width specification from the current default TableLook. You can create your own custom default TableLook to use when new tables are created, and you can control both row label column and data column widths with a TableLook.

- **Custom.** Overrides the default TableLook settings for data column width. Specify the minimum and maximum data column widths for the table and the measurement unit: points, inches, or centimeters.

Missing Values for Scale Variables. For tables with two or more scale variables, controls the handling of missing data for scale variable statistics.

- **Maximize use of available data (variable-by-variable deletion).** All cases with valid values for each scale variable are included in summary statistics for that scale variable.

- **Use consistent case base across scale variables (listwise deletion).** Cases with missing values for any scale variables in the table are excluded from the summary statistics for all scale variables in the table.

Count duplicate responses for multiple category sets. A duplicate response is the same response for two or more variables in the multiple category set. By default, duplicate responses are not counted, but this may be a perfectly valid condition that you do want to include in the count (such as a multiple category set representing the manufacturer of the last three cars purchased by a survey respondent).

Hide small counts. You can choose to hide counts that are less than a specified integer. Hidden values will be displayed as <N, where N is the specified integer. The specified integer must be greater than or equal to 2.

Custom Tables: Titles Tab

The Titles tab controls the display of titles, captions, and corner labels.

Figure 2-14
Custom Tables dialog box, Titles tab
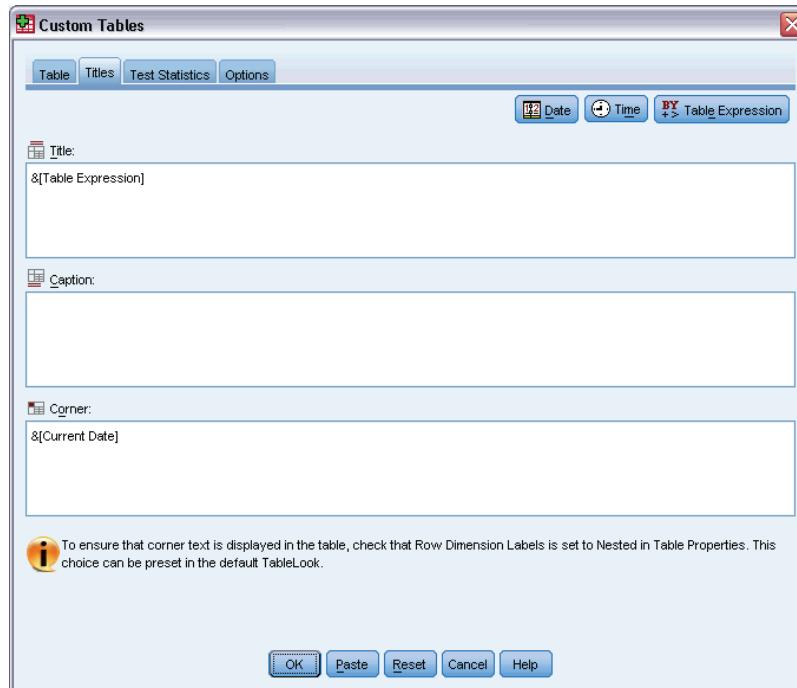
Title. Text that is displayed above the table.

Caption. Text that is displayed below the table and above any footnotes.

Corner. Text that is displayed in the upper left corner of the table. Corner text is displayed only if the table contains row variables and if the pivot table row dimension label property is set to Nested. This is *not* the default TableLook setting.

You can include the following automatically generated values in the table title, caption, or corner label:

Date. Current year, month, and day displayed in a format based on your current Windows Regional Options settings.

Time. Current hour, minute, and second displayed in a format based on your current Windows Regional Options settings.

Table Expression. Variables used in the table and how they are used in the table. If a variable has a defined variable label, the label is displayed. In the generated table, the following symbols indicate how variables are used in the table:

- **+** indicates stacked variables.
- **>** indicates nesting.
- **BY** indicates crosstabulation or layers.

Custom Tables: Test Statistics Tab

The Test Statistics tab allows you to request various significance tests for your custom tables, including:

■ Chi-square tests of independence.

■ Tests of the equality of column means.

■ Tests of the equality of column proportions.

■ Significance tests for multiple response sets and subtotals. (For information on significance testing for multiple response sets, see Significance Testing with Multiple Response Sets on p. 158.)

These tests are not available for tables in which category labels are moved out of their default table dimension.

Figure 2-15
Custom Tables dialog box, Test Statistics tab
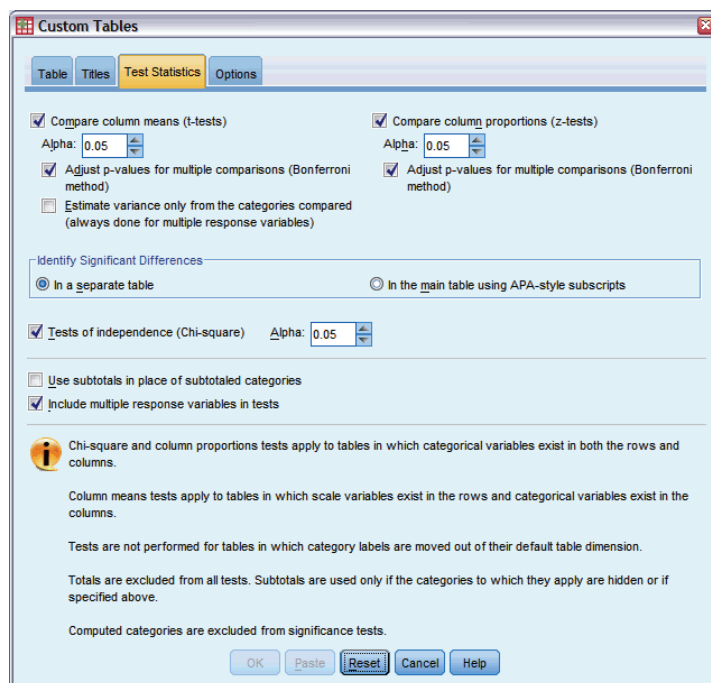
Compare column means (t-tests). This option produces pairwise tests of the equality of column means for tables in which at least one category variable exists in the columns and at least one scale variable exists in the rows. The table must include the mean as a summary statistic. You can select whether the *p* values of the tests are adjusted using the Bonferroni method. You can also specify the alpha level of the test, which should be a value greater than 0 and less than 1. Finally, while the variance for the means test is always based on just the categories compared for multiple response tests, for ordinary categorical variables it can be estimated from just the categories compared or all categories.

Compare column proportions (z-tests). This option produces pairwise tests of the equality of column proportions for tables in which at least one category variable exists in both the columns and rows. The table must include counts or simple column percentages. You can select whether the *p* values of the tests are adjusted using the Bonferroni method. You can also specify the alpha level of the test, which should be a value greater than 0 and less than 1.

Identify Significant Differences. If you select Compare column means or Compare column proportions, you can choose how to indicate significant differences.

■ **In a separate table.** Significance tests results are displayed in a separate table. If two values are significantly different, the cell corresponding to the larger value displays a key identifying the column of the smaller value. Following is an example.

Figure 2-16
Comparisons of column proportions in a separate table

		Marital status				
		Married	Widowed	Divorced	Separated	Never married
		(A)	(B)	(C)	(D)	(E)
Labor force status	Working full time	B		A B	B	B
	Working part-time					A B C
	Temporarily not working					
	Unemployed, laid off				.	A B
	Retired	E	A C D E	E		
	School					A B C
	Keeping house	C E	C E		C E	
	Other					

■ **In the main table using APA-style subscripts.** The main table itself identifies significant differences with APA-style formatting using subscript letters. If two values are significantly different, those values display *different* subscript letters. These subscripts are not footnotes. When this option is in effect, the defined footnote style in the current TableLook is overridden and footnotes are displayed as superscript numbers. Following is an example.

Figure 2-17
Comparisons of column proportions using APA-style subscripts

		Marital status				
		Married	Widowed	Divorced	Separated	Never married
		Column N %	Column N %	Column N %	Column N %	Column N %
Labor force status	Working full time	$57.8\%_a$	$15.5\%_b$	$66.1\%_c$	$62.4\%_{a,c}$	$59.1\%_{a,c}$
	Working part-time	$10.3\%_a$	$7.1\%_a$	$7.8\%_a$	$9.7\%_{a,b}$	$15.4\%_b$
	Temporarily not working	$1.7\%_a$	$.7\%_a$	$2.0\%_a$	$1.1\%_a$	$1.7\%_a$
	Unemployed, laid off	$1.0\%_a$	$1.1\%_a$	$2.2\%_{a,b}$	$.0\%$	$4.8\%_b$
	Retired	$12.5\%_a$	$53.0\%_b$	$11.9\%_a$	$6.5\%_{a,c}$	$2.6\%_c$
	School	$.7\%_a$	$.4\%_a$	$1.6\%_a$	$2.2\%_{a,b}$	$9.0\%_b$
	Keeping house	$14.9\%_a$	$19.4\%_a$	$5.6\%_b$	$14.0\%_a$	$5.3\%_b$
	Other	$1.2\%_a$	$2.8\%_a$	$2.7\%_a$	$4.3\%_a$	$2.1\%_a$

For the full example that demonstrates how to create and interpret these tables, see Comparing Column Proportions on p. 142.

Note: The APA-style table includes a caption that explains how to interpret the subscripts in the table. If you specify a caption on the Titles tab, the caption specified on the Titles tab will be displayed instead of the caption that explains the subscripts.

Tests of independence (chi-square). This option produces a chi-square test of independence for tables in which at least one category variable exists in both the rows and columns. You can also specify the alpha level of the test, which should be a value greater than 0 and less than 1.

Use subtotals in place of subtotaled categories. If selected, then each subtotal replaces its categories for significance testing. Otherwise, only subtotals for which the subtotaled categories are hidden replace their categories for testing.

Include multiple response variables. If selected, tests are performed using categories of multiple response sets. Otherwise multiple response sets are ignored when performing tests.

Simple Tables for Categorical Variables

Most tables you want to create will probably include at least one **categorical variable**. A categorical variable is one with a limited number of distinct values or categories (for example, gender or religion). Categorical variables can be either **nominal** or **ordinal**.

- **Nominal.** A variable can be treated as nominal when its values represent categories with no intrinsic ranking (for example, the department of the company in which an employee works). Examples of nominal variables include region, zip code, and religious affiliation.

- **Ordinal.** A variable can be treated as ordinal when its values represent categories with some intrinsic ranking (for example, levels of service satisfaction from highly dissatisfied to highly satisfied). Examples of ordinal variables include attitude scores representing degree of satisfaction or confidence and preference rating scores.

An icon next to each variable in the variable list identifies the variable type.

Measurement Level	Data Type			
	Numeric	String	Date	Time
Scale (Continuous)		n/a		
Ordinal				
Nominal				

Custom Tables is optimized for use with categorical variables that have defined **value labels**. For more information, see the topic Building Tables in Chapter 2 on p. 22.

Sample Data File

The examples in this chapter use the data file *survey_sample.sav*. For more information, see the topic Sample Files in Appendix A on p. 178.

All examples provided here display variable labels in dialog boxes, sorted in alphabetical order. Variable list display properties are set on the General tab in the Options dialog box (Edit menu, Options).

A Single Categorical Variable

Although a table of a single categorical variable may be one of the simplest tables you can create, it may often be all you want or need.

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ In the table builder, drag and drop *Age category* from the variable list to the Rows area on the canvas pane.

A preview of the table is displayed on the canvas pane. The preview doesn't display actual data values; it displays only placeholders where data will be displayed.

Figure 3-1
Single categorical variable in rows in table builder canvas preview
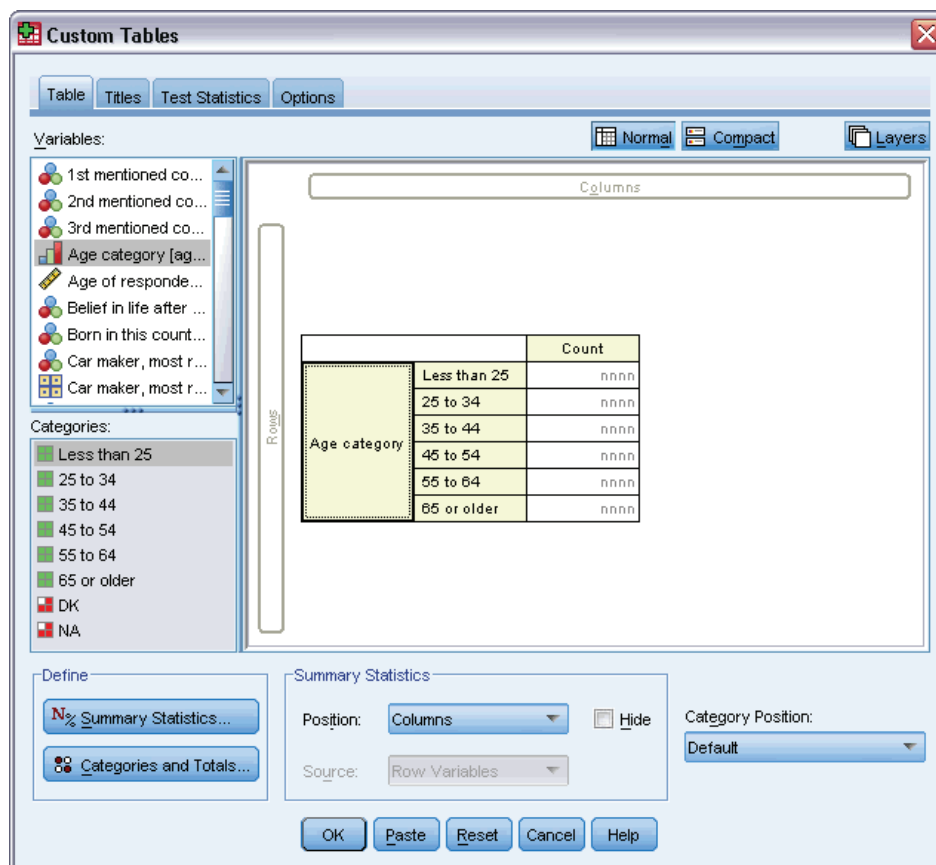
▶ Click OK to create the table.

The table is displayed in the Viewer window.

Figure 3-2
Single categorical variable in rows

		Count
Age category	Less than 25	242
	25 to 34	627
	35 to 44	679
	45 to 54	481
	55 to 64	320
	65 or older	479

In this simple table, the column heading *Count* isn't really necessary, and you can create the table without this column heading.

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ In the Summary Statistics group, select (click) Hide for Position.

▶ Click OK to create the table.

Figure 3-3
Single categorical variable without summary statistics column label

Age category	Less than 25	242
	25 to 34	627
	35 to 44	679
	45 to 54	481
	55 to 64	320
	65 or older	479

Percentages

In addition to counts, you can also display percentages. For a simple table of a single categorical variable, if the variable is displayed in rows, you probably want to look at column percentages. Conversely, for a variable displayed in columns, you probably want to look at row percentages.

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ In the Summary Statistics group, deselect (uncheck) Hide for Position. Since this table will have two columns, you want to display the column labels so you know what each column represents.

▶ Right-click *Age category* on the canvas pane and select Summary Statistics from the pop-up context menu.

Figure 3-4
Right-click context menu on canvas pane

▶ In the Summary Statistics dialog box, select Column N % in the Statistics list and click the arrow to add it to the Display list.

▶ In the Label cell in the Display list, delete the default label and type Percent.

Figure 3-5
Summary Statistics Categorical Variables dialog box
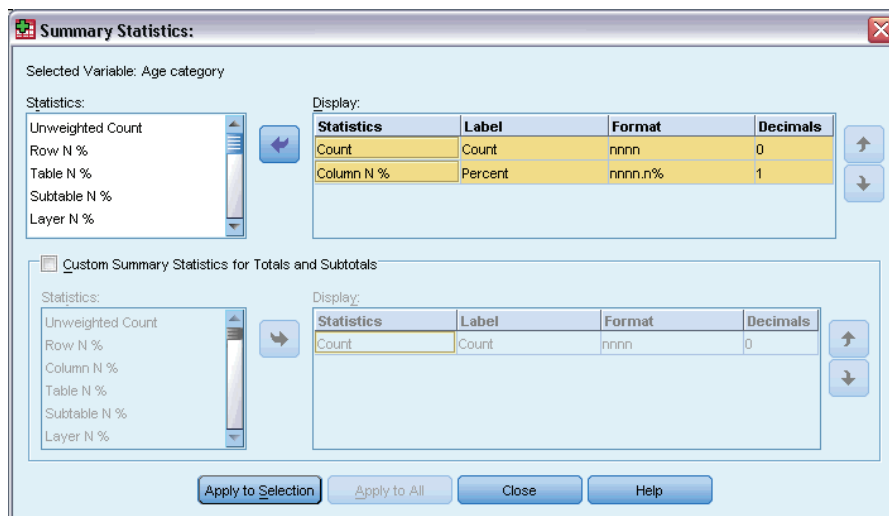
▶ Click Apply to Selection and then click OK in the table builder to create the table.

Figure 3-6
Counts and column percentages

		Count	Percent
Age category	Less than 25	242	8.6%
	25 to 34	627	22.2%
	35 to 44	679	24.0%
	45 to 54	481	17.0%
	55 to 64	320	11.3%
	65 or older	479	16.9%

Totals

Totals are not automatically included in custom tables, but it's easy to add totals to a table.

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Right-click *Age category* on the canvas pane and select Categories and Totals from the pop-up context menu.

▶ Select (click) Total in the Categories and Totals dialog box.

Figure 3-7
Categories and Totals dialog box
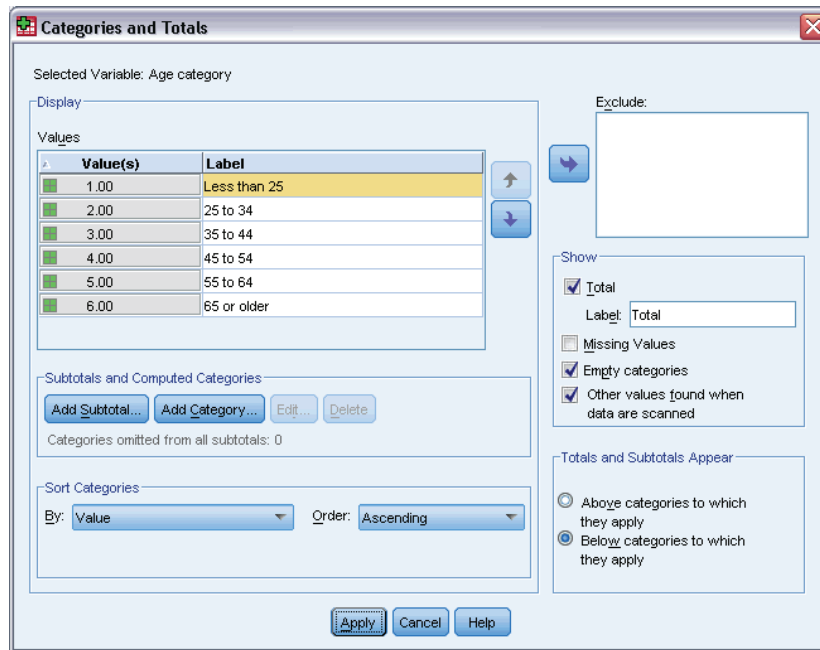
▶ Click Apply and then click OK in the table builder to create the table.

Figure 3-8
Counts, column percentages, and totals

		Count	Percent
Age category	Less than 25	242	8.6%
	25 to 34	627	22.2%
	35 to 44	679	24.0%
	45 to 54	481	17.0%
	55 to 64	320	11.3%
	65 or older	479	16.9%
	Total	2828	100.0%

For more information, see the topic Totals and Subtotals for Categorical Variables in Chapter 5 on p. 75.

Crosstabulation

Crosstabulation is a basic technique for examining the relationship between two categorical variables. For example, using *Age category* as a row variable and *Gender* as a column variable, you can create a two-dimensional crosstabulation that shows the number of males and females in each age category.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Reset to delete any previous selections in the table builder.

▶ In the table builder, drag and drop *Age category* from the variable list to the Rows area on the canvas pane.

▶ Drag and drop *Gender* from the variable list to the Columns area on the canvas pane. (You may have to scroll down through the variable list to find this variable.)

Figure 3-9
Crosstabulation in table builder canvas preview
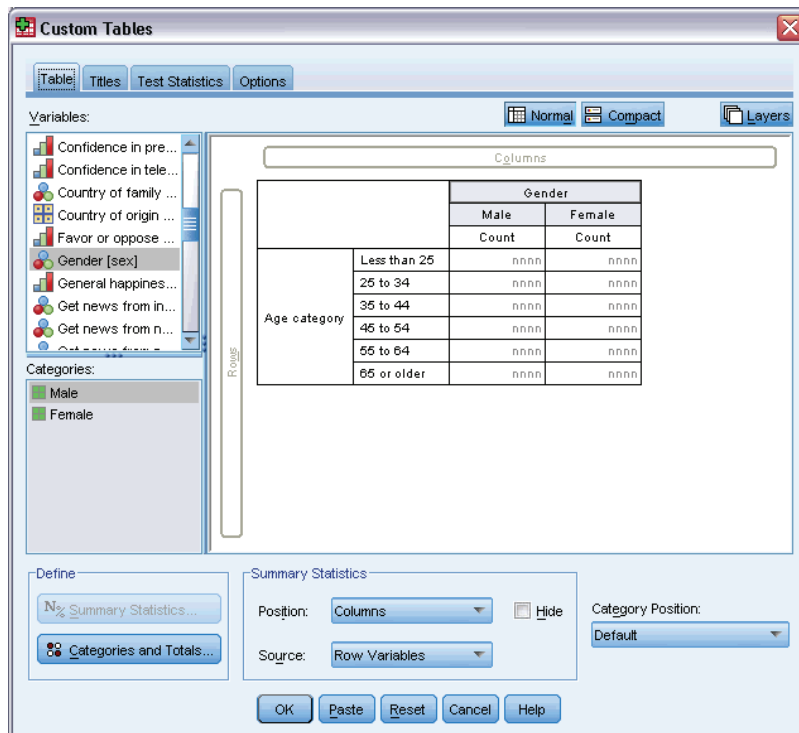
▶ Click OK to create the table.

Figure 3-10
Crosstabulation of Age category and Gender

		Gender	
		Male	Female
		Count	Count
Age category	Less than 25	108	134
	25 to 34	276	351
	35 to 44	309	370
	45 to 54	221	260
	55 to 64	136	184
	65 or older	178	301

Percentages in Crosstabulations

In a two-dimensional crosstabulation, both row and column percentages may provide useful information.

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Right-click *Gender* on the canvas pane.

You may notice that Summary Statistics is disabled in the pop-up context menu. This is because you can select summary statistics only for the innermost variable in the statistics source dimension. The default statistics source dimension (row or column) for categorical variables is based on the order in which you drag and drop variables onto the canvas pane. In this example, we dragged *Age category* to the rows dimension first—and since there aren't any other variables in the rows dimension, *Age category* is the statistics source variable. You can change the statistics source dimension, but in this example, you don't need to do that. For more information, see the topic Summary Statistics in Chapter 2 on p. 29.

▶ Right-click *Age category* on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select Column N % in the Statistics list and click the arrow to add it to the Display list.

▶ Select Row N % in the Statistics list and click the arrow to add it to the Display list.

▶ Click Apply to Selection and then click OK in the table builder to create the table.

Figure 3-11
Crosstabulation with row and column percentages

| | | Gender | | | | | |
| | | Male | | | Female | | |
		Count	Column N %	Row N %	Count	Column N %	Row N %
Age category	Less than 25	108	8.8%	44.6%	134	8.4%	55.4%
	25 to 34	276	22.5%	44.0%	351	21.9%	56.0%
	35 to 44	309	25.2%	45.5%	370	23.1%	54.5%
	45 to 54	221	18.0%	45.9%	260	16.3%	54.1%
	55 to 64	136	11.1%	42.5%	184	11.5%	57.5%
	65 or older	178	14.5%	37.2%	301	18.8%	62.8%

Controlling Display Format

You can control the display format, including the number of decimals displayed in summary statistics. For example, by default, percentages are displayed with one decimal and a percent sign. But what if you want the cell values to show two decimals and no percent sign?

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click *Age category* on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ For the two selected percentage summary statistics (Column N % and Row N %), select nnnn.n from the Format drop-down list and type 2 in the Decimals cell for both of them.

Figure 3-12
Summary Statistics dialog box
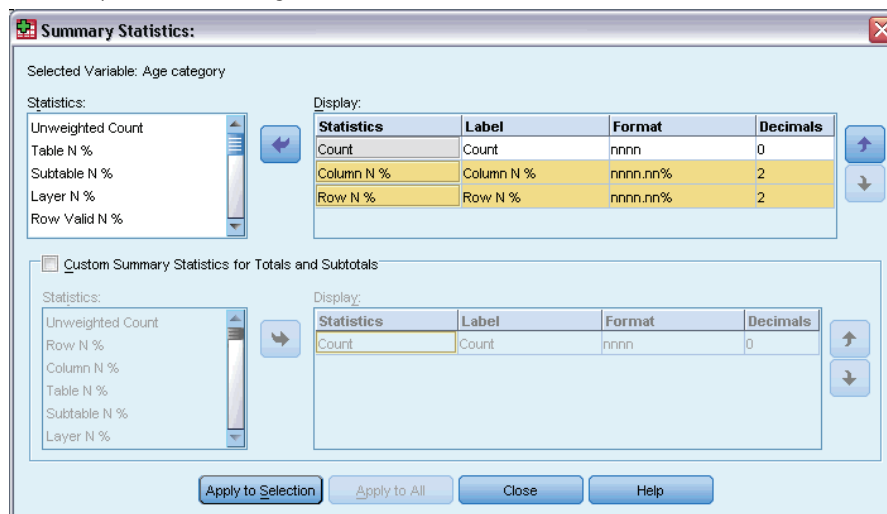
▶ Click OK to create the table.

Figure 3-13
Formatted cell display for row and column percentages

		Gender					
		Male			Female		
		Count	Column N %	Row N %	Count	Column N %	Row N %
Age category	Less than 25	108	8.79	44.63	134	8.38	55.37
	25 to 34	276	22.48	44.02	351	21.94	55.98
	35 to 44	309	25.16	45.51	370	23.13	54.49
	45 to 54	221	18.00	45.95	260	16.25	54.05
	55 to 64	136	11.07	42.50	184	11.50	57.50
	65 or older	178	14.50	37.16	301	18.81	62.84

Marginal Totals

It's fairly common in crosstabulations to display **marginal totals**—totals for each row and column. Since these aren't included in Custom Tables by default, you need to explicitly add them to your tables.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Reset to delete any previous selections in the table builder.

▶ In the table builder, drag and drop *Age category* from the variable list to the Rows area on the canvas pane.

▶ Drag and drop *Gender* from the variable list to the Columns area on the canvas pane. (You may have to scroll down through the variable list to find this variable.)

▶ Right-click *Age category* on the canvas pane and select Categories and Totals from the pop-up context menu.

▶ Select (click) Total in the Categories and Totals dialog box and then click Apply.

▶ Right-click *Gender* on the canvas pane and select Categories and Totals from the pop-up context menu.

▶ Select (click) Total in the Categories and Totals dialog box and then click Apply.

▶ In the Summary Statistics group, select (click) Hide for Position. (Since you're displaying only counts, you don't need to identify the "statistic" displayed in the data cells of the table.)

▶ Click OK to create the table.

Figure 3-14
Crosstabulation with marginal totals

| | | Gender | | |
		Male	Female	Total
Age category	Less than 25	108	134	242
	25 to 34	276	351	627
	35 to 44	309	370	679
	45 to 54	221	260	481
	55 to 64	136	184	320
	65 or older	178	301	479
	Total	1228	1600	2828

Sorting and Excluding Categories

By default, categories are displayed in the ascending order of the data values that the category value labels represent. For example, although value labels of *Less than 25*, *25 to 34*, *35 to 44*, *...*, etc., are displayed for age categories, the actual underlying data values are 1, 2, 3, ..., etc., and it is those underlying data values that control the default display order of the categories.

You can easily change the order of the categories and also exclude categories that you don't want to be displayed in the table.

Sorting Categories

You can manually rearrange categories or sort categories in ascending or descending order of:

■ Data values.

■ Value labels.

■ Cell counts.

■ Summary statistics. The available summary statistics for sorting depends on the summary statistics you have selected to display in the table.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ If *Age category* isn't already displayed in the Rows area on the canvas pane, drag and drop it there.

► Right-click *Age category* on the canvas pane and select Categories and Totals from the pop-up context menu.

Both data values and the associated value labels are displayed in the current sort order, which in this case is still ascending order of data values.

Figure 3-15
Default category order, ascending by data values
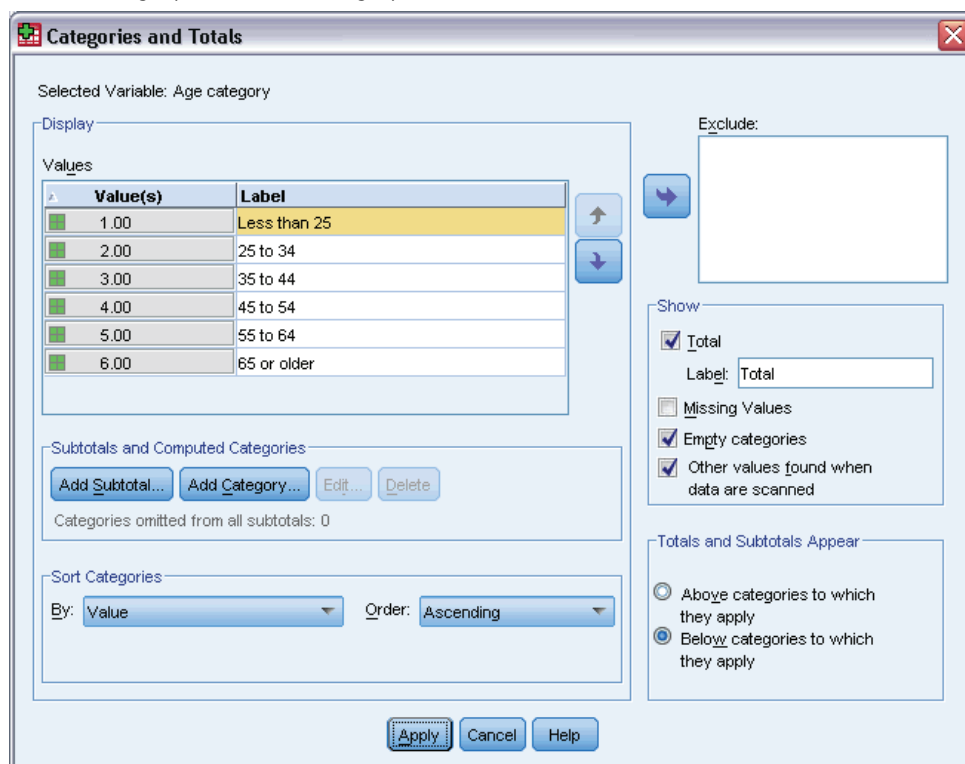
► In the Sort Categories group, select Descending from the Order drop-down list.

The sort order is now reversed.

► Select Labels from the By drop-down list.

The categories are now sorted in descending alphabetical order of the value labels.

Figure 3-16
Descending alphabetical sort order
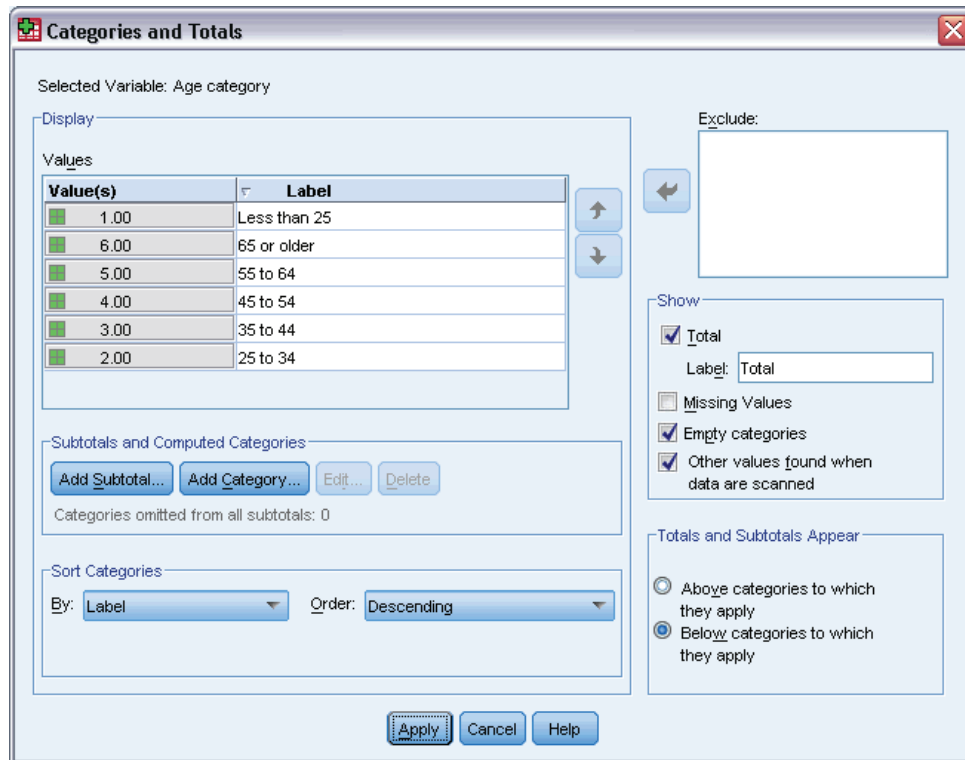
Notice that the category labeled *Less than 25* is at the top of the list. In alphabetical sorting, letters come after numbers. Since this is the only label that starts with a letter and since the list is sorted in descending (reverse) order, this category sorts to the top of the list.

If you want a particular category to appear at a different location in the list, you can easily move it.

▶ Click the category labeled *Less than 25* in the Label list.

▶ Click the down arrow to the right of the list. The category moves down one row in the list.

▶ Keep clicking the down arrow until the category is at the bottom of the list.

Figure 3-17
Manually arranged categories
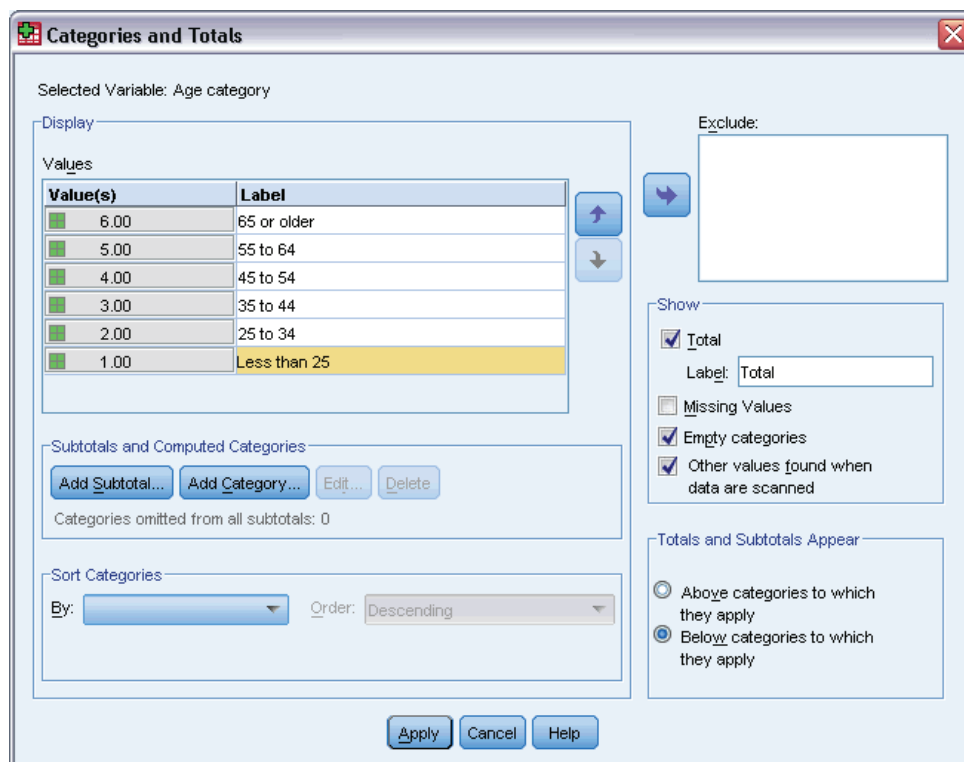
Excluding Categories

If there are some categories that you don't want to appear in the table, you can exclude them.

▶ Click the category labeled *Less than 25* in the Label list.

▶ Click the arrow key to the left of the Exclude list.

▶ Click the category labeled *65 or older* in the Label list.

▶ Click the arrow key to the left of the Exclude list again.

The two categories are moved from the Display list to the Exclude list. If you change your mind, you can easily move them back to the Display list.

Figure 3-18
Manually excluded categories
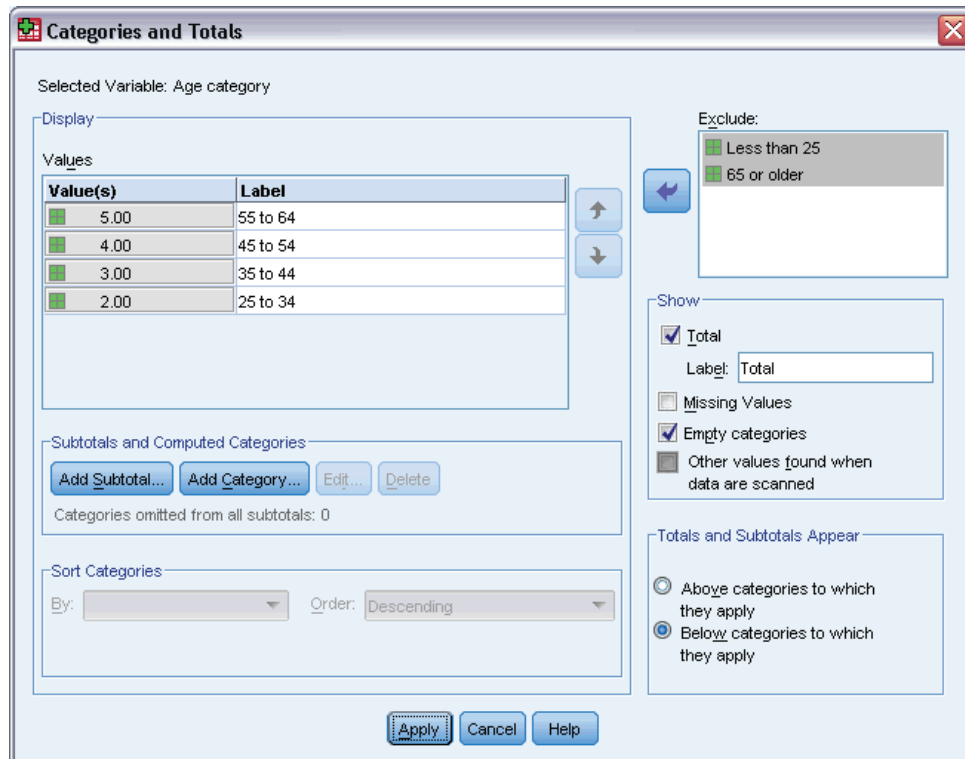
▶ Click Apply and then click OK in the table builder to create the table.

Figure 3-19
Table sorted by descending value label, some categories excluded

		Gender		
		Male	Female	Total
Age category	55 to 64	136	184	320
	45 to 54	221	260	481
	35 to 44	309	370	679
	25 to 34	276	351	627
	Total	942	1165	2107

Notice that the totals are lower than they were before the two categories were excluded. This is because totals are based on the categories included in the table. Any excluded categories are excluded from the total calculation. For more information, see the topic Totals and Subtotals for Categorical Variables in Chapter 5 on p. 75.

Stacking, Nesting, and Layers with Categorical Variables

Stacking, nesting, and layers are all methods for displaying multiple variables in the same table. This chapter focuses on using these techniques with categorical variables, although they can also be used with scale variables.

Sample Data File

The examples in this chapter use the data file *survey_sample.sav*. For more information, see the topic Sample Files in Appendix A on p. 178.

All examples provided here display variable labels in dialog boxes, sorted in alphabetical order. Variable list display properties are set on the General tab in the Options dialog box (Edit menu, Options).

Stacking Categorical Variables

Stacking can be thought of as taking separate tables and pasting them together into the same display. For example, you could display information on *Gender* and *Age category* in separate sections of the same table.

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ In the table builder, drag and drop *Gender* from the variable list to the Rows area on the canvas pane.

▶ Drag and drop *Age category* from the variable list to the Rows area below *Gender*.

The two variable are now stacked in the row dimension.

Figure 4-1
Stacked categorical variables displayed on the canvas pane
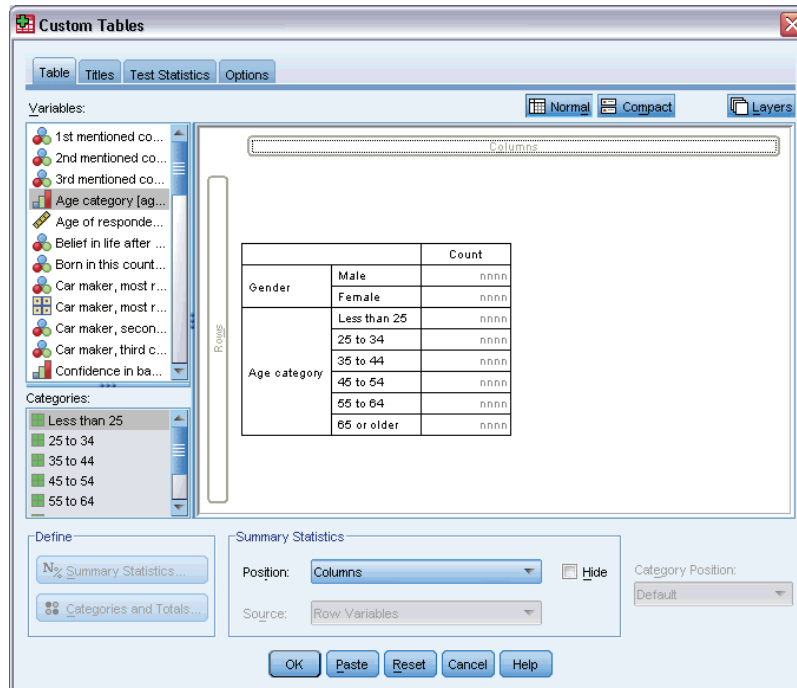
▶ Click OK to create the table.

Figure 4-2
Table of categorical variables stacked in rows

		Count
Gender	Male	1232
	Female	1600
Age category	Less than 25	242
	25 to 34	627
	35 to 44	679
	45 to 54	481
	55 to 64	320
	65 or older	479

You can also stack variables in columns in a similar fashion.

Stacking with Crosstabulation

A stacked table can include other variables in other dimensions. For example, you could crosstabulate two variables stacked in the rows with a third variable displayed in the column dimension.

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ If *Age category* and *Gender* aren't already stacked in the rows, follow the directions above for stacking them.

▶ Drag and drop *Get news from internet* from the variable list to the Columns area on the canvas pane.

▶ Click OK to create the table.

Figure 4-3
Two stacked row variables crosstabulated with a column variable

| | | Get news from internet | |
| | | No | Yes |
		Count	Count
Gender	Male	873	359
	Female	1092	508
Age category	Less than 25	146	96
	25 to 34	368	259
	35 to 44	435	244
	45 to 54	346	135
	55 to 64	252	68
	65 or older	416	63

Note: There are several variables with labels that start with *Get news from ...*, so it may be difficult to distinguish between them in the variable list (since the labels may be too wide to be displayed completely in the variable list). There are two ways to see the entire variable label:

■ Position the mouse pointer on a variable in the list to display the entire label in a pop-up ToolTip.

■ Click and drag the vertical bar that separates the variable and Categories lists from the canvas pane to make the lists wider.

Figure 4-4
Variable list widened to display more complete variable labels
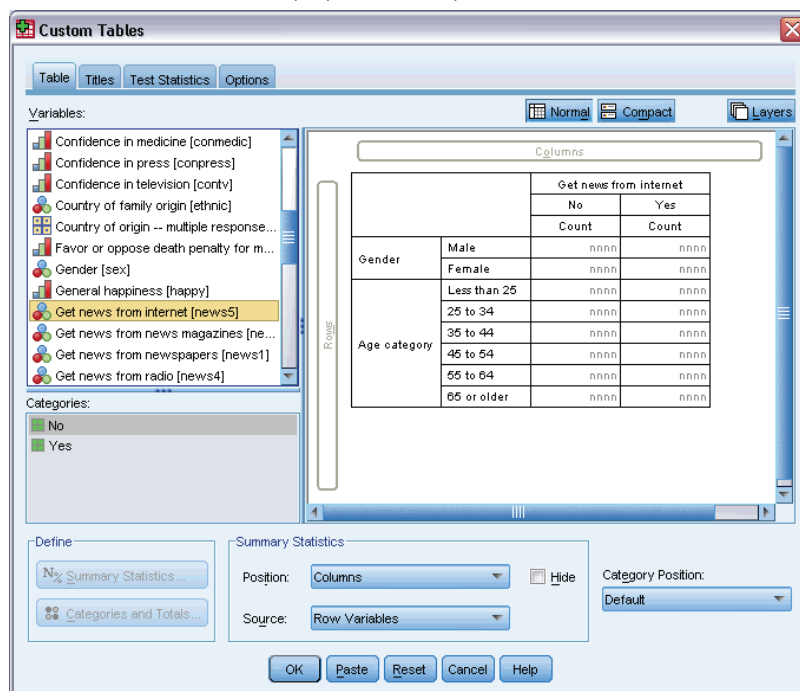
Nesting Categorical Variables

Nesting, like crosstabulation, can show the relationship between two categorical variables, except that one variable is nested within the other in the same dimension. For example, you could nest *Gender* within *Age category* in the row dimension, showing the number of males and females in each age category.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Reset to delete any previous selections in the table builder.

▶ In the table builder, drag and drop *Age category* from the variable list to the Rows area on the canvas pane.

▶ Drag and drop *Gender* from the variable list to the right of *Age category* in the Rows area.

The preview on the canvas pane now shows that the nested table will contain a single column of counts, with each cell containing the number of males or females in each age category.

Figure 4-5
Gender nested within Age category
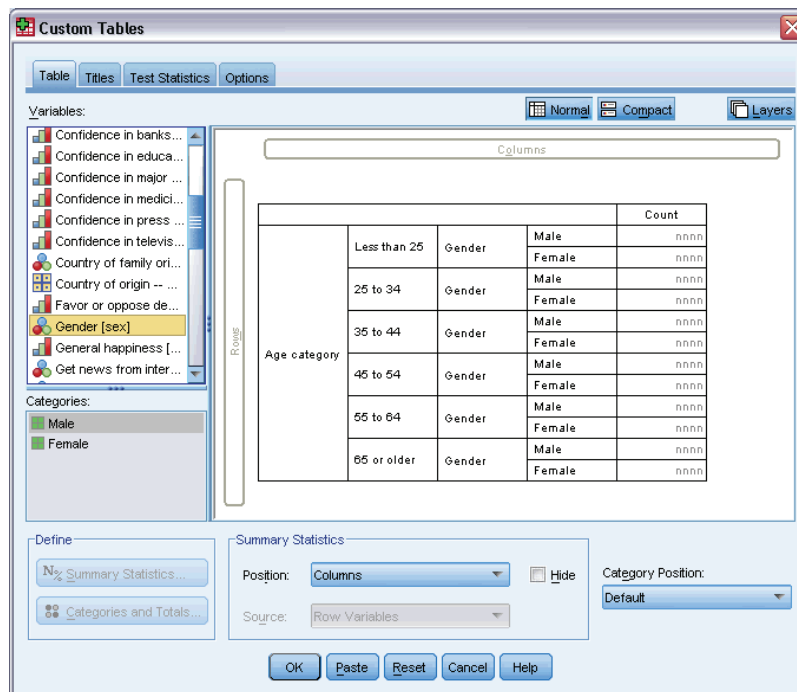
You may notice that the variable label *Gender* is displayed repeatedly, once for each age category. You can minimize this kind of repetition by placing the variable with the fewest categories at the outermost level of the nesting.

▶ Click the variable label *Gender* on the canvas pane.

▶ Drag and drop the variable as far to the left in the Rows area as you can.

Now instead of *Gender* being repeated six times, *Age category* is repeated twice. This is a less-cluttered table that will produce essentially the same results.

Figure 4-6
Age category nested within Gender in table builder preview
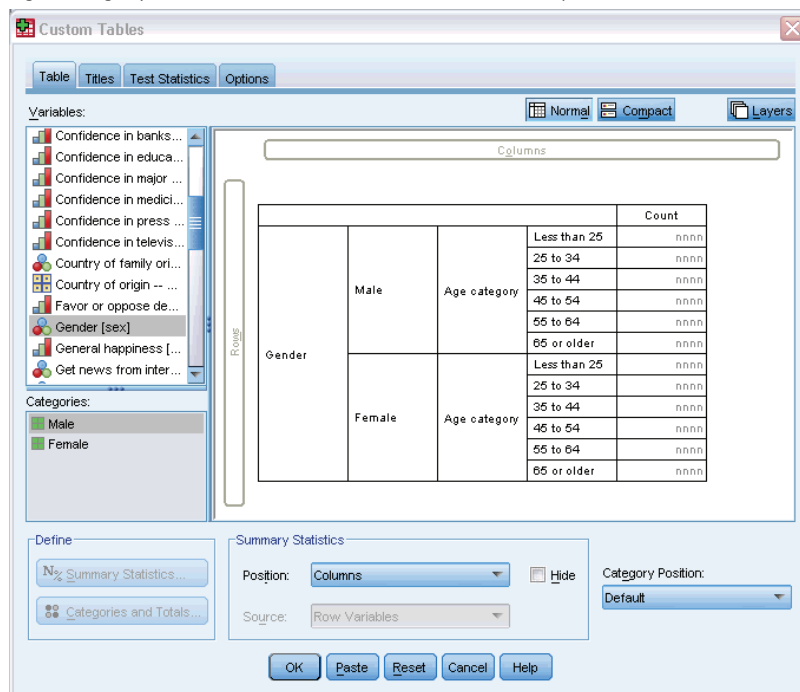
▶ Click OK to create the table.

Figure 4-7
Table of Age category nested within Gender

				Count
Gender	Male	Age category	Less than 25	108
			25 to 34	276
			35 to 44	309
			45 to 54	221
			55 to 64	136
			65 or older	178
	Female	Age category	Less than 25	134
			25 to 34	351
			35 to 44	370
			45 to 54	260
			55 to 64	184
			65 or older	301

Note: Custom Tables do not honor layered split file processing. To achieve the same result as layered split files, place the split file variables in the outermost nesting layers of the table.

Suppressing Variable Labels

Another solution to redundant variable labels in nested tables is simply to suppress the display of variable names or labels. Since the value labels for both *Gender* and *Age category* are probably sufficiently descriptive without the variable labels, we can eliminate the labels for both variables.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click *Age category* on the canvas pane and deselect (uncheck) Show Variable Label on the pop-up context menu.

▶ Do the same for *Gender*.

Figure 4-8
Suppressing variable labels via the context menu in the table builder
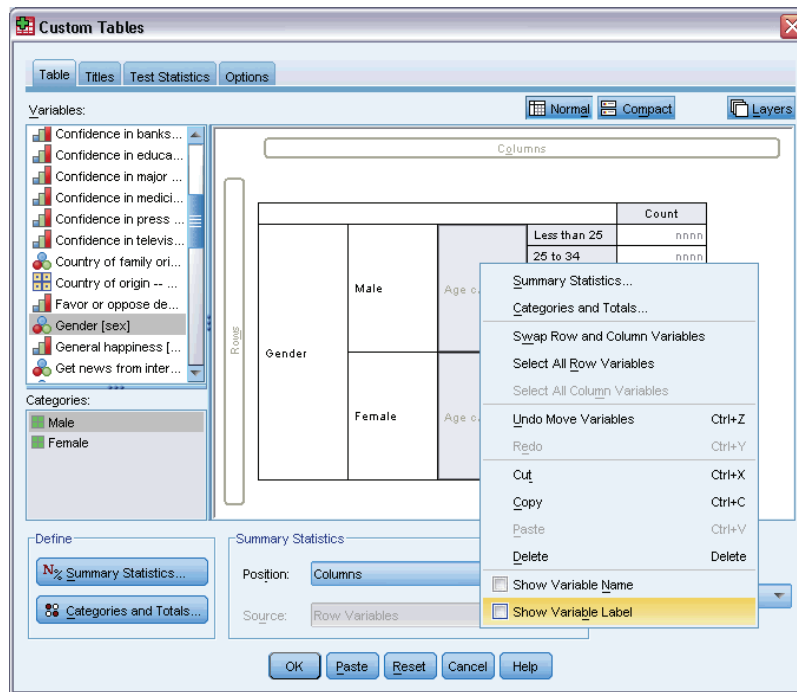
The variable labels are still displayed in the table preview, but they won't be included in the table.

▶ Click OK to create the table.

Figure 4-9
Nested table without variable labels

		Count
Male	Less than 25	108
	25 to 34	276
	35 to 44	309
	45 to 54	221
	55 to 64	136
	65 or older	178
Female	Less than 25	134
	25 to 34	351
	35 to 44	370
	45 to 54	260
	55 to 64	184
	65 or older	301

If you want the variable labels included with the table somewhere—without displaying them multiple times in the body of the table—you can include them in the table title or corner label.

► Open the table builder (Analyze menu, Tables, Custom Tables).

► Click the Titles tab.

► Click anywhere in the Title text box.

► Click Table Expression. The text *&[Table Expression]* is displayed in the Title text box. This will generate a table title that includes the variable labels for the variables used in the table.

► Click OK to create the table.

Figure 4-10
Variable labels in table title

Gender > Age category

		Count
Male	Less than 25	108
	25 to 34	276
	35 to 44	309
	45 to 54	221
	55 to 64	136
	65 or older	178
Female	Less than 25	134
	25 to 34	351
	35 to 44	370
	45 to 54	260
	55 to 64	184
	65 or older	301

The greater than sign (>) in the title indicates that *Age category* is nested within *Gender*.

Nested Crosstabulation

A nested table can contain other variables in other dimensions. For example, you could nest *Age category* within *Gender* in the rows and crosstabulate the nested rows with a third variable in the column dimension.

► Open the table builder (Analyze menu, Tables, Custom Tables).

▶ If *Age category* isn't already nested within *Gender* in the rows, follow the directions above for nesting them.

▶ Drag and drop *Get news from internet* from the variable list to the Columns area on the canvas pane.

You may notice that the table is too large to display completely on the canvas pane. You can scroll up/down or right/left on the canvas pane to see more of the table preview, or you can:

- Click Compact in the table builder to see a compact view. This displays only the variable labels, without any information on categories or summary statistics included in the table.

- Increase the size of the table builder by clicking and dragging any of the sides or corners of the table builder.

Figure 4-11
Compact view on the canvas pane
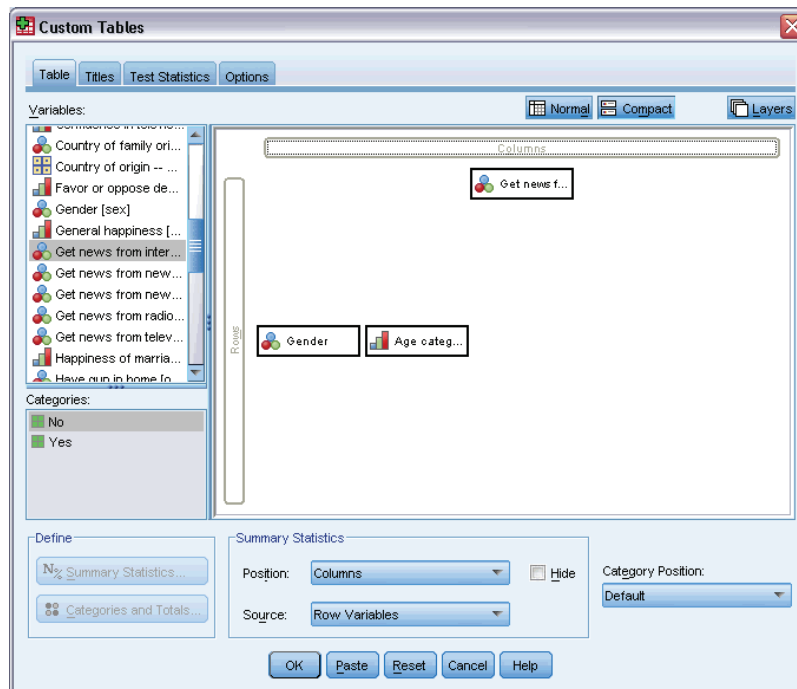
▶ Click OK to create the table.

Figure 4-12
Nested crosstabulation

| | | | | Get news from internet | |
| | | | | No | Yes |
				Count	Count
Gender	Male	Age category	Less than 25	59	49
			25 to 34	159	117
			35 to 44	217	92
			45 to 54	169	52
			55 to 64	112	24
			65 or older	155	23
	Female	Age category	Less than 25	87	47
			25 to 34	209	142
			35 to 44	218	152
			45 to 54	177	83
			55 to 64	140	44
			65 or older	261	40

Swapping Rows and Columns

What do you do if you spend a lot of time setting up a complex table and then decide it's absolutely perfect—except that you want to switch the orientation, putting all of the row variables in the columns and vice versa? For example, you've created a nested crosstabulation with *Age category* and *Gender* nested in the rows, but now you want these two demographic variables nested in the columns instead.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click anywhere on the canvas pane and select Swap Row and Column Variables from the pop-up context menu.

Figure 4-13
Swapping row and column variables
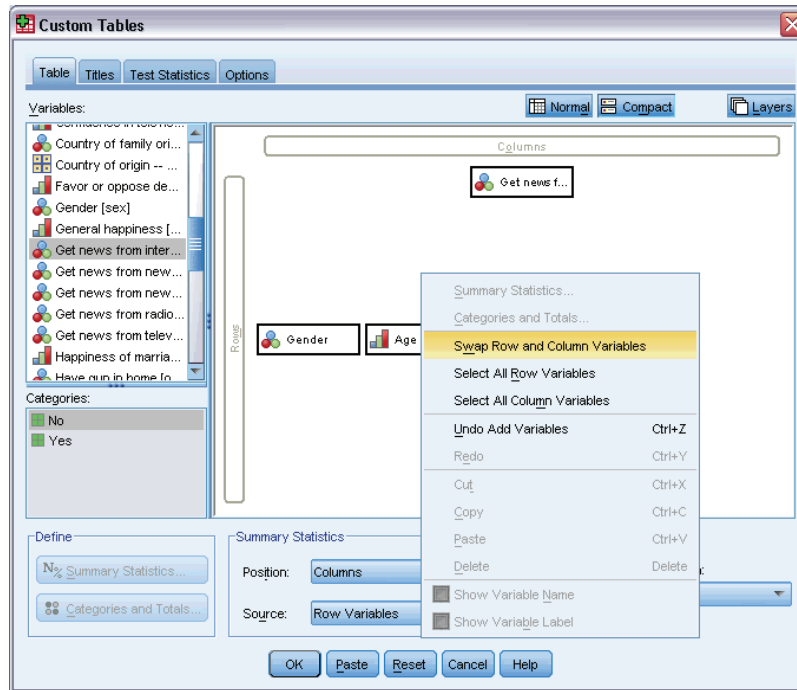
The row and column variables have now been switched.

Before creating the table, let's make a few modifications to make the display less cluttered.

▶ Select Hide to suppress the display of the summary statistics column label.

▶ Right-click *Gender* on the canvas pane and deselect (uncheck) Show Variable Label.

▶ Now click OK to create the table.

Figure 4-14
Crosstabulation with demographic variables nested in columns

		Male						Female					
		Age category						Age category					
		Less than 25	25 to 34	35 to 44	45 to 54	55 to 64	65 or older	Less than	25 to 34	35 to 44	45 to 54	55 to 64	65 or older
Get news from internet	No	59	159	217	169	112	155	87	209	218	177	140	261
	Yes	49	117	92	52	24	23	47	142	152	83	44	40

Layers

You can use layers to add a dimension of depth to your tables, creating three-dimensional "cubes." Layers are, in fact, quite similar to nesting or stacking; the primary difference is that only one layer category is visible at a time. For example, using *Age category* as the row variable and *Gender* as a layer variable produces a table in which information for males and females is displayed in different layers of the table.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Reset to delete any previous selections in the table builder.

▶ In the table builder, drag and drop *Age category* from the variable list to the Rows area on the canvas pane.

▶ Click Layers at the top of the table builder to display the Layers list.

▶ Drag and drop *Gender* from the variable list to the Layers list.

Figure 4-15
Age category in rows, Gender in layers
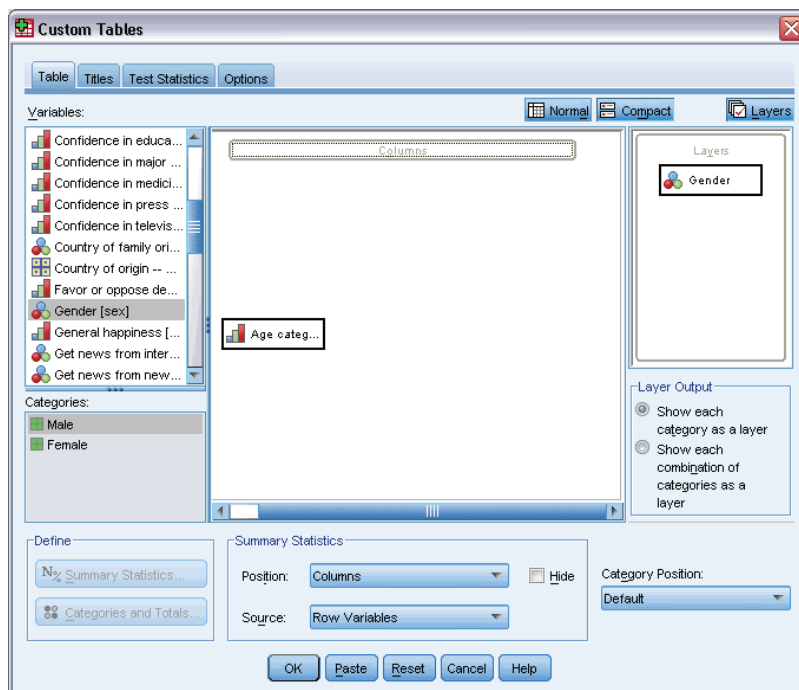
At this point, you might notice that adding a layer variable has no visible effect on the preview displayed on the canvas pane. Layer variables do not affect the preview on the canvas pane unless the layer variable is the statistics source variable and you change the summary statistics.

▶ Click OK to create the table.

Figure 4-16
Simple layered table

Gender Male

		Count
Age category	Less than 25	108
	25 to 34	276
	35 to 44	309
	45 to 54	221
	55 to 64	136
	65 or older	178

At first glance, this table doesn't look any different than a simple table of a single categorical variable. The only difference is the presence of the label *Gender Male* at the top of the table.

▶ Double-click the table in the Viewer window to activate it.

▶ You can now see that the label *Gender Male* is actually a choice in a drop-down list.

▶ Click the down arrow on the drop-down list to display the whole list of layers.

Figure 4-17
List of layers in activated pivot table

In this table, there is only one other choice in the list.

▶ Select *Gender Female* from the drop-down list.

Figure 4-18
Simple layered table with different layer displayed

Two Stacked Categorical Layer Variables

If you have more than one categorical variable in the layers, you can either stack or nest the layer variables. By default, layer variables are stacked. (*Note*: If you have any scale layer variables, layer variables can only be stacked.)

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ If you don't already have *Age category* in the rows and *Gender* in the layers, follow the directions above for creating a layered table.

▶ Drag and drop *Highest degree* from the variable list to the Layer list below *Gender*.

Figure 4-19
Stacked layer variables in table builder
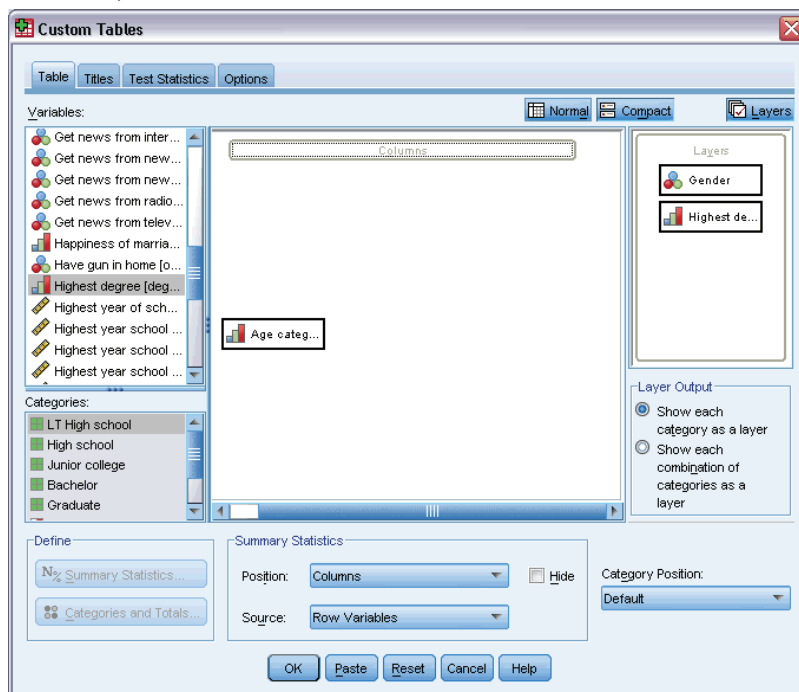
The two radio buttons below the Layer list in the Layer Output group are now activated. The default selection is Show each category as a layer. This is equivalent to stacking.

▶ Click OK to create the table.

▶ Double-click the table in the Viewer window to activate it.

▶ Click the down arrow on the drop-down list to display the whole list of layers.

Figure 4-20
List of stacked layers in activated pivot table

There are seven layers in the table: two layers for the two *Gender* categories and five layers for the five *Highest degree* categories. For stacked layers, the total number of layers is the sum of the number of categories for the layer variables (including any total or subtotal categories you have requested for the layer variables).

Two Nested Categorical Layer Variables

Nesting categorical layer variables creates a separate layer for each combination of layer variable categories.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ If you haven't done so already, follow the directions above for creating a table of stacked layers.

▶ In the Layer Output group, select Show each combination of categories as a layer. This is equivalent to nesting.

▶ Click OK to create the table.

▶ Double-click the table in the Viewer window to activate it.

▶ Click the down arrow on the drop-down list to display the whole list of layers.

Figure 4-21
List of nested layers in activated pivot table

There are 10 layers in the table (you have to scroll through the list to see all of them), one for each combination of *Gender* and *Highest degree*. For nested layers, the total number of layers is the *product* of the number of categories for each layer variable (in this example, 5 x 2 = 10).

Printing Layered Tables

By default, only the currently visible layer is printed. To print all layers of a table:

▶ Double-click the table in the Viewer window to activate it.

▶ From the Viewer window menus, choose:
Format > Table Properties...

▶ Click the Printing tab.

▶ Select Print all layers.

You can also save this setting as part of a TableLook, including the default TableLook.

Totals and Subtotals for Categorical Variables

You can include both totals and subtotals in custom tables. Totals and subtotals can be applied to categorical variables at any nesting level in any dimension—row, column, or layer.

Sample Data File

The examples in this chapter use the data file *survey_sample.sav*. For more information, see the topic Sample Files in Appendix A on p. 178.

All examples provided here display variable labels in dialog boxes, sorted in alphabetical order. Variable list display properties are set on the General tab in the Options dialog box (Edit menu, Options).

Simple Total for a Single Variable

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ In the table builder, drag and drop *Age category* from the variable list to the Rows area on the canvas pane.

▶ Right-click *Age category* on the canvas pane and choose Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select Column N % in the Statistics list and click the arrow to add it to the Display list.

▶ In the Label cell in the Display list, delete the default label and type Percent.

▶ Click Apply to Selection.

▶ Right-click *Age category* on the canvas pane and choose Categories and Totals from the pop-up context menu.

▶ Select (click) Total in the Categories and Totals dialog box.

Figure 5-1
Categories and Totals dialog box
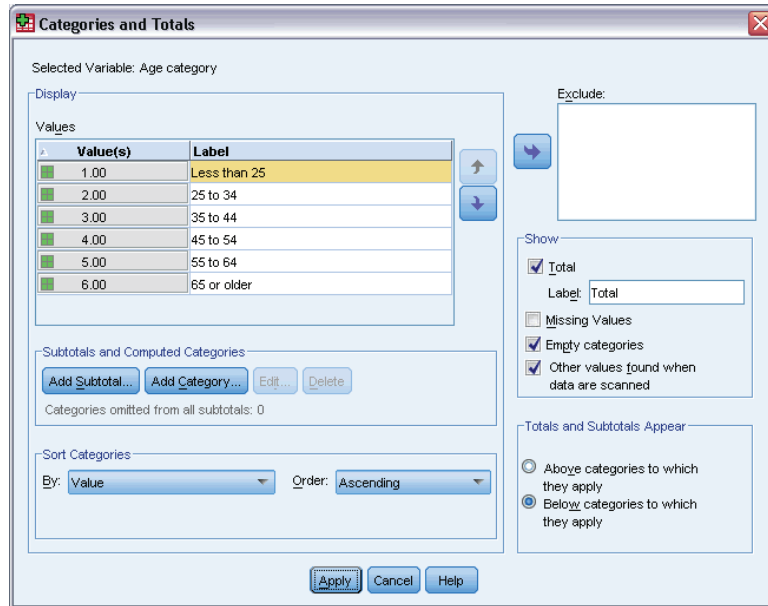
► Click Applyand then click OK in the table builder to create the table.

Figure 5-2
Simple total for a single categorical variable

		Count	Percent
Age category	Less than 25	242	8.6%
	25 to 34	627	22.2%
	35 to 44	679	24.0%
	45 to 54	481	17.0%
	55 to 64	320	11.3%
	65 or older	479	16.9%
	Total	2828	100.0%

What You See Is What Gets Totaled

Totals are based on categories displayed in the table. If you choose to exclude some categories from a table, cases from those categories are not included in total calculations.

► Open the table builder (Analyze menu, Tables, Custom Tables).

► Right-click *Age category* on the canvas pane and choose Categories and Totals from the pop-up context menu.

► Click the category labeled *Less than 25* in the Label list.

► Click the arrow key to the left of the Exclude list.

► Click the category labeled *65 or older* in the Label list.

▶ Click the arrow key to the left of the Exclude list again.

The two categories are moved from the Display list to the Exclude list.

Figure 5-3
Manually excluded categories
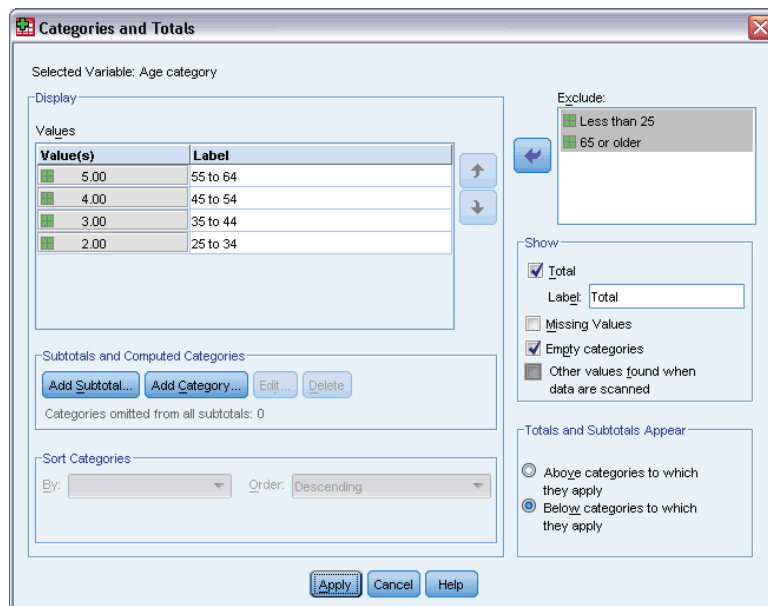
▶ Click Apply and then click OK in the table builder to create the table.

Figure 5-4
Total in table with excluded categories

		Count	Percent
Age category	25 to 34	627	29.8%
	35 to 44	679	32.2%
	45 to 54	481	22.8%
	55 to 64	320	15.2%
	Total	2107	100.0%

The total count in this table is only 2,107, compared to 2,828 when all of the categories are included. Only the categories that are used in the table are included in the total. (The percentage total is still 100% because all of the percentages are based on the total number of cases used in the table, not the total number of cases in the data file.)

Display Position of Totals

By default, totals are displayed below the categories being totaled. You can change the display position of totals to show them above the categories being totaled.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click *Age category* on the canvas pane and choose Categories and Totals from the pop-up context menu.

▶ In the Totals and Subgroups Appear group, select Above categories to which they apply.

► Click Apply and then click OK in the table builder to create the table.

Figure 5-5
Total displayed above totaled categories

		Count	Percent
Age category	Total	2107	100.0%
	25 to 34	627	29.8%
	35 to 44	679	32.2%
	45 to 54	481	22.8%
	55 to 64	320	15.2%

Totals for Nested Tables

Since totals can be applied to categorical variables at any level of the nesting, you can create tables that contain group totals at multiple nesting levels.

Group Totals

Totals for categorical variables nested within other categorical variables represent group totals.

► Open the table builder (Analyze menu, Tables, Custom Tables).

► Drag and drop *Gender* to the left of *Age category* on the canvas pane.

► Right-click *Age category* on the canvas pane and choose Categories and Totals from the pop-up context menu.

Before creating the table, let's move the totals back below the totaled categories.

► In the Totals and Subgroups Appear group, select Below categories to which they apply.

► Click Apply to save the setting and return to the table builder.

Figure 5-6
Age category nested within Gender in the table builder
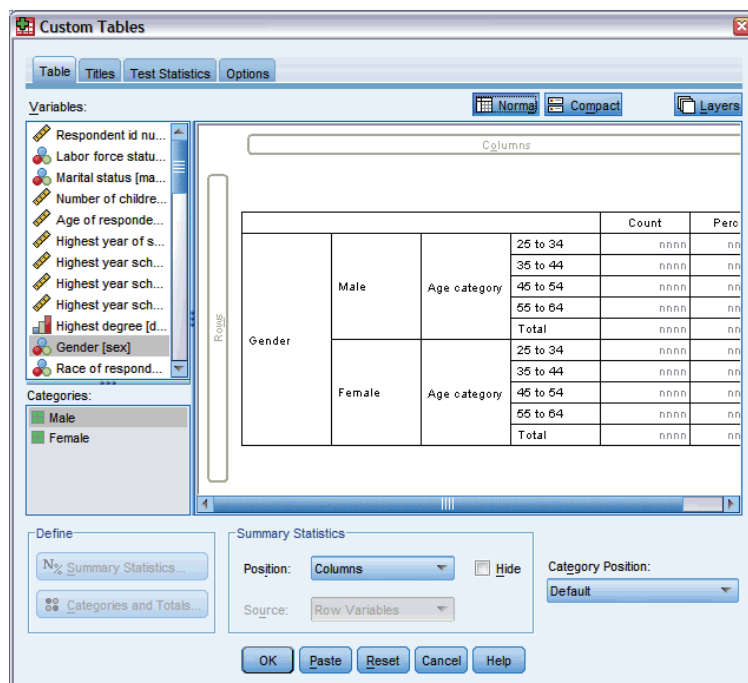
▶ Click OK to create the table.

Figure 5-7
Age category totals within Gender categories

Gender				Count	Percent
Gender	Male	Age category	25 to 34	276	29.3%
			35 to 44	309	32.8%
			45 to 54	221	23.5%
			55 to 64	136	14.4%
			Total	942	100.0%
	Female	Age category	25 to 34	351	30.1%
			35 to 44	370	31.8%
			45 to 54	260	22.3%
			55 to 64	184	15.8%
			Total	1165	100.0%

The table now displays two group totals: one for males and one for females.

Grand Totals

Totals applied to nested variables are always group totals, not grand totals. If you want totals for the entire table, you can apply totals to the variable at the outermost nesting level.

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Right-click *Gender* on the canvas pane and choose Categories and Totals from the pop-up context menu.

▶ Select (click) Total in the Categories and Totals dialog box.

▶ Click Apply and then click OK in the table builder to create the table.

Figure 5-8
Grand totals for a nested table

				Count	Percent
Gender	Male	Age category	25 to 34	276	29.3%
			35 to 44	309	32.8%
			45 to 54	221	23.5%
			55 to 64	136	14.4%
			Total	942	100.0%
	Female	Age category	25 to 34	351	30.1%
			35 to 44	370	31.8%
			45 to 54	260	22.3%
			55 to 64	184	15.8%
			Total	1165	100.0%
	Total	Age category	25 to 34	627	29.8%
			35 to 44	679	32.2%
			45 to 54	481	22.8%
			55 to 64	320	15.2%
			Total	2107	100.0%

Notice that the grand total is only 2,107, not 2,828. Two age categories are still excluded from the table, so the cases in those categories are excluded from all totals.

Layer Variable Totals

Totals for layer variables are displayed as separate layers in the table.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Layers in the table builder to display the Layers list.

▶ Drag and drop *Gender* from the row area on the canvas pane to the Layers list.

Figure 5-9
Layer variable in table builder
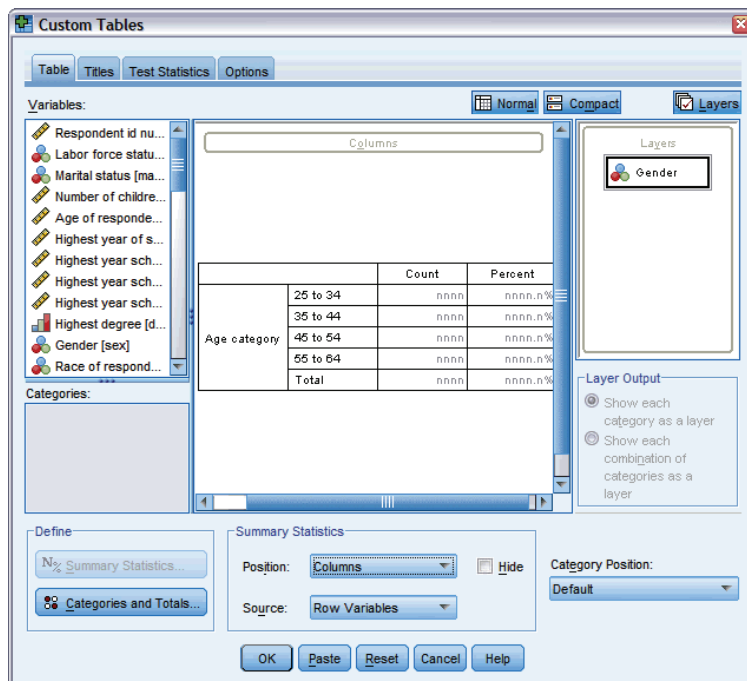
Note: Since you already specified totals for *Gender*, you don't need to do so now. Moving the variable between dimensions does not affect any of the settings for that variable.

▶ Click OK to create the table.

▶ Double-click the table in the Viewer to activate it.

▶ Click the down arrow in the Layer drop-down list to display a list of all the layers in the table.

There are three layers in the table: *Gender Male*, *Gender Female*, and *Gender Total*.

Figure 5-10
Total layer in Layer list in activated pivot table

Display Position of Layer Totals

For layer variable totals, the display position (above or below) for totals determines the layer position for the totals. For example, if you specify Above categories to which they apply for a layer variable total, the total layer is the first layer displayed.
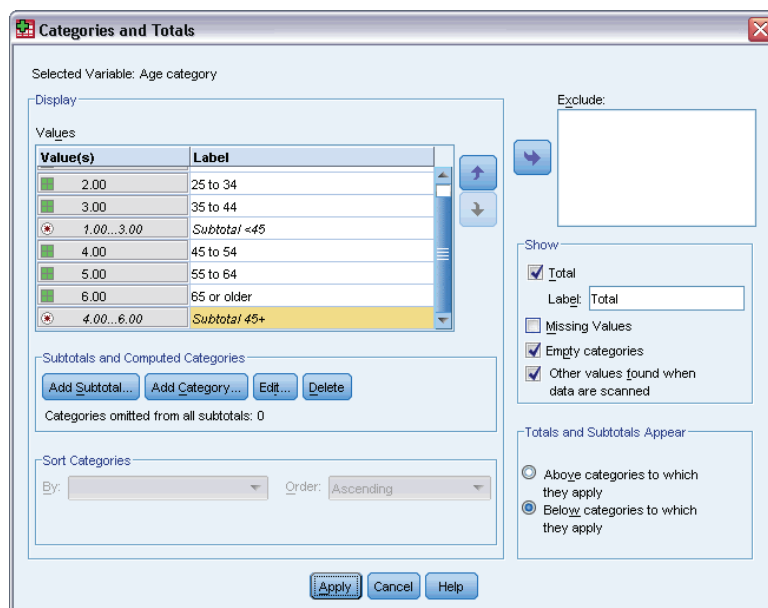
Subtotals

You can include subtotals for subsets of categories of a variable. For example, you could include subtotals for age categories that represent all of the respondents in the sample survey under and over age 45.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Reset to clear any previous settings in the table builder.

▶ In the table builder, drag and drop *Age category* from the variable list to the Rows area on the canvas pane.

▶ Right-click *Age category* on the canvas pane and choose Categories and Totals from the pop-up context menu.

▶ Select 3.00 in the Value(s) list.

▶ Click Add Subtotal to display the Define Subtotal dialog box.

▶ In the Label text field, type Subtotal < 45.

▶ Then click Continue.

This inserts a row containing the subtotal for the first three age categories.

▶ Select 6.00 in the Value(s) list.

▶ Click Add Subtotal to display the Define Subtotal dialog box.

▶ In the Label text field, type Subtotal 45+.

▶ Then click Continue.

Figure 5-11
Defining subtotals

Important note: You should select the display position for totals and subtotals (Above categories to which they apply or Below categories to which they apply) before defining any subtotals. Changing the display position affects all subtotals (not just the currently selected subtotal), and it also *changes the categories included in the subtotals*.

▶ Click Apply and then click OK in the table builder to create the table.

Figure 5-12
Subtotals for Age category

		Count
Age category	Less than 25	242
	25 to 34	627
	35 to 44	679
	Subtotal < 45	1548
	45 to 54	481
	55 to 64	320
	65 or older	479
	Subtotal 45+	1280

What You See Is What Gets Subtotaled

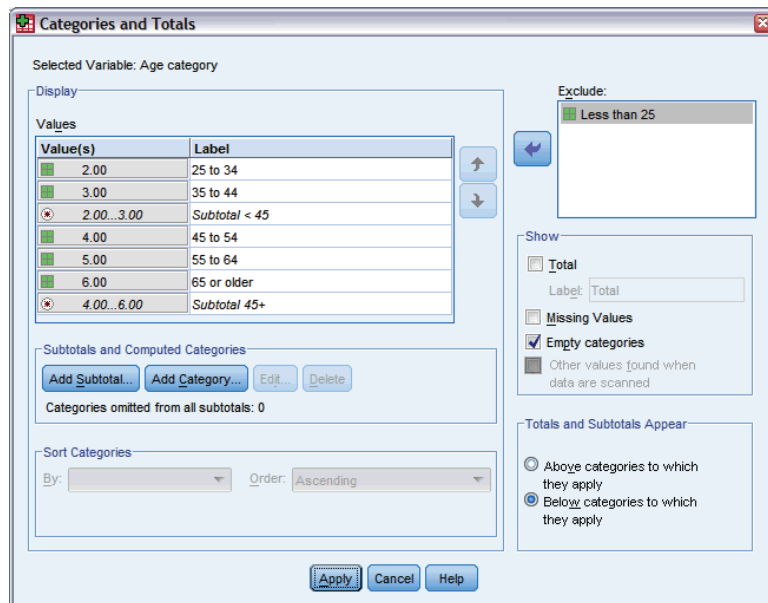Just like totals, subtotals are based on the categories included in the table.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click *Age category* on the canvas pane and choose Categories and Totals from the pop-up context menu.

Note: The value (not the label) displayed for the first subtotal is 1.00...3.00, indicating that the subtotal includes all of the values in the list between 1 and 3.

▶ Select 1.00 in the Value(s) list (or click the label *Less than 25*).

▶ Click the arrow key to the left of the Exclude list.

Figure 5-13
Subtotals when categories are excluded

The first age category is now excluded, and the value displayed for the first subtotal changes to 2.00...3.00, indicating the fact that the excluded category will not be included in the subtotal because subtotals are based on the categories included in the table. Excluding a category automatically excludes it from any subtotals, so you cannot, for example, display only subtotals without the categories on which the subtotals are based.

Hiding Subtotaled Categories

You can suppress the display of the categories that define a subtotal and display only the subtotal, effectively "collapsing" categories without affecting the underlying data.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Reset to clear any previous settings in the table builder.

▶ In the table builder, drag and drop *Age category* from the variable list to the Rows area on the canvas pane.

▶ Right-click *Age category* on the canvas pane and choose Categories and Totals from the pop-up context menu.

▶ Select 3.00 in the Value(s) list.

▶ Click Add Subtotal to display the Define Subtotal dialog box.

▶ In the Label text field, type Less than 45.

▶ Select (check) Hide subtotaled categories from the table.

▶ Then click Continue.

This inserts a row containing the subtotal for the first three age categories.

▶ Select 6.00 in the Value(s) list.

▶ Click Add Subtotal to display the Define Subtotal dialog box.

▶ In the Label text field, type 45 or older.

▶ Select (check) Hide subtotaled categories.

▶ Then click Continue.

▶ To include a total with the subtotals, select (check) Total in the Show group.

Figure 5-14
Hiding subtotaled categories
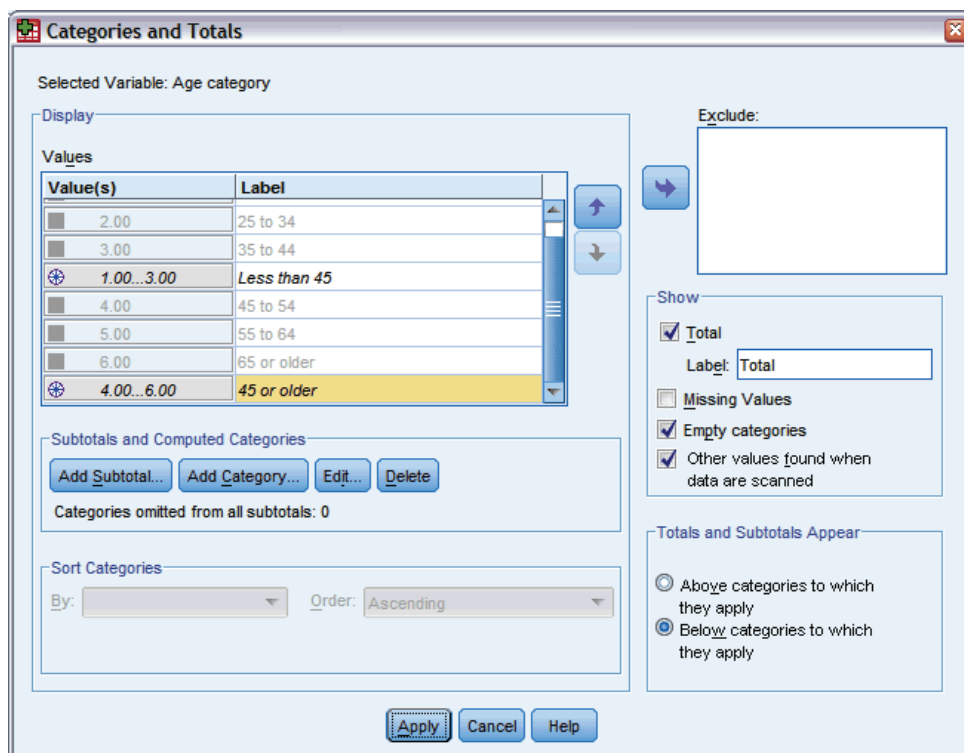
▶ Click Apply.

The canvas reflects the fact that subtotals will be displayed but the categories that define the subtotals will be excluded.

Figure 5-15
Canvas displaying subtotals without subtotaled categories
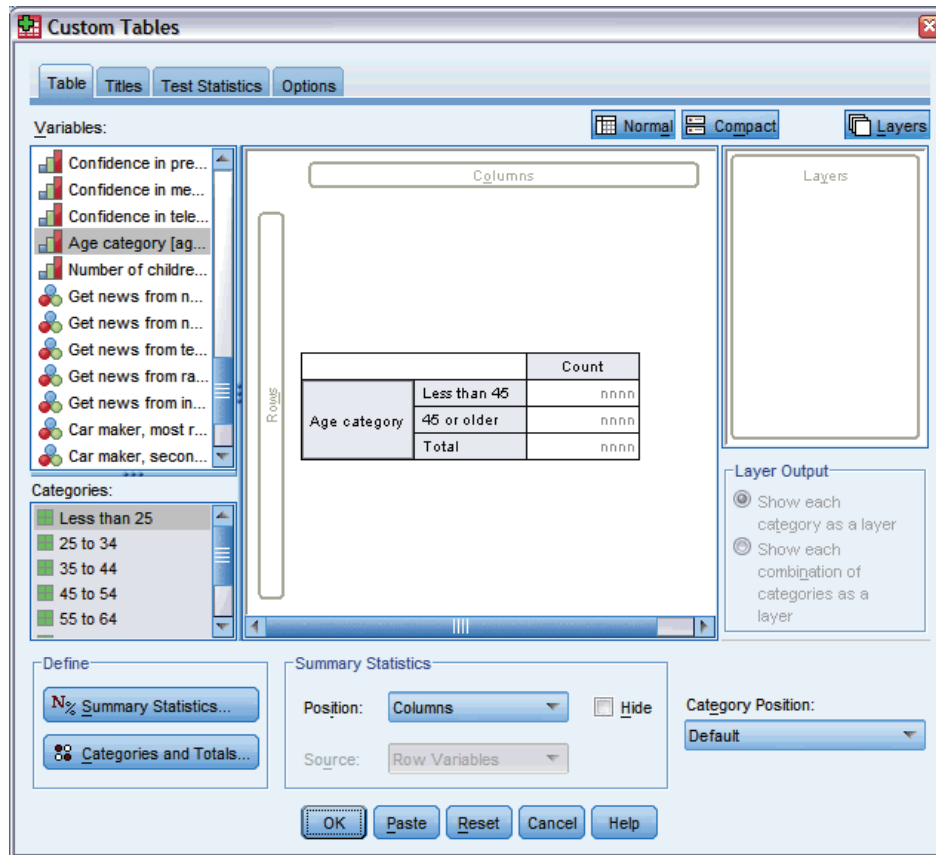
▶ Click OK to produce the table.

Figure 5-16
Table displaying only subtotals and totals

		Count
Age category	Less than 45	1548
	45 or older	1280
	Total	2828

Layer Variable Subtotals

Just like totals, subtotals for layer variables are displayed as separate layers in the table. Essentially, the subtotals are treated as categories. Each category is a separate layer in the table, and the display order of the layer categories is determined by the category order specified in the Categories and Totals dialog box, including the display position of the subtotal categories.

Computed Categories for Categorical Variables

You can include computed categories in custom tables. These are new categories that are calculated from categories of the same variable at any nesting level in any dimension—row, column, or layer. For example, you could include a computed category that shows the difference between two categories.
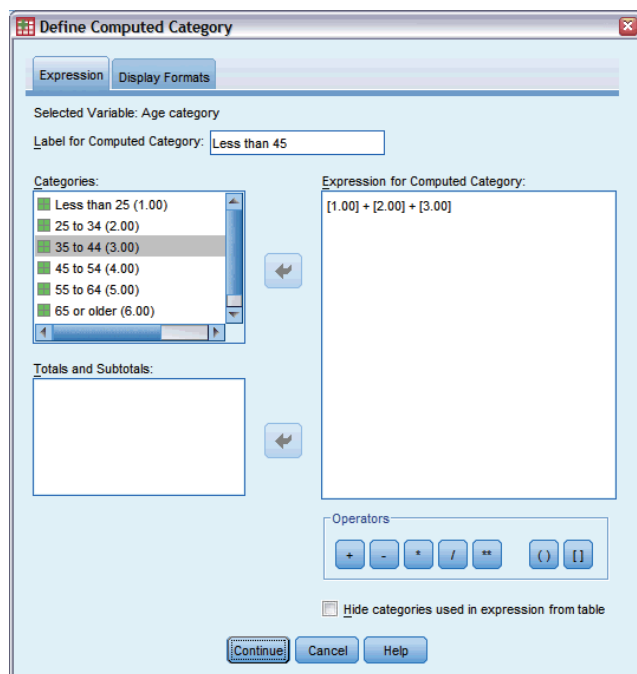
Sample Data File

The examples in this chapter use the data file *survey_sample.sav*. For more information, see the topic Sample Files in Appendix A on p. 178.

Simple Computed Category

▶ From the menus, choose:
Analyze > Tables > Custom Tables...
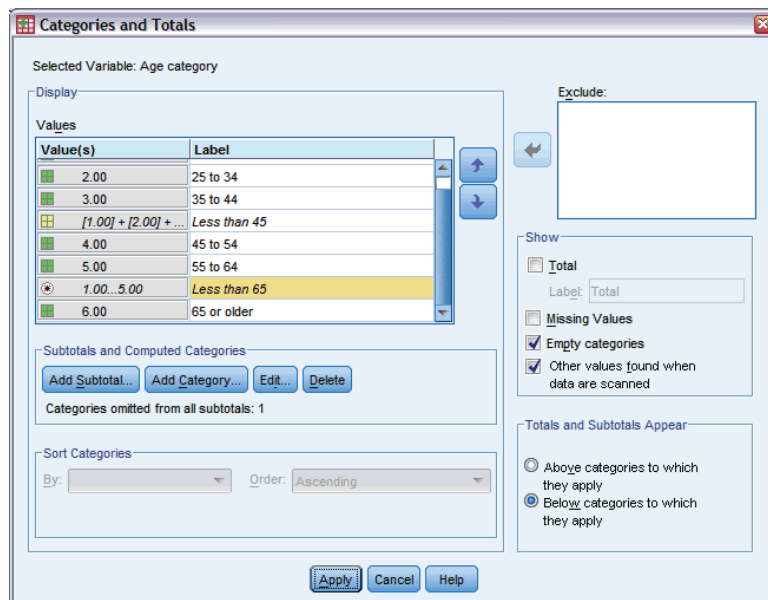
Figure 6-1
Define Computed Category dialog box

▶ In the table builder, drag and drop *Age category* from the variable list to the Rows area on the canvas pane.

▶ Right-click *Age category* on the canvas pane and choose Categories and Totals from the pop-up context menu.

▶ Select 3.00 in the Value(s) list.

▶ Click Add Category to display the Define Compute Category dialog box.

▶ In the Label for Computed Category text field, type Less than 45.

▶ Select Less than 25 (1.00) in the Categories list and click the arrow button to copy it to the Expression for Computed Category text box. [1] is displayed in the expression.

▶ Click the plus (+) operator button in the dialog box (or press the + key on the keyboard).

▶ Select 25 to 34 (2.00) in the Categories list and click the arrow button to copy it to the Expression for Computed Category text box.

▶ Click the plus (+) operator button in the dialog box (or press the + key on the keyboard).

▶ Select 35 to 44 (3.00) in the Categories list and click the arrow button to copy it to the Expression for Computed Category text box.

▶ Then click Continue.

This inserts a row containing the subtotal for the first three age categories.

▶ Select 5.00 in the Value(s) list.

▶ Click Add Subtotal to display the Define Subtotal dialog box.

▶ In the Label text field, type Less than 65.

▶ Then click Continue.

This inserts a row containing the subtotal for the first the first five categories.

Figure 6-2
Categories and Totals dialog box

▶ Click Apply and then click OK in the table builder to create the table.

Figure 6-3
Computed category with subtotal

		Count
Age category	Less than 25	242
	25 to 34	627
	35 to 44	679
	Less than 45	1548
	45 to 54	481
	55 to 64	320
	Less than 65	2349
	65 or older	479

The table includes a computed category (*Less than 45*) and a subtotal (*Less than 65*). The subtotal includes categories also included in the computed category. You could not create the same table with subtotals alone, because subtotals cannot share the same categories.
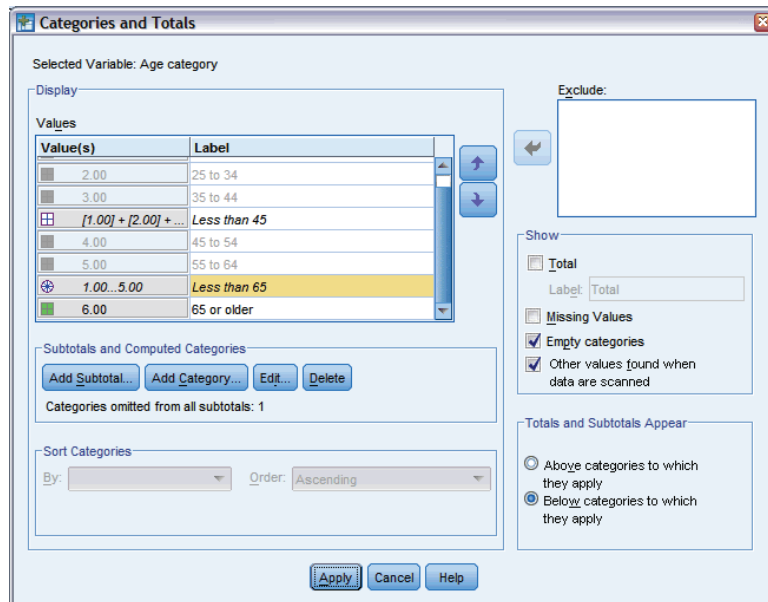
Hiding Categories in a Computed Category

As with subtotals, you can suppress the display of the categories that are used in a computed category's expression and display only the computed category itself. The following example builds on the previous one.

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ Right-click *Age category* on the canvas pane and choose Categories and Totals from the pop-up context menu.

▶ Select the *Less than 45* computed category in the Value(s) list.

▶ Click Edit to display the Define Compute Category dialog box.

▶ Select Hide categories used in expression from table.

▶ Then click Continue.

▶ Select the *Less than 65* subtotal in the Value(s) list.

▶ Click Edit to display the Define Subtotal dialog box.

▶ Select Hide subtotaled categories from the table.

▶ Then click Continue.

Figure 6-4
Categories and Totals dialog box showing hidden categories

▶ Click Apply and then click OK in the table builder to create the table.

Figure 6-5
Computed category with subtotal and hidden categories

		Count
Age category	Less than 45	1548
	Less than 65	2349
	65 or older	479

Like the previous example, the table includes a computed category and a subtotal. But in this case the categories in each are hidden so that only these totals are shown.

Referencing Subtotals in a Computed Category

You can include subtotals in a computed category's expression.

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ Click Reset to clear any previous settings in the table builder.

▶ In the table builder, drag and drop *Labor force status* from the variable list into the Rows area of the canvas pane.

▶ Drag and drop *Marital status* from the variable list into the Columns area.

▶ Right-click *Labor force status* on the canvas pane and choose Categories and Totals from the pop-up context menu.

▶ Select 2 in the Value(s) list.

▶ Click Add Subtotal to display the Define Subtotal dialog box.

▶ In the Label text field, type Working.

▶ Select Hide subtotaled categories from the table.

▶ Then click Continue.

This inserts a row containing the subtotal for the first two working status categories.

▶ Select 8 in the Value(s) list.

▶ Click Add Subtotal to display the Define Subtotal dialog box.

▶ In the Label text field, type Not Working.

▶ Select Hide subtotaled categories.

▶ Then click Continue.

This inserts a row containing the subtotal for the other working status categories.

Figure 6-6
Categories and Totals dialog box
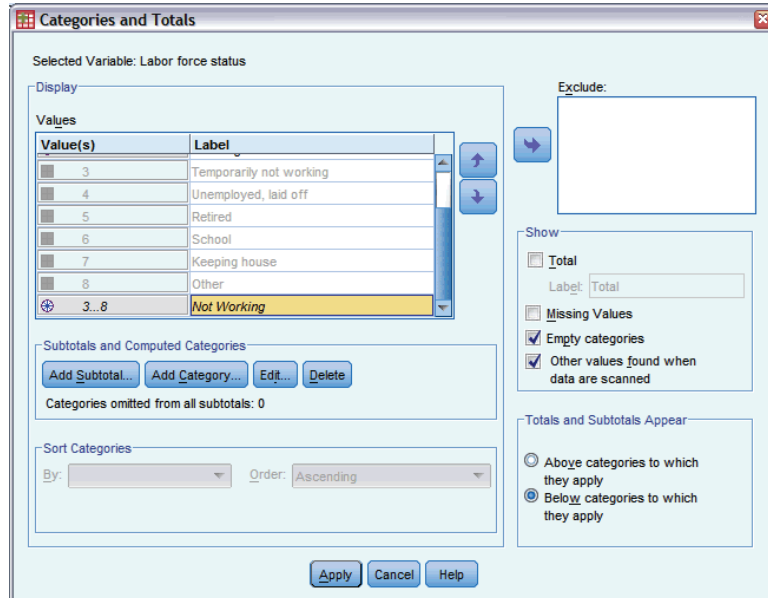
▶ Select the *Not Working* subtotal in the Value(s) list.

▶ Click Add Category to display the Define Compute Category dialog box.

▶ In the Label for Computed Category text field, type Working / Not Working.

▶ Select Working (Working #1) in the Totals and Subtotals list and click the arrow button to copy it to the Expression for Computed Category text box.

▶ Click the division (/) operator button in the dialog box (or press the / key on the keyboard).

▶ Select Not Working (Not Working #2) in the Totals and Subtotals list and click the arrow button to copy it to the Expression for Computed Category text box.

Figure 6-7
Define Computed Category dialog box
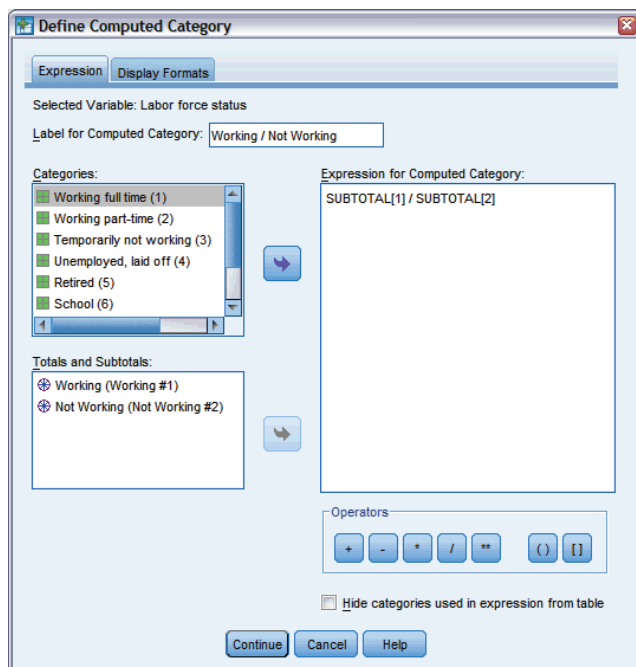
By default, the computed category uses the same format as the variable's statistic, which is Count in this case. Because we want to show decimal places resulting from the division in the computed category's expression and the default format for Count does not include decimal places, we need to change the format.

▶ Click the Display Formats tab.

▶ Change the Decimals setting for Count to 2.

Figure 6-8
Display Formats tab in Define Computed Category dialog box
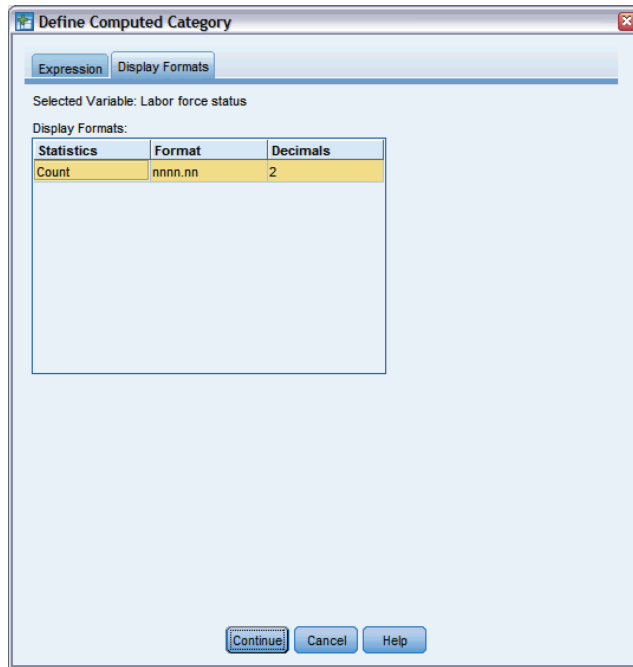
▶ Then click Continue.

▶ Click Apply and then click OK in the table builder to create the table.

Figure 6-9
Computed category showing ratio of subtotals

		Marital status				
		Married	Widowed	Divorced	Separated	Never married
		Count	Count	Count	Count	Count
Labor force status	Working	916	64	330	67	494
	Not Working	429	219	116	26	169
	Working / Not Working	2.14	.29	2.84	2.58	2.92

The table includes two subtotals and a computed category. The computed category shows the ratio of the subtotals so that you can easily compare the groups represented by each subtotal. There's a much lower ratio of working to not working widowed respondents compared to the other groups. Also, there is a slightly lower ratio of married respondents, perhaps resulting from spouses who leave the workforce to stay home with a child.

## *Using Computed Categories to Display Nonexhaustive Subtotals*

Subtotals are exhaustive. That is, all subtotals in a table include all values above or below their positions in the table. Computed categories, on the other hand, are not exhaustive and allow you to sum a mix of categories in a table.

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ Click Reset to clear any previous settings in the table builder.

▶ In the table builder, drag and drop *Think of self as liberal or conservative* from the variable list into the Rows area of the canvas pane.

▶ Right-click *Think of self as liberal or conservative* on the canvas pane and choose Categories and Totals from the pop-up context menu.

▶ Select 3 in the Value(s) list.

▶ Click Add Category to display the Define Computed Category dialog box.

▶ In the Label for Computed Category text field, type   Liberal Subtotal. Note that there are four spaces before the text. These spaces are used for indentation in the resulting table.

▶ Select Extremely liberal (1) in the Categories list and click the arrow button to copy it to the Expression for Computed Category text box.

▶ Click the plus (+) operator button in the dialog box (or press the + key on the keyboard).

▶ Select Liberal (2) in the Categories list and click the arrow button to copy it to the Expression for Computed Category text box.

▶ Click the plus (+) operator button in the dialog box (or press the + key on the keyboard).

▶ Select Slightly liberal (3) in the Categories list and click the arrow button to copy it to the Expression for Computed Category text box.

▶ Click Continue.

This inserts a row containing the subtotal for the liberal categories.

Figure 6-10
Categories and Totals dialog box
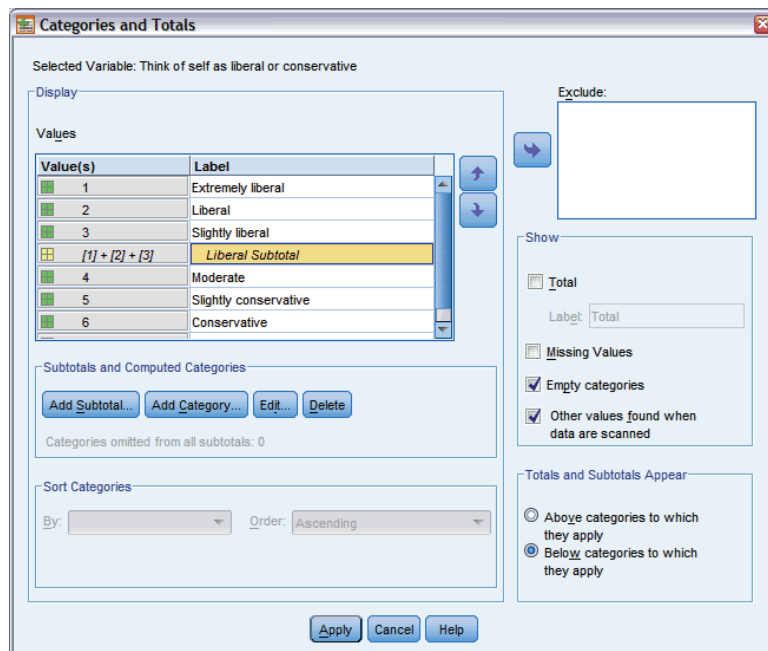
▶ Select 7 in the Value(s) list.

▶ Click Add Category to display the Define Computed Category dialog box.

▶ In the Label for Computed Category text field, type   Conservative Subtotal. Note that there are four spaces before the text. These spaces are used for indentation in the resulting table.

▶ Select Slight conservative (5) in the Categories list and click the arrow button to copy it to the Expression for Computed Category text box.

▶ Click the plus (+) operator button in the dialog box (or press the + key on the keyboard).

▶ Select Conservative (6) in the Categories list and click the arrow button to copy it to the Expression for Computed Category text box.

▶ Click the plus (+) operator button in the dialog box (or press the + key on the keyboard).

▶ Select Extremely conservative (7) in the Categories list and click the arrow button to copy it to the Expression for Computed Category text box.

▶ Click Continue.

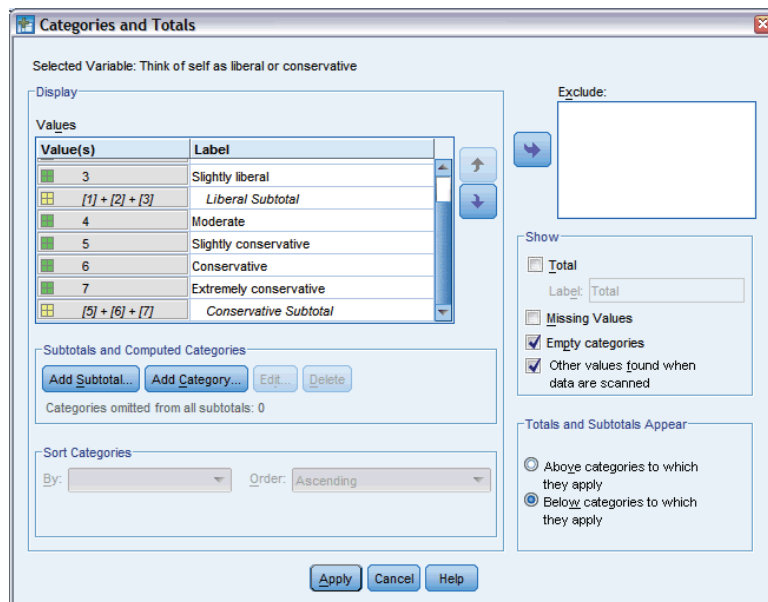This inserts a row containing the subtotal for the conservative categories.

Figure 6-11
Categories and Totals dialog box

▶ Click Apply and then click OK in the table builder to create the table.

Figure 6-12
Computed categories displaying nonexhaustive subtotals

		Count
Think of self as liberal or conservative	Extremely liberal	64
	Liberal	357
	Slightly liberal	351
	Liberal Subtotal	772
	Moderate	986
	Slightly conservative	432
	Conservative	415
	Extremely conservative	86
	Conservative Subtotal	933

The table includes two computed categories that do not include all the categories displayed in the table. The *Moderate* category is not included in either computed category. You cannot create the same table with subtotals because subtotals are exhaustive.

Tables for Variables with Shared Categories

Surveys often contain many questions with a common set of possible responses. For example, our sample survey contains a number of variables concerning confidence in various public and private institutions and services, all with the same set of response categories: 1 = *A great deal*, 2 = *Only some*, and 3 = *Hardly any*. You can use stacking to display these related variables in the same table—and you can display the shared response categories in the columns of the table. These features are also available if you use computed categories, with the provision that any computed category's label and expression are the same in all variables.

Figure 7-1
Table of variables with shared categories

	A great deal	Only some	Hardly any
Confidence in banks & financial institutions	490	1068	306
Confidence in education	511	1055	315
Confidence in major companies	500	1078	243
Confidence in medicine	844	864	167
Confidence in press	176	878	808
Confidence in television	196	936	744

Note: In the previous version of Custom Tables, this was known as a "table of frequencies."

Sample Data File

The examples in this chapter use the data file *survey_sample.sav*. For more information, see the topic Sample Files in Appendix A on p. 178.

All examples provided here display variable labels in dialog boxes, sorted in alphabetical order. Variable list display properties are set on the General tab in the Options dialog box (Edit menu, Options).

Table of Counts

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ In the variable list in the table builder, click *Confidence in banks...* and then Shift-click *Confidence in television* to select all of the "confidence" variables. (*Note*: This assumes that variable labels are displayed in alphabetical order, not file order, in the variable list.)

▶ Drag and drop the six confidence variables to the Rows area on the canvas pane.

Figure 7-2
Confidence variables stacked in rows
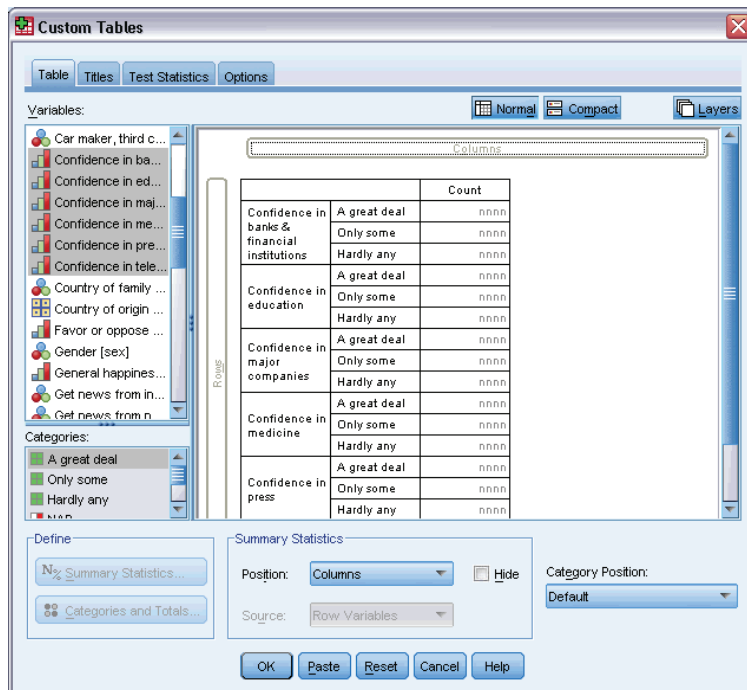
This stacks the variables in the row dimension. By default, the category labels for each variable are also displayed in the rows, resulting in a very long, narrow table (6 variables x 3 categories = 18 rows)—but since all six variables share the same defined category labels (value labels), you can put the category labels in the column dimension.

▶ From the Category Position drop-down list, select Row Labels in Columns.

Now the table has only six rows, one for each of the stacked variables, and the defined categories become columns in the table.

▶ Before creating the table, select (click) Hide for Position in the Summary Statistics group, since the summary statistic label *Count* isn't really necessary.

Figure 7-3
Category labels in columns
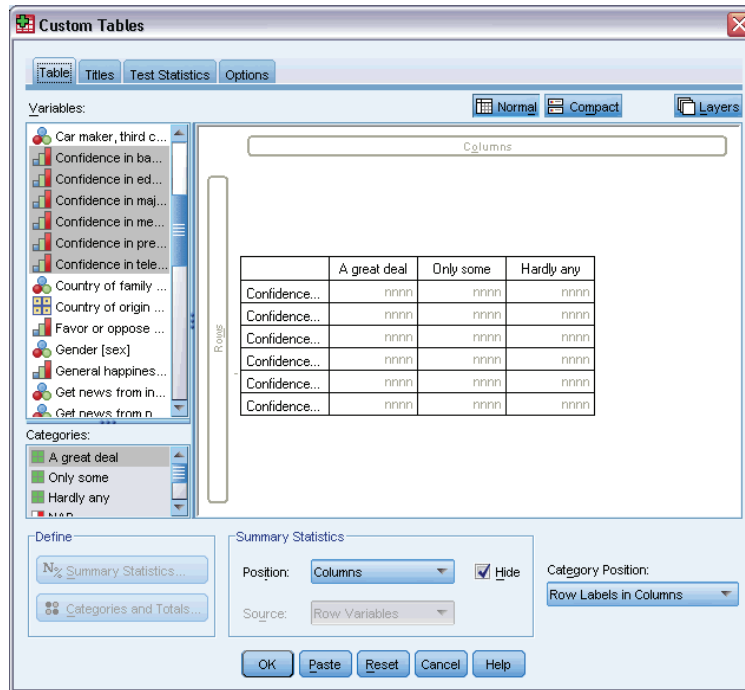
▶ Click OK to create the table.

Figure 7-4
Table of stacked row variables with shared category labels in columns

	A great deal	Only some	Hardly any
Confidence in banks & financial institutions	490	1068	306
Confidence in education	511	1055	315
Confidence in major companies	500	1078	243
Confidence in medicine	844	864	167
Confidence in press	176	878	808
Confidence in television	196	936	744

Instead of displaying the variables in the rows and categories in the columns, you could create a table with the variables stacked in the columns and the categories displayed in the rows. This might be a better choice if there were more categories than variables, whereas in our example there are more variables than categories.

Table of Percentages

For a table with variables stacked in rows and categories displayed in columns, the most meaningful (or at least easiest to understand) percentage to display is row percentages. (For a table with variables stacked in the columns and categories displayed in the rows, you would probably want column percentages.)

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Right-click any one of the confidence variables in the table preview on the canvas pane and choose Summary Statistics from the pop-up context menu.

▶ Select Row N % in the Statistics list and click the arrow button to move it to the Display list.

▶ Click any cell in the *Count* row in the Display list and click the arrow button to move it back to the Statistics list, removing it from the Display list.

Figure 7-5
Row percentages selected
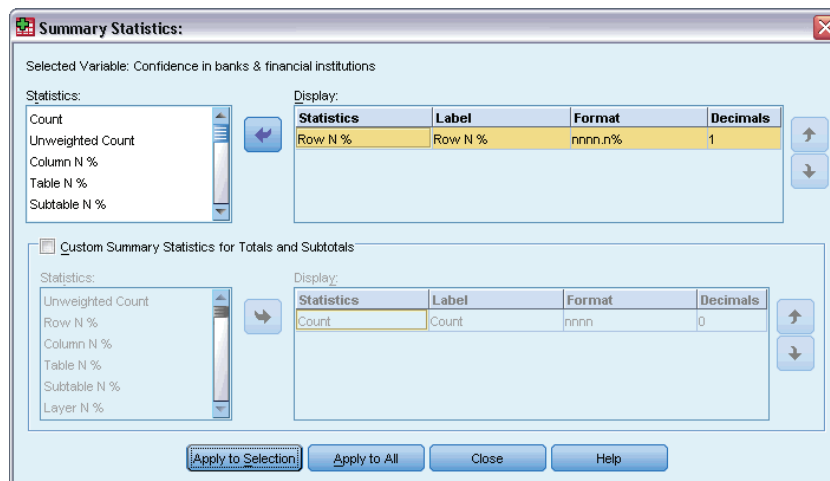
▶ Click Apply to All to apply the summary statistic change to all of the stacked variables in the table.

Figure 7-6
Row percentages in table preview on canvas pane
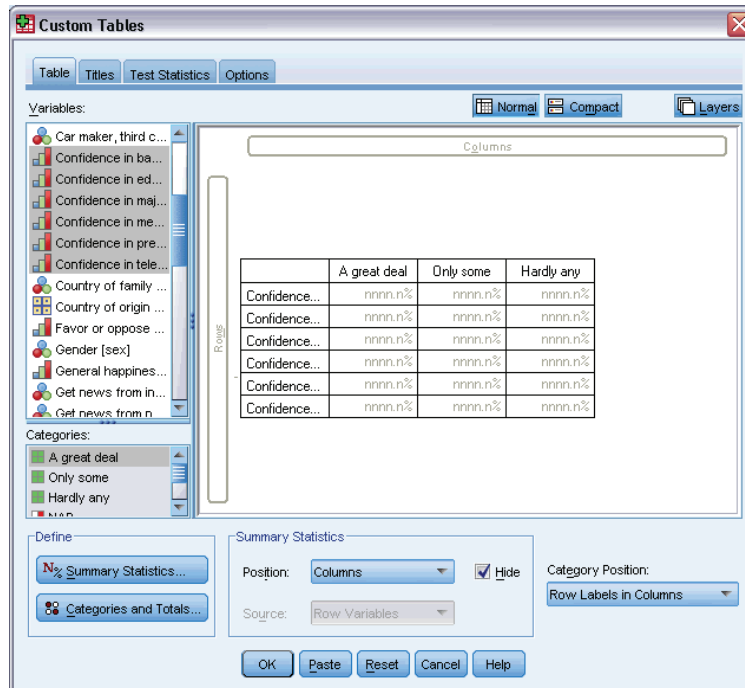
Note: If your table preview doesn't look like this figure, you probably clicked Apply to Selection instead of Apply to All, which applies the new summary statistic only to the selected variable. In this example, that would result in two columns for each category: one with count placeholders displayed for all of the other variables and one with a row percentage placeholder displayed for the selected variable. This is exactly the table that would be produced but *not* the one that we want in this example.

▶ Click OK to create the table.

Figure 7-7
Table of row percentages for variables stacked in rows, categories displayed in columns

|	A great deal	Only some	Hardly any
Confidence in banks & financial institutions	26.3%	57.3%	16.4%
Confidence in education	27.2%	56.1%	16.7%
Confidence in major companies	27.5%	59.2%	13.3%
Confidence in medicine	45.0%	46.1%	8.9%
Confidence in press	9.5%	47.2%	43.4%
Confidence in television	10.4%	49.9%	39.7%

Note: You can include any number of summary statistics in a table of variables with shared categories. Our examples show only one at a time to keep them simple.

Totals and Category Control

You can create tables with categories in the opposite dimension from the variables only if all of the variables in the table have the same categories, displayed in the same order. This includes totals, subtotals, and any other category adjustments you make. This means that any modifications you make in the Categories and Totals dialog box must be made for all variables in the table that share the categories.

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Right-click the first confidence variable in the table preview on the canvas pane and choose Categories and Totals from the pop-up context menu.

▶ Select (check) Total in the Categories and Totals dialog box and then click Apply.

Figure 7-8
Probably not the results you want
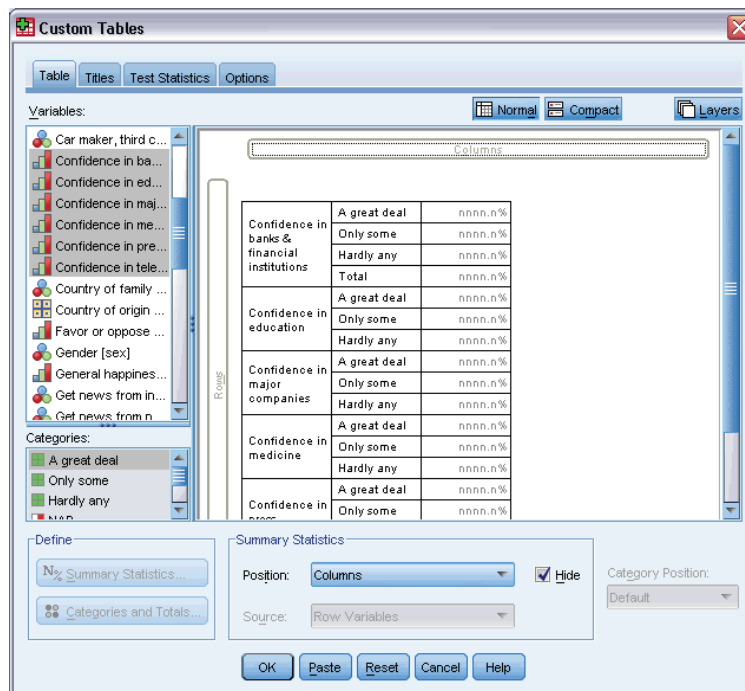
The first thing you'll probably notice is that the category labels have moved from the columns back to the rows. You may also notice that the Category Position control is now disabled. This is because the variables no longer share the exact same set of "categories." One of the variables now has a total category.

▶ Right-click any one of the confidence variables on the canvas pane and select Select All Row Variables from the pop-up context menu—or Ctrl-click each stacked variable on the canvas pane until they are all selected (you may have to scroll down the pane or expand the table builder window).

▶ Click Categories and Totals in the Define group.

▶ If Total isn't already selected (checked) in the Categories and Totals dialog box, select it now and then click Apply.

▶ The Category Position drop-down list should be enabled again, since now all of the variables have the additional total category, so select Row Labels in Columns.

Figure 7-9
Categories and totals in columns
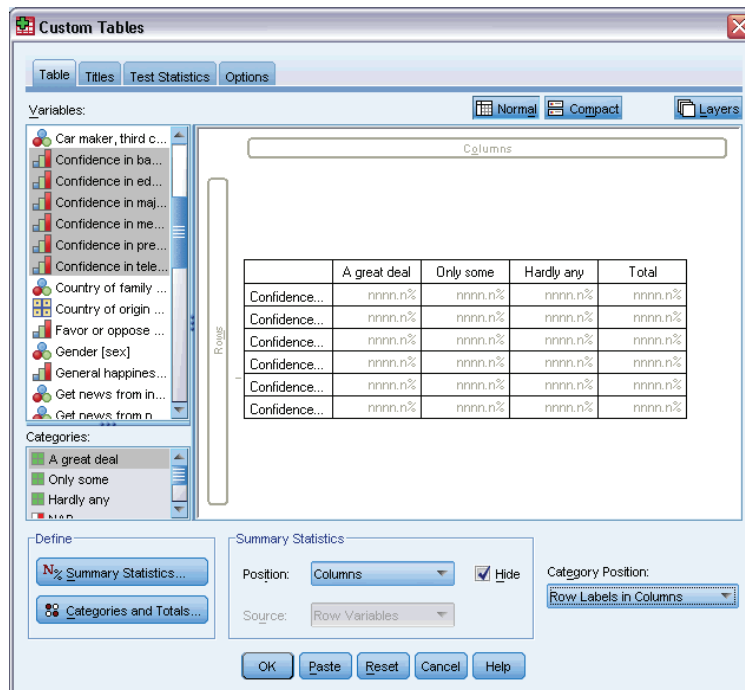
▶ Click OK to create the table.

Figure 7-10
Table of row percentages for variables stacked in rows, categories and totals displayed in columns

	A great deal	Only some	Hardly any	Total
Confidence in banks & financial institutions	26.3%	57.3%	16.4%	100.0%
Confidence in education	27.2%	56.1%	16.7%	100.0%
Confidence in major companies	27.5%	59.2%	13.3%	100.0%
Confidence in medicine	45.0%	46.1%	8.9%	100.0%
Confidence in press	9.5%	47.2%	43.4%	100.0%
Confidence in television	10.4%	49.9%	39.7%	100.0%

Nesting in Tables with Shared Categories

In nested tables, the stacked variables with the shared categories must be at the innermost nesting level of their dimension if you want to display the category labels in the opposite dimension.

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Drag and drop *Gender* from the variable list to the left side of the Rows area.

Figure 7-11
Nested variables with shared categories
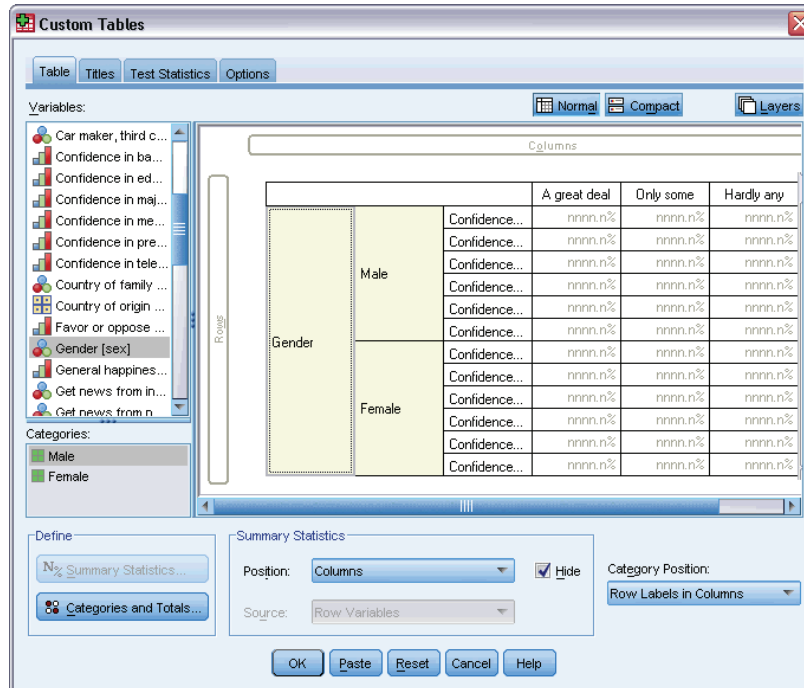
The stacked variables with shared categories are now nested within gender categories in the table preview.

▶ Now drag and drop *Gender* to the right of one of the stacked confidence variables in the table preview.

Figure 7-12
Another example of results you probably do not want
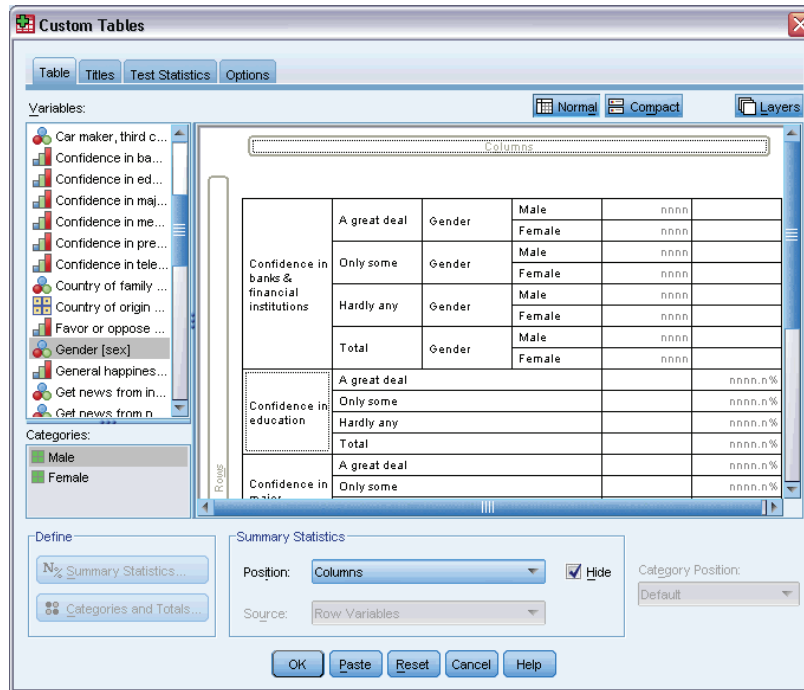
Once again, the category labels have reverted to the row dimension, and the Category Position control is disabled. You now have one stacked variable that also has *Gender* nested within it, while the other stacked variables contain no nested variables. You could add *Gender* as a nested variable to each of the stacked variables, but then moving row labels to columns would result in the category labels for *Gender* being displayed in the columns, not the category labels for the stacked variables with the shared categories. This is because *Gender* would now be the innermost nested variable, and changing the category position always applies to the innermost nested variable.

Summary Statistics

Summary statistics include everything from simple counts for categorical variables to measures of dispersion, such as the standard error of the mean for scale variables. It does *not* include significance tests available on the Test Statistics tab in the Custom Tables dialog box. For more information, see the topic Test Statistics in Chapter 10 on p. 132.

Summary statistics for categorical variables and multiple response sets include counts and a wide variety of percentage calculations, including:

- Row percentages
- Column percentages
- Subtable percentages
- Table percentages
- Valid N percentages

In addition to the summary statistics available for categorical variables, summary statistics for scale variables and custom total summaries for categorical variables include:

- Mean
- Median
- Percentiles
- Sum
- Standard deviation
- Range
- Minimum and maximum values

Additional summary statistics are available for multiple response sets. For more information, see the topic Counts, Responses, Percentages, and Totals in Chapter 11 on p. 150. A complete list of summary statistics is also available. For more information, see the topic Summary Statistics in Chapter 2 on p. 29.

Sample Data File

The examples in this chapter use the data file *survey_sample.sav*. For more information, see the topic Sample Files in Appendix A on p. 178.

All examples provided here display variable labels in dialog boxes, sorted in alphabetical order. Variable list display properties are set on the General tab in the Options dialog box (Edit menu, Options).

Summary Statistics Source Variable

Available summary statistics depend on the measurement level of the summary statistics source variable. The source of summary statistics (the variable on which the summary statistics are based) is determined by:

■ **Measurement level**. If a table (or a table section in a stacked table) contains a scale variable, summary statistics are based on the scale variable.

■ **Variable selection order**. The default statistics source dimension (row or column) for categorical variables is based on the order in which you drag and drop variables onto the canvas pane. For example, if you drag a variable to the rows area first, the row dimension is the default statistics source dimension.

■ **Nesting**. For categorical variables, summary statistics are based on the innermost variable in the statistics source dimension.

A stacked table may have multiple summary statistics source variables (both scale and categorical), but each table section has only one summary statistics source.

Summary Statistics Source for Categorical Variables

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ In the table builder, drag and drop *Age category* from the variable list into the Rows area of the canvas pane.

▶ Right-click on *Age category* on the canvas pane and select Summary Statistics from the pop-up context menu. (Since this is the only variable in the table, it is the statistics source variable.)

▶ In the Summary Statistics dialog box, select *Column N %* in the Statistics list and click the arrow to add it to the Display list.

Figure 8-1
Summary Statistics dialog box for categorical variables
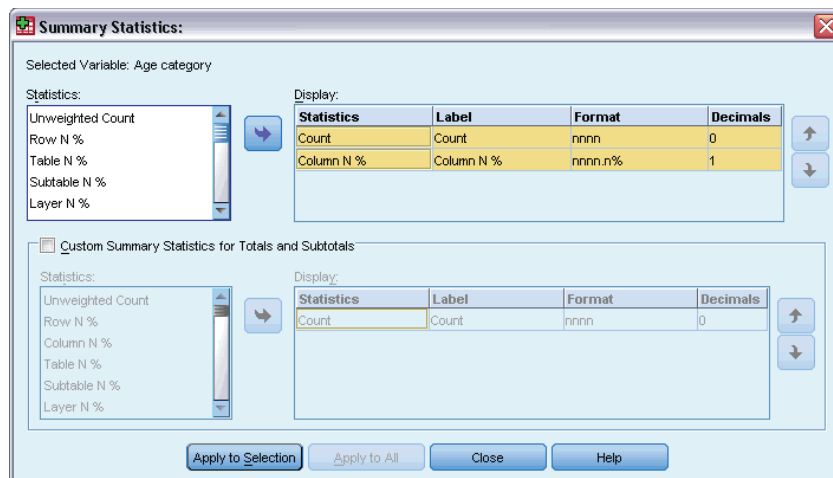
▶ Click Apply to Selection.

▶ In the table builder, drag and drop *Get news from internet* to the right of *Age category* on the canvas pane.

Figure 8-2
Nested categorical variables
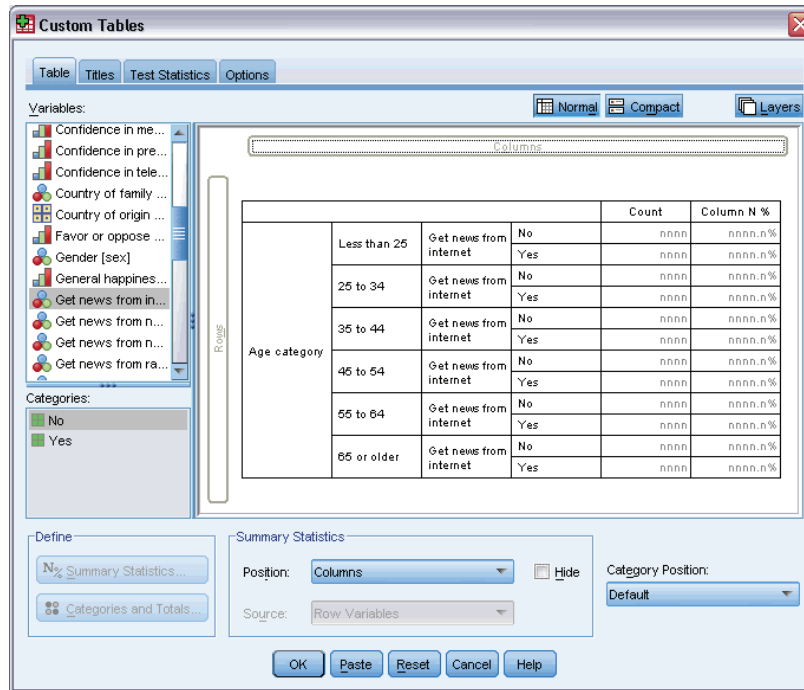
▶ Right-click on *Age category* on the canvas pane again. The Summary Statistics item on the context menu is now disabled because *Age category* is not the innermost nested variable in the statistics source dimension.

▶ Right-click on *Get news from internet* on the canvas pane. The Summary Statistics item is enabled because it is now the summary statistics source variable, since it is the innermost nested variable in the statistics source dimension. (Since the table has only one dimension—rows—it is the statistics source dimension.)

▶ Drag and drop *Get news from internet* from the Rows area on the canvas pane into the Columns area.

Figure 8-3
Crosstabulated categorical variables
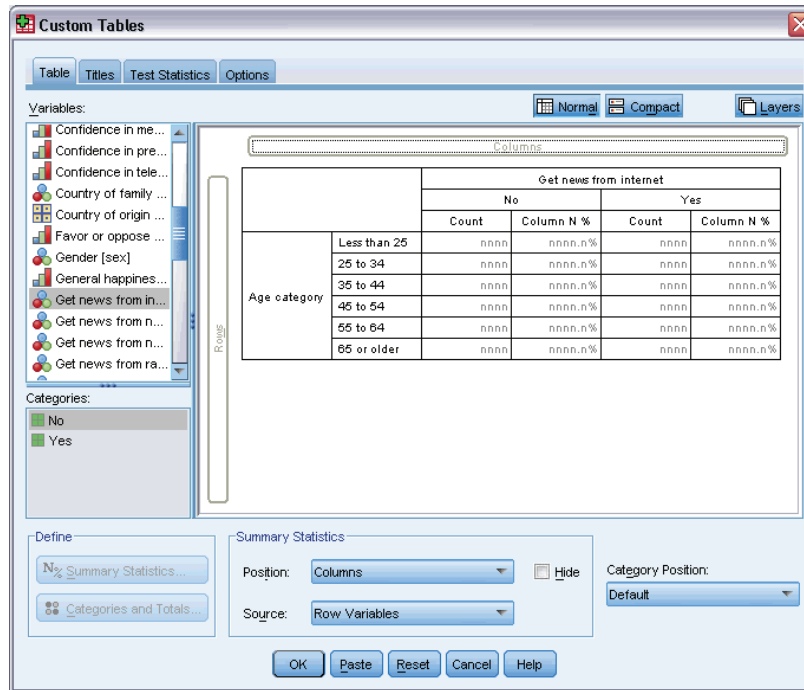
▶ Right-click on *Get news from internet* on the canvas pane again. The Summary Statistics item on the pop-up context menu is now disabled because the variable is no longer in the statistics source dimension.

Age category is once again the statistics source variable because the default statistics source dimension for categorical variables is the first dimension where you put variables when creating the table. In this example, the first thing we did was put variables in the row dimension. Thus, the row dimension is the default statistics source dimension; and since *Age category* is now the only variable in that dimension, it is the statistics source variable.

Summary Statistics Source for Scale Variables

▶ Drag and drop the scale variable *Hours per day watching TV* to the left of *Age category* in the Rows area of the canvas pane.

Figure 8-4
Crosstabulation with scale summary statistics variable
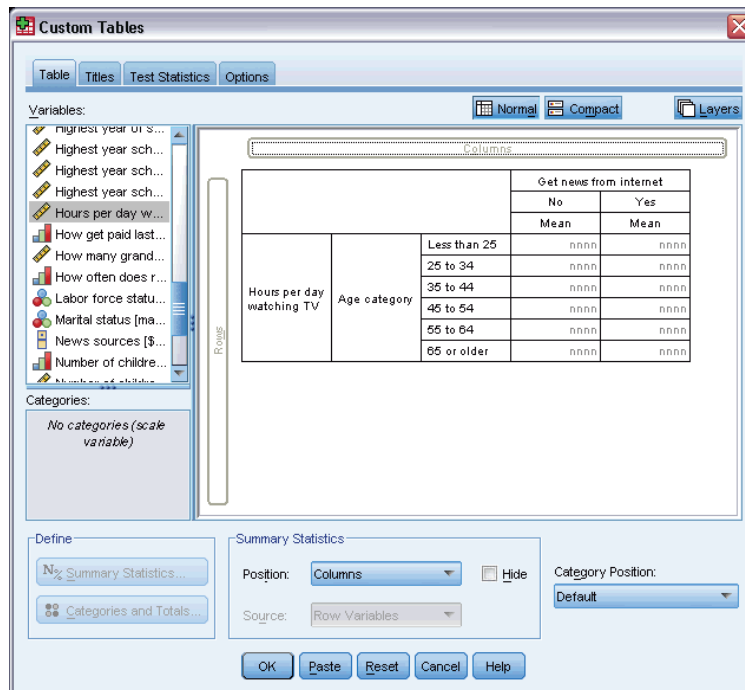
The first thing you may notice is that the *Count* and *Column N %* summaries have been replaced with *Mean*—and if you right-click on *Hours per day watching TV* on the canvas pane, you'll see that it is now the summary statistics source variable. For a table with a scale variable, the scale variable is always the statistics source variable regardless of its nesting level or dimension, and the default summary statistic for scale variables is the mean.

▶ Drag and drop *Hours per day watching TV* from the Rows area into the Columns area above *Get news from internet*.

▶ Right-click on *Hours per day watching TV* and select Summary Statistics from the pop-up context menu. (It's still the statistics source variable even when you move it to a different dimension.)

▶ In the Summary Statistics dialog box, click the Format cell for the mean in the Display list and select nnnn from the Format drop-down list. (You may have to scroll up the list to find this choice.)

▶ In the Decimals cell, type 2.

Figure 8-5
Summary Statistics dialog box for scale variables
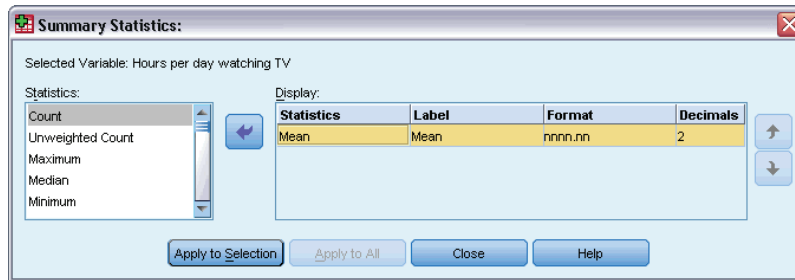
▶ Click Apply to Selection.

Figure 8-6
Scale summary statistic with two decimals
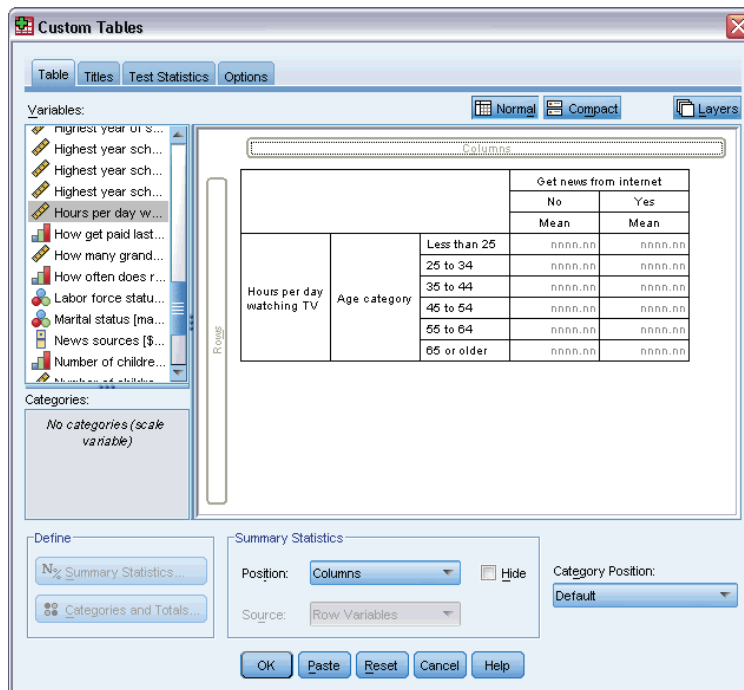
The table preview on the canvas pane now shows that the mean values will be displayed with two decimals.

▶ Click OK to create the table.

Figure 8-7
Scale variable summarized within crosstabulated categorical variables

		Hours per day watching TV	
		Get news from internet	
		No	Yes
		Mean	Mean
Age category	Less than 25	3.54	2.12
	25 to 34	3.42	2.14
	35 to 44	3.00	2.01
	45 to 54	2.83	2.06
	55 to 64	3.24	2.37
	65 or older	3.82	2.33

Stacked Variables

Since a stacked table can contain multiple statistics source variables and you can specify different summary statistics for each of those statistics source variables, there are a few special considerations for specifying summary statistics in stacked tables.

► Open the table builder (Analyze menu, Tables, Custom Tables).

► Click Reset to clear any previous settings in the table builder.

► Click *Get news from internet* in the variable list and then shift-click *Get news from television* in the variable list to select all of the "news" variables. (*Note*: This assumes that variable labels are displayed in alphabetical order, not file order, in the variable list.)

► Drag and drop the five news variables into the Rows area of the canvas pane.

Figure 8-8
News variables stacked in rows
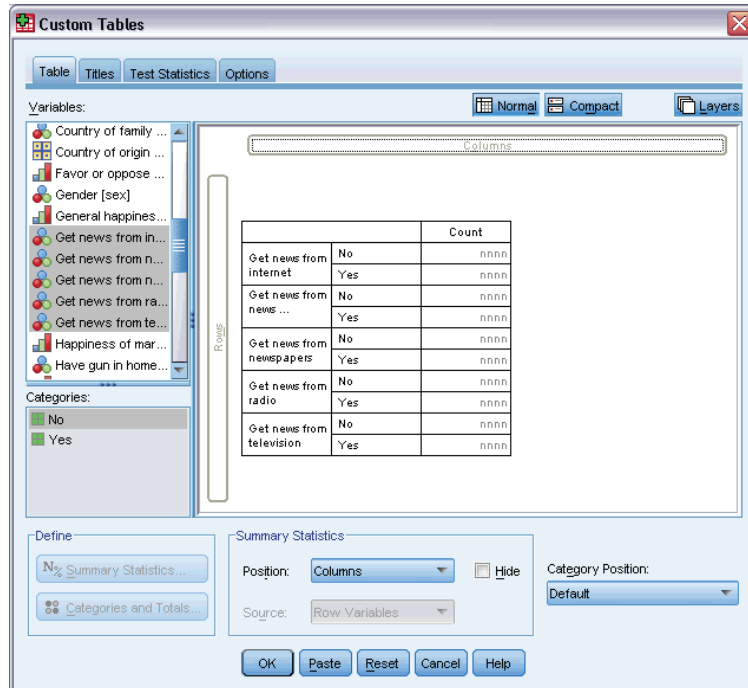
The five news variables are stacked in the row dimension.

▶ Click *Get news from internet* on the canvas pane so that only that variable is selected.

▶ Now right-click *Get news from internet* and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select *Column N %* from the Statistics list and click the arrow to add it to the Display list. (You can use the arrow to move selected statistics from the Statistics list into the Display list, or you can drag and drop selected statistics from the Statistics list into the Display list.)

▶ Then click Apply to Selection.

Figure 8-9

Additional statistic applied to one variable in a stacked table
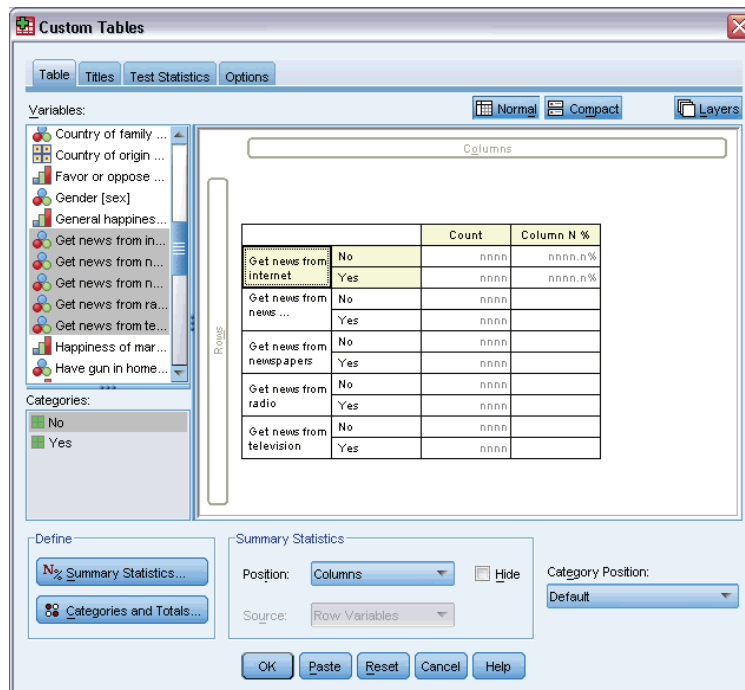
A column is added for column percentages—but the table preview on the canvas pane indicates that column percentages will be displayed for only one variable. This is because in a stacked table there are multiple statistics source variables, and each one can have different summary statistics. In this example, however, we want to display the same summary statistics for all variables.

▶ Right-click *Get news from newspapers* on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select *Column N %* from the Statistics list and click the arrow to add it to the Display list.

▶ Then click Apply to All.

Figure 8-10
Additional statistic applied to all variables in a stacked table
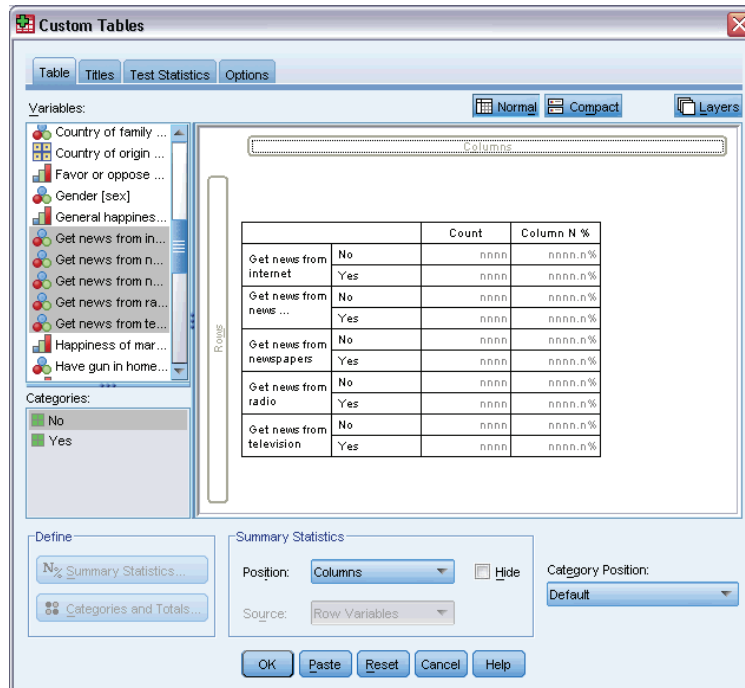
Now the table preview indicates that column percentages will be displayed for all of the stacked variables.

Custom Total Summary Statistics for Categorical Variables

For categorical statistics source variables, you can include custom total summary statistics that are different from the statistics displayed for the categories of the variable. For example, for an ordinal variable, you could display percentages for each category and the mean or median for the custom total summary statistic.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Reset to clear any previous settings in the table builder.

▶ Click *Confidence in press* in the variable list, and then Ctrl-click *Confidence in TV* to select both variables.

▶ Drag and drop the two variables into the Rows area of the canvas pane. This stacks the two variables in the row dimension.

▶ Right-click either variable on the canvas pane and select Select All Row Variables from the pop-up context menu. (They may both already be selected, but we want to make sure.)

▶ Right-click the variable again and select Categories and Totals from the pop-up context menu.

▶ In the Categories and Totals dialog box, click (check) Total, and then click Apply.

The table preview on the canvas pane now displays a total row for both variables. In order to display custom total summary statistics, totals and/or subtotals must be specified for the table.

▶ Right-click either variable on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, click *Count* in the Display list and click the arrow to move it to the Statistics list, removing it from the Display list.

▶ Click *Column N %* in the Statistics list and click the arrow key to move it to the Display list.

▶ Click (check) Custom Summary Statistics for Totals and Subtotals.

▶ Click *Count* in the custom summary Display list and click the arrow to move it to the custom summary Statistics list, removing it from the Display list.

▶ Click *Mean* in the custom summary Statistics list and click the arrow to move it to the custom summary Display list.

▶ Click the Format cell for the mean in the Display list and select nnnn from the drop-down list of formats. (You may have to scroll up the list to find this choice.)

▶ In the Decimals cell, type 2.

Figure 8-11
Selecting custom summary statistics for totals
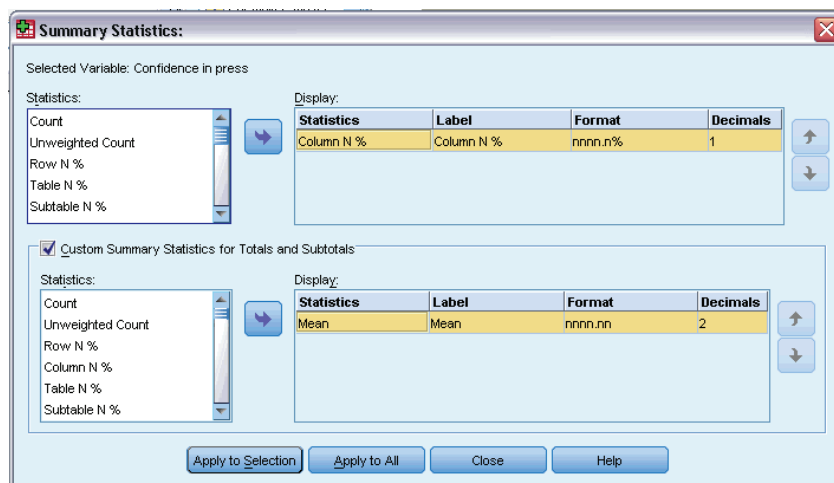
▶ Click Apply to All to apply these settings to both variables in the table.

Figure 8-12
Custom total summary statistics for row variables displayed in columns
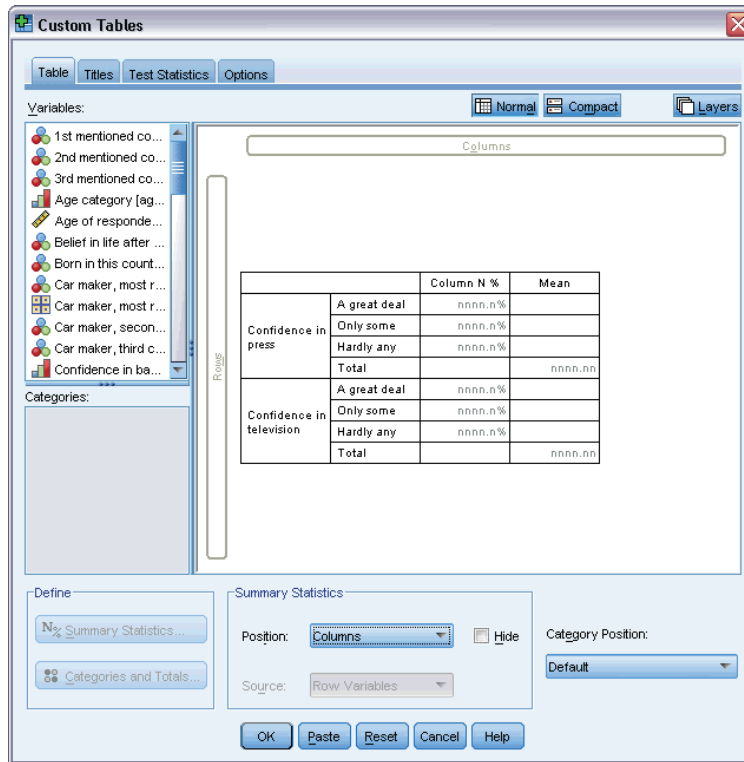
A new column has been added for the custom total summary statistic, which may not be what you want, since the preview on the canvas pane clearly indicates that this will result in a table with many empty cells.

▶ In the table builder, in the Summary Statistics group, select Rows from the Position drop-down list.

Figure 8-13
Summary statistics for row variables displayed in rows
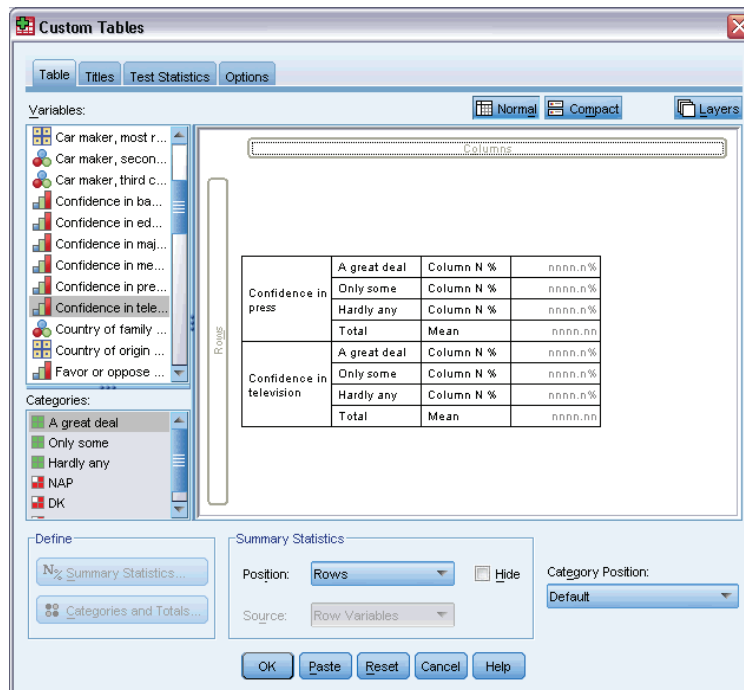
This moves all the summary statistics to the row dimension, displaying all summary statistics in a single column in the table.

▶ Click OK to create the table.

Figure 8-14
Categorical variables with custom total summary statistics

Confidence in press	A great deal	Column N %	9.5%
	Only some	Column N %	47.2%
	Hardly any	Column N %	43.4%
	Total	Mean	2.34
Confidence in television	A great deal	Column N %	10.4%
	Only some	Column N %	49.9%
	Hardly any	Column N %	39.7%
	Total	Mean	2.29

Displaying Category Values

There's only one small problem with the preceding table—it may be hard to interpret the mean value without knowing the underlying category values on which it is based. Is a mean of 2.34 somewhere between *A great deal* and *Only some*—or is it somewhere between *Only some* and *Hardly any*?

Although we can't address this problem directly in Custom Tables, we can address it in a more general way.

▶ From the menus, choose:
Edit > Options...

▶ In the Options dialog box, click the Output Labels tab.

▶ In the Pivot Table Labeling group, select Values and Labels from the Variable values in labels shown as drop-down list.

Figure 8-15
Output labeling options
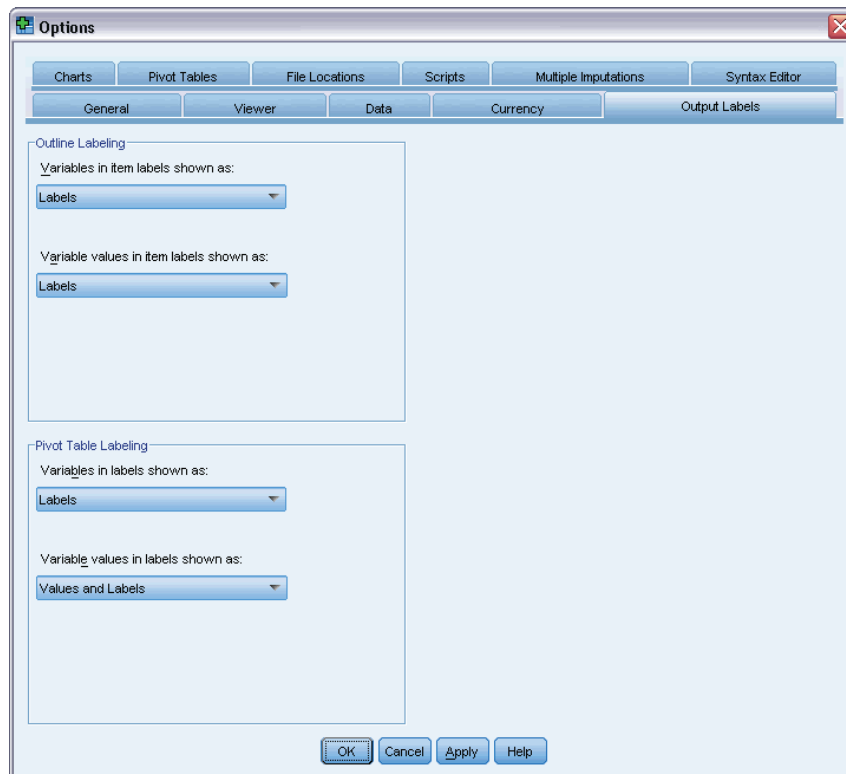
▶ Click OK to save this setting.

▶ Open the table builder (Analyze menu, Tables, Custom Tables) and click OK to create the table again.

Figure 8-16
Values and labels displayed for variable categories

Confidence in press	1 A great deal	Column N %	9.5%
	2 Only some	Column N %	47.2%
	3 Hardly any	Column N %	43.4%
	Total	Mean	2.34
Confidence in television	1 A great deal	Column N %	10.4%
	2 Only some	Column N %	49.9%
	3 Hardly any	Column N %	39.7%
	Total	Mean	2.29

The category values make it clear that a mean of 2.34 is somewhere between *Only some* and *Hardly any*. Displaying the category values in the table makes it much easier to interpret the value of custom total summary statistics, such as the mean.

This display setting is a global setting that affects all pivot table output from all procedures and persists across sessions until you change it. To change the setting back to display only value labels:

▶ From the menus, choose:
Edit > Options...

▶ In the Options dialog box, click the Output Labels tab.

▶ In the Pivot Table Labeling group, select Labels from the Variable values in labels shown as drop-down list.

▶ Click OK to save this setting.

Summarizing Scale Variables

A wide range of summary statistics are available for scale variables. In addition to the counts and percentages available for categorical variables, summary statistics for scale variables also include:

- Mean
- Median
- Percentiles
- Sum
- Standard deviation
- Range
- Minimum and maximum values

For more information, see the topic Summary Statistics for Scale Variables and Categorical Custom Totals in Chapter 2 on p. 32.

Sample Data File

The examples in this chapter use the data file *survey_sample.sav*. For more information, see the topic Sample Files in Appendix A on p. 178.

All examples provided here display variable labels in dialog boxes, sorted in alphabetical order. Variable list display properties are specified on the General tab in the Options dialog box (Edit menu, Options).

Stacked Scale Variables

You can summarize multiple scale variables in the same table by stacking them in the table.

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ In the table builder, click *Age of respondent* in the variable list, Ctrl-click *Highest year of school completed*, and Ctrl-click *Hours per day watching TV* to select all three variables.

▶ Drag and drop the three selected variables to the Rows area of the canvas pane.

Figure 9-1
Stacked scale variables in table builder
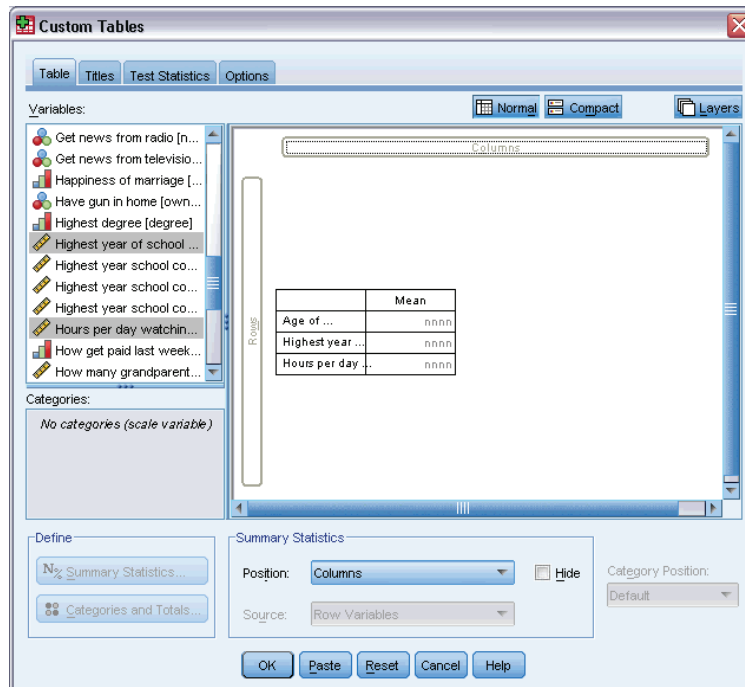
The three variables are stacked in the row dimension. Since all three variables are scale variables, no categories are displayed, and the default summary statistic is the mean.

▶ Click OK to create the table.

Figure 9-2
Table of mean values of stacked scale variables

	Mean
Age of respondent	46
Highest year of school completed	13
Hours per day watching TV	3

Multiple Summary Statistics

By default, the mean is displayed for scale variables; however, you can choose other summary statistics for scale variables, and you can display more than one summary statistic.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click any one of the three scale variables in the table preview on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select *Median* in the Statistics list and click the arrow to add it to the Display list. (You can use the arrow to move selected statistics from the Statistics list to the Display list, or you can drag and drop selected statistics from the Statistics list into the Display list.)

▶ Click the Format cell for the median in the Display list and select nnnn from the drop-down list of formats.

▶ In the Decimals cell, type 1.

▶ Make the same changes for the mean in the Display list.

Figure 9-3
Mean and median selected in Summary Statistics dialog box
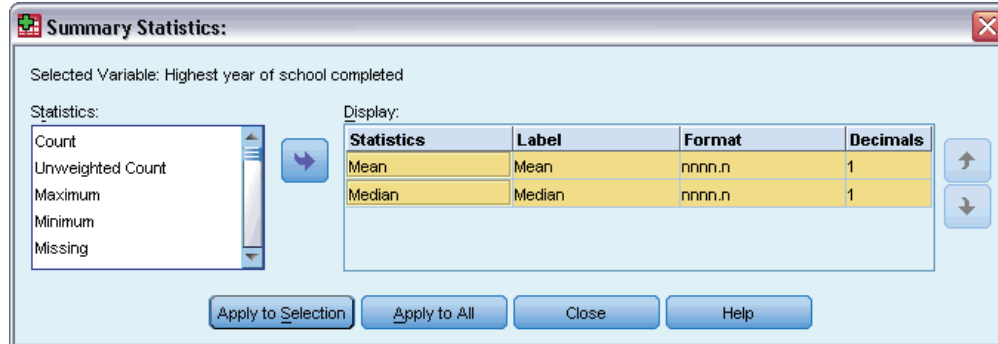
▶ Click Apply to All to apply these changes to all three scale variables.

▶ Click OK in the table builder to create the table.

Figure 9-4
Mean and median displayed in table of stacked scale variables

	Mean	Median
Age of respondent	45.6	42.0
Highest year of school completed	13.3	13.0
Hours per day watching TV	2.9	2.0

Count, Valid N, and Missing Values

It is often useful to display the number of cases used to compute summary statistics, such as the mean, and you might assume (not unreasonably) that the summary statistic *Count* would provide that information. However, this will not give you an accurate case base if there are any missing values. To obtain an accurate case base, use *Valid N*.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click any one of the three scale variables in the table preview on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select Count in the Statistics list and click the arrow to add it to the Display list.

▶ Then select Valid N in the Statistics list and click the arrow to add it to the Display list.

▶ Click Apply to All to apply these changes to all three scale variables.

▶ Click OK in the table builder to create the table.

Figure 9-5
Count versus Valid N

	Mean	Median	Count	Valid N
Age of respondent	45.6	42.0	2832	2828
Highest year of school completed	13.3	13.0	2832	2820
Hours per day watching TV	2.9	2.0	2832	2337

For all three variables, *Count* is the same: 2,832. Not coincidentally, this is the total number of cases in the data file. Since the scale variables aren't nested within any categorical variables, *Count* simply represents the total number of cases in the data file.

Valid N, on the other hand, is different for each variable and differs quite a lot from *Count* for *Hours per day watching TV*. This is because there is a large number of **missing values** for this variable—that is, cases with no value recorded for this variable or values defined as representing missing data (such as a code of 99 to represent *Not Applicable* for pregnancy in males).

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click any one of the three scale variables in the table preview on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select Valid N in the Display list and click the arrow key to move it back to the Statistics list, removing it from the Display list.

▶ Select Count in the Display list and click the arrow key to move it back to the Statistics list, removing it from the Display list.

▶ Select Missing in the Statistics list and click the arrow key to add it to the Display list.

▶ Click Apply to All to apply these changes to all three scale variables.

▶ Click OK in the table builder to create the table.

Figure 9-6
Number of missing values displayed in table of scale summary statistics

	Mean	Median	Missing
Age of respondent	45.6	42.0	4
Highest year of school completed	13.3	13.0	12
Hours per day watching TV	2.9	2.0	495

The table now displays the number of missing values for each scale variable. This makes it quite apparent that *Hours per day watching TV* has a large number of missing values, whereas the other two variables have very few. This may be a factor to consider before putting a great deal of faith in the summary values for that variable.
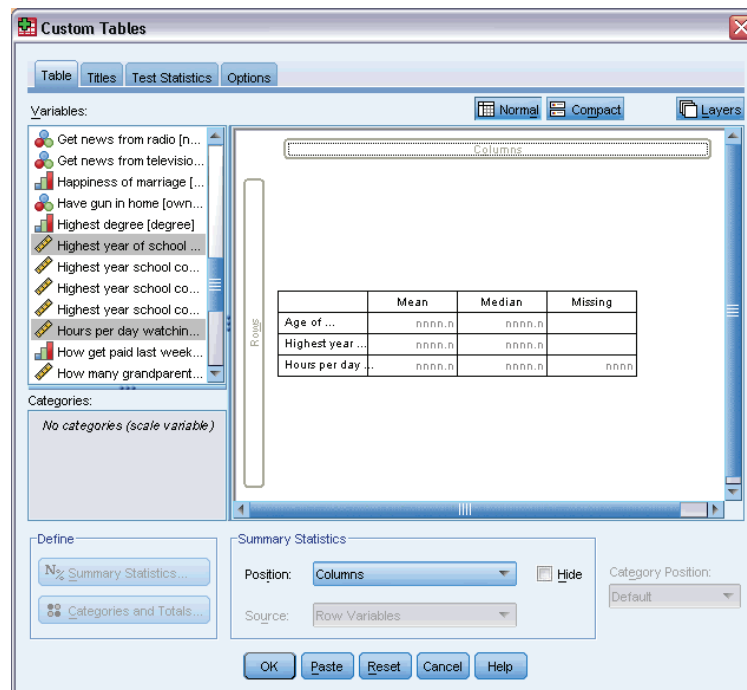
Different Summaries for Different Variables

In addition to displaying multiple summary statistics, you can display different summary statistics for different scale variables in a stacked table. For example, the previous table revealed that only one of the three variables has a large number of missing values; so you might want to show the number of missing values for only that one variable.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click *Age of respondent* in the table preview on the canvas pane, and then Ctrl-click *Highest year of school completed* to select both variables.

▶ Right-click either of the two selected variables and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select Missing in the Display list and click the arrow key to move it back to the Statistics list, removing it from the Display list.

▶ Click Apply to Selection to apply the change to only the two selected variables.

Figure 9-7
Table preview for different summary statistics for different variables

The placeholders in the data cells of the table indicate that the number of missing values will be displayed only for *Hours per day watching TV*.

▶ Click OK to create the table.

Figure 9-8
Table of different summary statistics for different variables

	Mean	Median	Missing
Age of respondent	45.6	42.0	
Highest year of school completed	13.3	13.0	
Hours per day watching TV	2.9	2.0	495

Although this table provides the information that we want, the layout may make it difficult to interpret the table. Somebody reading the table might think that the blank cells in the *Missing* column indicate zero missing values for those variables.

► Open the table builder (Analyze menu, Tables, Custom Tables).

► In the Summary Statistics group in the table builder, select Rows from the Position drop-down list.

Figure 9-9
Moving summary statistics from the column dimension to the row dimension
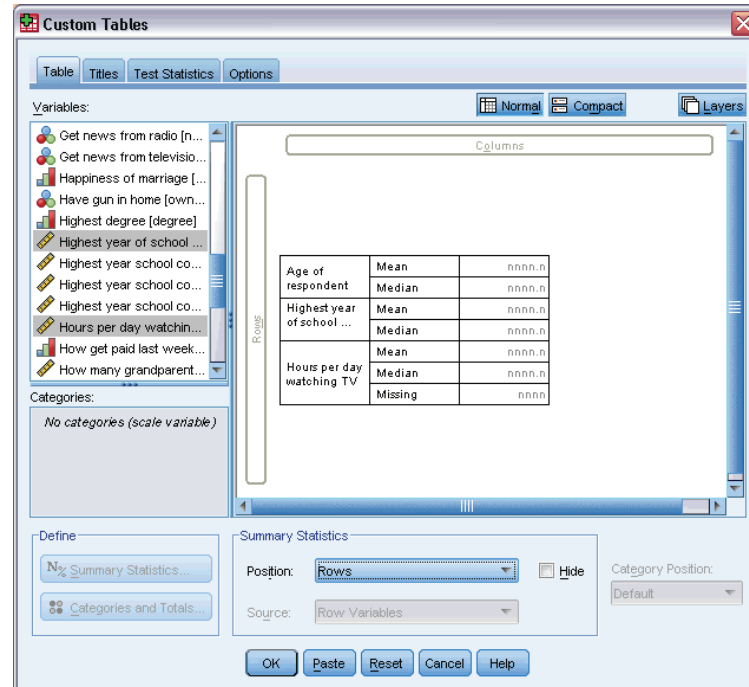
► Click OK to create the table.

Figure 9-10
Summary statistics and variables both displayed in the row dimension

Age of respondent	Mean	45.6
	Median	42.0
Highest year of school completed	Mean	13.3
	Median	13.0
Hours per day watching TV	Mean	2.9
	Median	2.0
	Missing	495

Now it's clear that the table reports the number of missing values for only one variable.

Group Summaries in Categories

You can use categorical variables as grouping variables to display scale variable summaries within groups defined by the categories of the categorical variable.

► Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Drag and drop *Gender* from the variable list into the Columns area of the canvas pane.

If you right-click *Gender* in the table preview on the canvas pane, you will see that Summary Statistics is disabled on the pop-up context menu. This is because in a table with scale variables, the scale variables are always the statistics source variables.

▶ Click OK to create the table.

Figure 9-11
Grouped scale summaries using a categorical column variable

		Gender	
		Male	Female
Age of respondent	Mean	44.6	46.3
	Median	42.0	43.0
Highest year of school completed	Mean	13.4	13.2
	Median	13.0	13.0
Hours per day watching TV	Mean	2.8	2.9
	Median	2.0	2.0
	Missing	213	282

This table makes it easy to compare the averages (mean and median) for males and females, and it clearly shows that there isn't much difference between them—which may not be terribly interesting but might be useful information.

Multiple Grouping Variables

You can subdivide the groups further by nesting and/or using both row and column categorical grouping variables.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Drag and drop *Get news from internet* from the variable list to the far left side of the Rows area of the canvas pane. Make sure to position it so that all three scale variables are nested within it, not just one of them.

Figure 9-12
Correct: All three scale variables nested within the categorical variable

Figure 9-13
Wrong: Only one scale variable nested within the categorical variable

Although there may be times when you want something like the second example above, it's not what we want in this case.

▶ Click OK to create the table.

Figure 9-14
Scale summaries grouped by categorical row and column variables

				Gender	
				Male	Female
Get news from internet	No	Age of respondent	Mean	47.0	48.8
			Median	45.0	46.0
		Highest year of school completed	Mean	13.4	13.1
			Median	13.0	12.0
		Hours per day watching TV	Mean	3.2	3.4
			Median	2.0	3.0
			Missing	213	282
	Yes	Age of respondent	Mean	38.7	41.1
			Median	35.0	38.0
		Highest year of school completed	Mean	13.2	13.3
			Median	13.0	13.0
		Hours per day watching TV	Mean	2.1	2.1
			Median	2.0	2.0
			Missing	0	0

Nesting Categorical Variables within Scale Variables

Although the above table may provide the information you want, it may not provide it in the easiest format to interpret. For example, you can compare the average age of men who use the Internet to get news and those who don't—but it would be easier to do if the values were next to each other rather than separated. Swapping the positions of the two row variables and nesting the categorical grouping variable within the three scale variables might improve the table. With scale variables, nesting level has no effect on the statistics source variable. The scale variable is always the statistics source variable regardless of nesting level.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click *Age of respondent* in the table preview on the canvas pane, Ctrl-click *Highest year of school completed*, and Ctrl-click *Hours per day watching TV* to select all three scale variables.

▶ Drag and drop the three scale variables onto the far left side of the Rows area, nesting the categorical variable *Get news from internet* within each of the three scale variables.

▶ Click OK to create the table.

Figure 9-15
Categorical row variable nested within stacked scale variables

				Gender	
				Male	Female
Age of respondent	Get news from internet	No	Mean	47.0	48.8
			Median	45.0	46.0
		Yes	Mean	38.7	41.1
			Median	35.0	38.0
Highest year of school completed	Get news from internet	No	Mean	13.4	13.1
			Median	13.0	12.0
		Yes	Mean	13.2	13.3
			Median	13.0	13.0
Hours per day watching TV	Get news from internet	No	Mean	3.2	3.4
			Median	2.0	3.0
			Missing	213	282
		Yes	Mean	2.1	2.1
			Median	2.0	2.0
			Missing	0	0

The choice of nesting order depends on the relationships or comparisons that you want to emphasize in the table. Changing the nesting order of the scale variables doesn't change the summary statistics values; it changes only their relative positions in the table.

Test Statistics

Three different tests of significance are available for studying the relationship between row and column variables. This chapter discusses the output of each of these tests, with special attention to the effects of nesting and stacking. For more information, see the topic Stacking, Nesting, and Layers with Categorical Variables in Chapter 4 on p. 61.

Sample Data File

The examples in this chapter use the data file *survey_sample.sav*. For more information, see the topic Sample Files in Appendix A on p. 178.

Tests of Independence (Chi-Square)

The chi-square test of independence is used to determine whether there is a relationship between two categorical variables. For example, you may want to determine whether *Labor force status* is related to *Marital status*.

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ In the table builder, drag and drop *Labor force status* from the variable list into the Rows area of the canvas pane.

▶ Drag and drop *Marital status* from the variable list into the Columns area.

Figure 10-1
Variables displayed on canvas pane
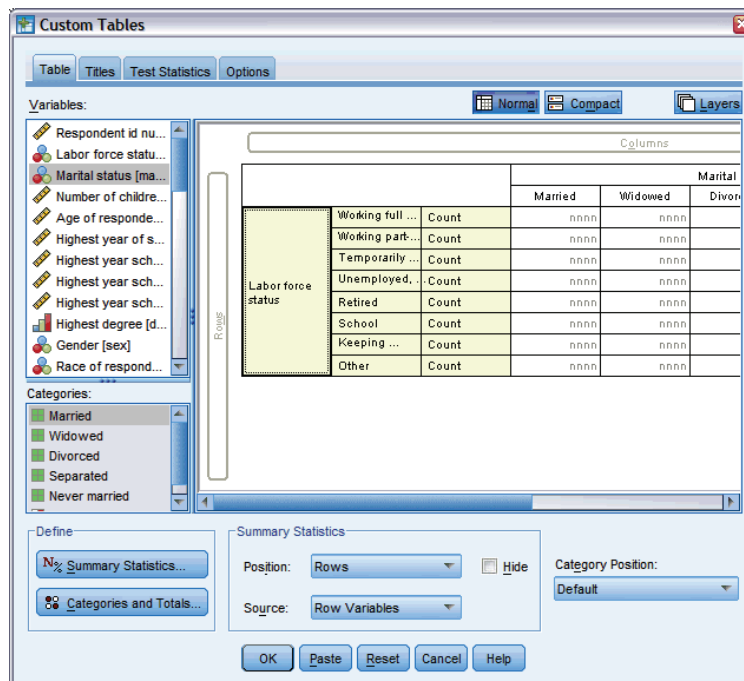
▶ Select Rows as the position for the summary statistics.

▶ Select *Labor force status* and click Summary Statistics in the Define group.

Figure 10-2
Summary Statistics dialog box
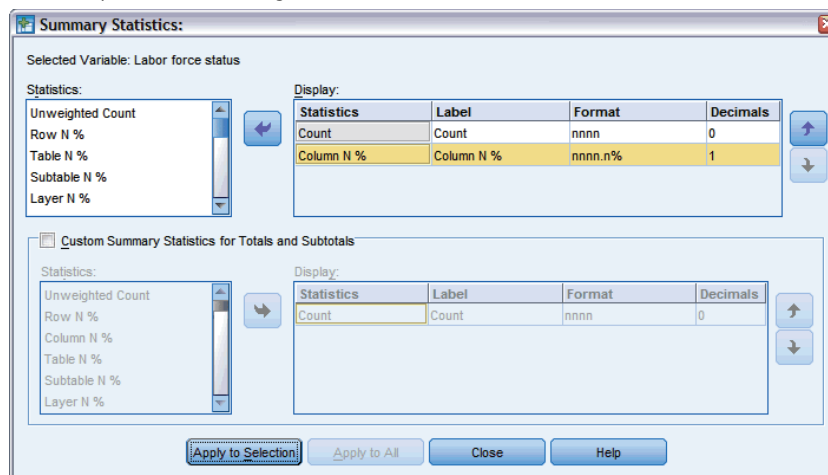
▶ Select Column N % in the Statistics list and add it to the Display list.

▶ Click Apply to Selection.

▶ In the Custom Tables dialog box, click the Test Statistics tab.

Figure 10-3
Test Statistics tab with the Tests of independence (chi-square) selected
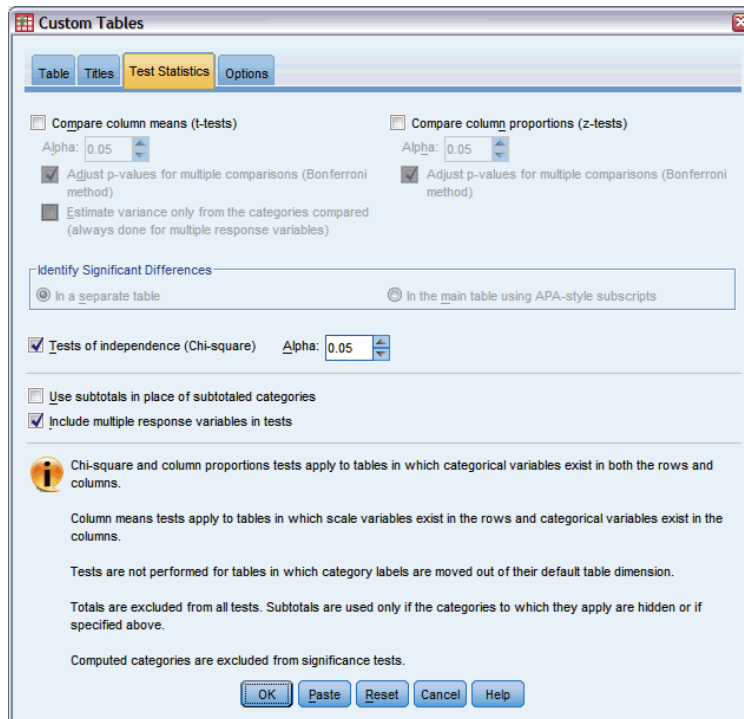
▶ Select Tests of independence (Chi-square).

▶ Click OK to create the table and obtain the chi-square test.

Figure 10-4
Labor force status by Marital status

			Marital status				
			Married	Widowed	Divorced	Separated	Never married
Labor force status	Working full time	Count	778	44	295	58	392
		Column %	57.8%	15.5%	66.1%	62.4%	59.1%
	Working part-time	Count	138	20	35	9	102
		Column %	10.3%	7.1%	7.8%	9.7%	15.4%
	Temporarily not working	Count	23	2	9	1	11
		Column %	1.7%	.7%	2.0%	1.1%	1.7%
	Unemployed, laid off	Count	13	3	10	0	32
		Column %	1.0%	1.1%	2.2%	.0%	4.8%
	Retired	Count	168	150	53	6	17
		Column %	12.5%	53.0%	11.9%	6.5%	2.6%
	School	Count	9	1	7	2	60
		Column %	.7%	.4%	1.6%	2.2%	9.0%
	Keeping house	Count	200	55	25	13	35
		Column %	14.9%	19.4%	5.6%	14.0%	5.3%
	Other	Count	16	8	12	4	14
		Column %	1.2%	2.8%	2.7%	4.3%	2.1%

This table is a crosstabulation of *Labor force status* by *Marital status*, with counts and column proportions shown as the summary statistics. Column proportions are computed so that they sum to 100% down each column. If these two variables are unrelated, then in each row the proportions

should be similar across columns. There appear to be differences in the proportions, but you can check the chi-square test to be sure.

Figure 10-5
Pearson's chi-square test

		Marital status
Labor force status	Chi-square	729.242
	df	28
	Sig.	.000*

*. The Chi-square statistic is significant at the 0.05 level.

The test of independence hypothesizes that *Labor force status* and *Marital status* are unrelated—that is, that the column proportions are the same across columns, and any observed discrepancies are due to chance variation. The chi-square statistic measures the overall discrepancy between the observed cell counts and the counts you would expect if the column proportions were the same across columns. A larger chi-square statistic indicates a greater discrepancy between the observed and expected cell counts—greater evidence that the column proportions are not equal, that the hypothesis of independence is incorrect, and, therefore, that *Labor force status* and *Marital status* are related.

The computed chi-square statistic has a value of 729.242. In order to determine whether this is enough evidence to reject the hypothesis of independence, the significance value of the statistic is computed. The significance value is the probability that a random variate drawn from a chi-square distribution with 28 degrees of freedom is greater than 729.242. Since this value is less than the alpha level specified on the Test Statistics tab, you can reject the hypothesis of independence at the 0.05 level. Thus, *Labor force status* and *Marital status* are in fact related.

Effects of Nesting and Stacking on Tests of Independence

The rule for tests of independence is as follows: a separate test is performed for each innermost subtable. To see how nesting affects the tests, consider the previous example, but with *Marital status* nested within levels of *Gender*.

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Drag and drop *Gender* from the variable list into the Columns area of the canvas pane above *Marital status*.

▶ Click OK to create the table.

Figure 10-6
Pearson's chi-square test

		Gender	
		Male	Female
		Marital status	Marital status
Labor force status	Chi-square	246.637	542.589
	df	28	28
	Sig.	.000[*,1,2]	.000[*,1,2]

[*] The Chi-square statistic is significant at the 0.05 level.

[1] More than 20% of cells in this sub-table have expected cell counts less than 5.

[2] The minimum expected cell count in this sub-table is less than one.

With *Marital status* nested within levels of *Gender*, two tests are performed—one for each level of *Gender*. The significance value for each test indicates that you can reject the hypothesis of independence between *Marital status* and *Labor force status* for both males and females. However, the table notes that more than 20% of each table's cells have expected counts of less than 5, and the minimum expected cell count is less than 1. These notes indicate that the assumptions of the chi-square test may not be met by these tables, so the results of the tests are suspect.

Note: The footnotes may be cut off from view by the cell boundaries. You can make them visible by changing the alignment of these cells in the Cell Properties dialog box.

To see how stacking affects the tests:

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Drag and drop *Highest degree* from the variable list into the Rows area below *Labor force status*.

▶ Click OK to create the table.

Figure 10-7
Pearson's chi-square test

		Gender	
		Male	Female
		Marital status	Marital status
Labor force status	Chi-square	246.637	542.589
	df	28	28
	Sig.	.000[*,1,2]	.000[*,1,2]
Highest degree	Chi-square	43.844	105.506
	df	16	16
	Sig.	.000[*]	.000[*]

[*] The Chi-square statistic is significant at the 0.05 level.

[1] More than 20% of cells in this sub-table have expected cell counts less than 5.

[2] The minimum expected cell count in this sub-table is less than one.

With *Highest degree* stacked with *Labor force status*, four tests are performed—a test of the independence of *Marital status* and *Labor force status*, and a test of *Marital status* and *Highest degree* for each level of *Gender*. The test results for *Marital status* and *Labor force status* are the

same as before. The test results for *Marital status* and *Highest degree* indicate these variables are not independent.

Comparing Column Means

The column means tests are used to determine whether there is a relationship between a categorical variable in the Columns and a continuous variable in the Rows. Moreover, you can use the test results to determine the relative ordering of categories of the categorical variable in terms of the mean value of the continuous variable. For example, you may want to determine whether *Hours per day watching TV* is related to *Get news from newspapers*.

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ Click Reset to restore the default settings to all tabs.

▶ In the table builder, drag and drop *Hours per day watching TV* from the variable list into the Rows area of the canvas pane.

▶ Drag and drop *Get news from newspapers* from the variable list into the Columns area.

Figure 10-8
Variables displayed on canvas pane
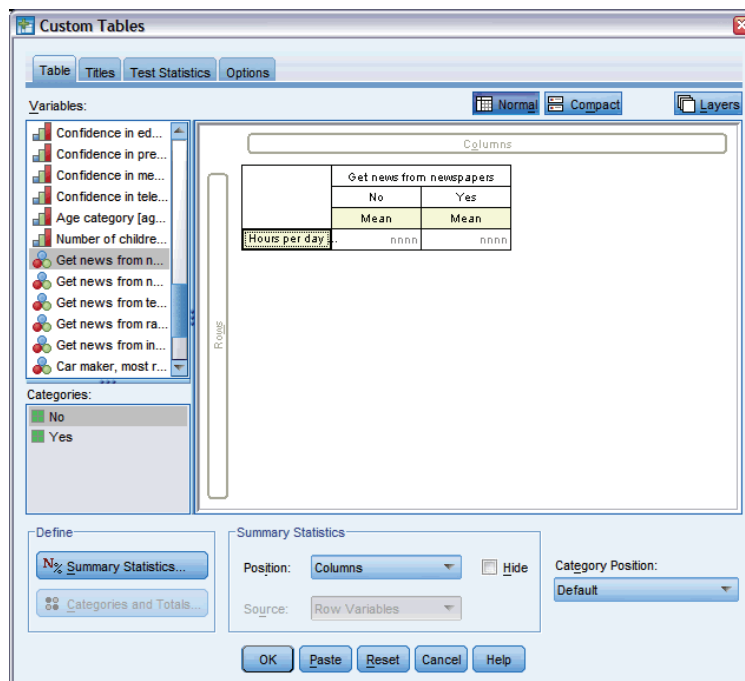
▶ Select *Hours per day watching TV* and click Summary Statistics in the Define group.

Figure 10-9
Summary Statistics dialog box
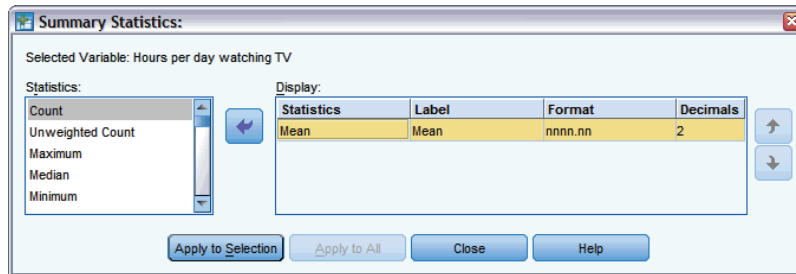
▶ Select nnnn as the format.

▶ Select 2 as the number of decimals to display. Notice that this causes the format to now read nnnn.nn.

▶ Click Apply to Selection.

▶ In the Custom Tables dialog box, click the Test Statistics tab.

Figure 10-10
Test Statistics tab with Compare column means (t tests) selected
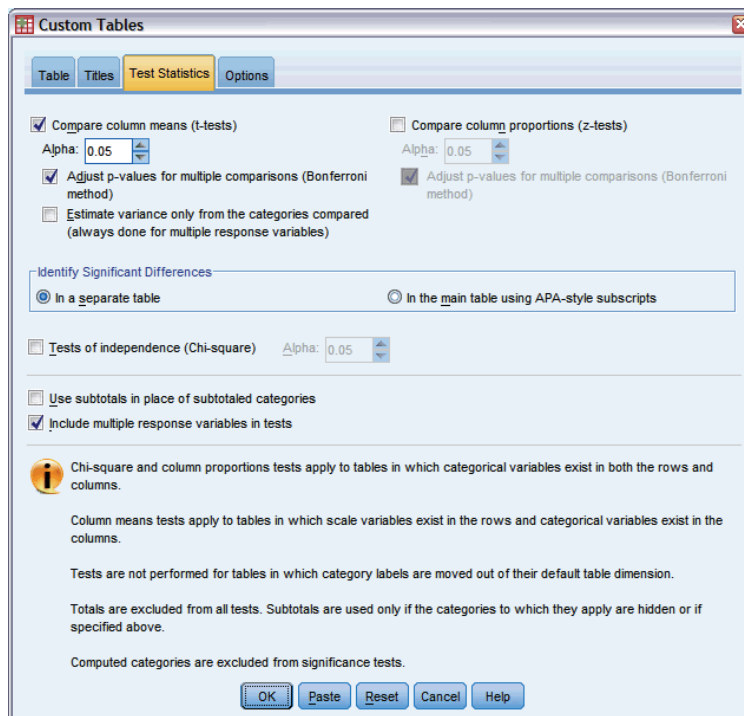
▶ Select Compare column means (t-tests).

▶ Click OK to create the table and obtain the column means tests.

Figure 10-11
Get news from newspapers by Hours per day watching TV

	Get news from newspapers	
	No	Yes
	Mean	Mean
Hours per day watching TV	2.92	2.74

This table shows the mean *Hours per day watching TV* for people who do and do not get their news from newspapers. The observed difference in these means suggests that people who do not get their news from newspapers spend approximately 0.18 more hours watching TV than people who do get their news from newspapers. To see whether this difference is due to chance variation, check the column means tests.

Figure 10-12
Comparisons of column means

	Get news from newspapers	
	No	Yes
	(A)	(B)
Hours per day watching TV		

The column means test table assigns a letter key to each category of the column variable. For *Get news from newspapers*, the category *No* is assigned the letter A, and *Yes* is assigned the letter B. For each pair of columns, the column means are compared using a *t* test. Since there are only two columns, only one test is performed. For each significant pair, the key of the category with the smaller mean is placed under the category with larger mean. Since no keys are reported in the cells of the table, this means that the column means are not statistically different.

Significance Results in APA-style Notation

If you do not want the significance results in a separate table, you can choose to display them in the main table. Significance results are identified using an APA-style notation with subscript letters. Complete the previous steps for comparing column means, but make the following change on the Test Statistics tab:

▶ In the Identify Significant Differences area, select In the main table using APA-style subscripts.

Figure 10-13
Test Statistics tab with Compare column means (t tests) and In the main table using subscripts selected
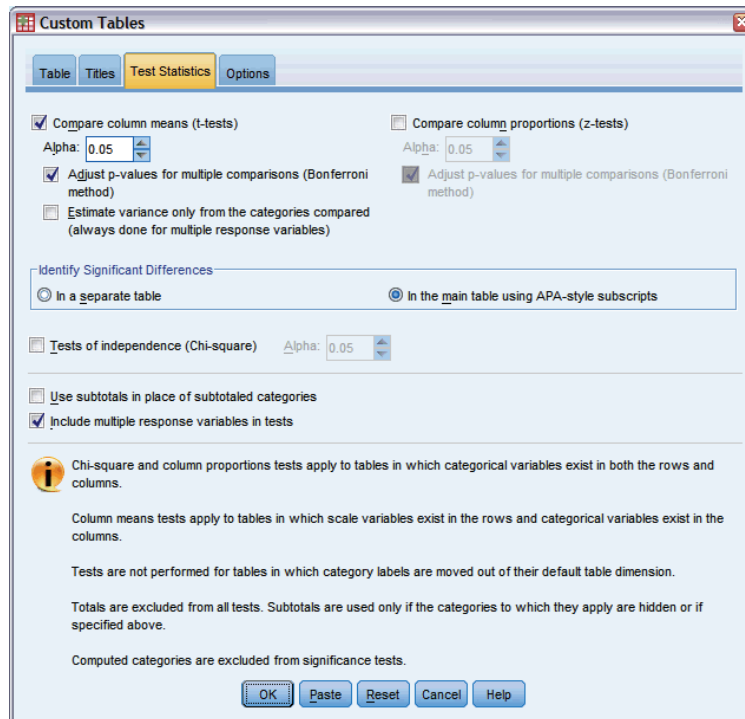
▶ Click OK to create the table and obtain the column means tests using APA-style notation.

Figure 10-14
Comparisons of column means using APA-style notation

	Get news from newspapers	
	No	Yes
	Mean	Mean
Hours per day watching TV	2.92_a	2.74_a

The column means test table assigns a subscript letter to the categories of the column variable. For each pair of columns, the column means are compared using a *t* test. If a pair of values is significantly different, the values have *different* subscript letters assigned to them. Since there are only two columns, only one test is performed. Because the column means in this example share the same subscript letter, the column means are not statistically different.

Effects of Nesting and Stacking on Column Means Tests

The rule for column means tests is as follows: a separate set of pairwise tests is performed for each innermost subtable. To see how nesting affects the tests, consider the previous example, but with *Hours per day watching TV* nested within levels of *Labor force status*.

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Drag and drop *Labor force status* from the variable list into the Rows area of the canvas pane.

▶ Click OK to create the table.

Figure 10-15
Comparisons of column means

			Get news from newspapers	
			No	Yes
			(A)	(B)
Labor force status	Working full time	Hours per day watching TV	B	
	Working part-time	Hours per day watching TV		
	Temporarily not working	Hours per day watching TV		
	Unemployed, laid off	Hours per day watching TV		
	Retired	Hours per day watching TV		
	School	Hours per day watching TV		
	Keeping house	Hours per day watching TV		
	Other	Hours per day watching TV		

With *Hours per day watching TV* nested within levels of *Labor force status*, seven sets of column means tests are performed: one for each level of *Labor force status*. The same letter keys are assigned to the categories of *Get news from newspapers*. For respondents *working full time*, the B key appears in the A column. This means that for full-time employees, the mean value of *Hours per day watching TV* is lower for people who get their news from newspapers. No other keys appear in the columns, so you can conclude that there are no other statistically significant differences in the column means.

Bonferroni adjustments. When multiple tests are performed, the Bonferroni adjustment is applied to column means tests to ensure that the alpha level (or false positive rate) specified on the Test Statistics tab applies to each *set* of tests. Thus, in this table, no Bonferroni adjustments were applied because although seven sets of tests are performed, within each set only one pair of columns is compared.

To see how stacking affects the tests:

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Drag and drop *Get news from internet* from the variable list into the Columns area to the left of *Get news from newspapers*.

▶ Click OK to create the table.

Figure 10-16
Comparisons of column means

|			Get news from internet		Get news from newspapers	
			No	Yes	No	Yes
			(A)	(B)	(A)	(B)
Labor force status	Working full time	Hours per day watching TV	B		B	
	Working part-time	Hours per day watching TV	B			
	Temporarily not working	Hours per day watching TV				
	Unemployed, laid off	Hours per day watching TV	B			
	Retired	Hours per day watching TV	B			
	School	Hours per day watching TV	B			
	Keeping house	Hours per day watching TV	B			
	Other	Hours per day watching TV	B			

With *Get news from internet* stacked with *Get news from newspapers*, 14 sets of column means tests are performed—one for each level of *Labor force status* for *Get news from internet* and *Get news from newspapers*. Again, no Bonferroni adjustments are applied because within each set, only one pair of columns is compared. The tests for *Get news from newspapers* are the same as before. For *Get news from internet*, the category *No* is assigned the letter A and *Yes* is assigned the letter B. The B key is reported in the A column for each set of column means tests except for those respondents temporarily not working. This means that the mean value of *Hours per day watching TV* is lower for people who get their news from the Internet than for people who do not get their news from newspapers. No keys are reported for the *Temporarily not working* set; thus, the column means are not statistically different for these respondents.

Comparing Column Proportions

The column proportions tests are used to determine the relative ordering of categories of the Columns categorical variable in terms of the category proportions of the Rows categorical variable. For example, after using a chi-square test to find that *Labor force status* and *Marital status* are not independent, you may want to see which rows and columns are responsible for this relationship.

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ Click Reset to restore the default settings to all tabs.

▶ In the table builder, drag and drop *Labor force status* from the variable list into the Rows area of the canvas pane.

▶ Drag and drop *Marital status* from the variable list into the Columns area.

Figure 10-17
Variables displayed on canvas pane
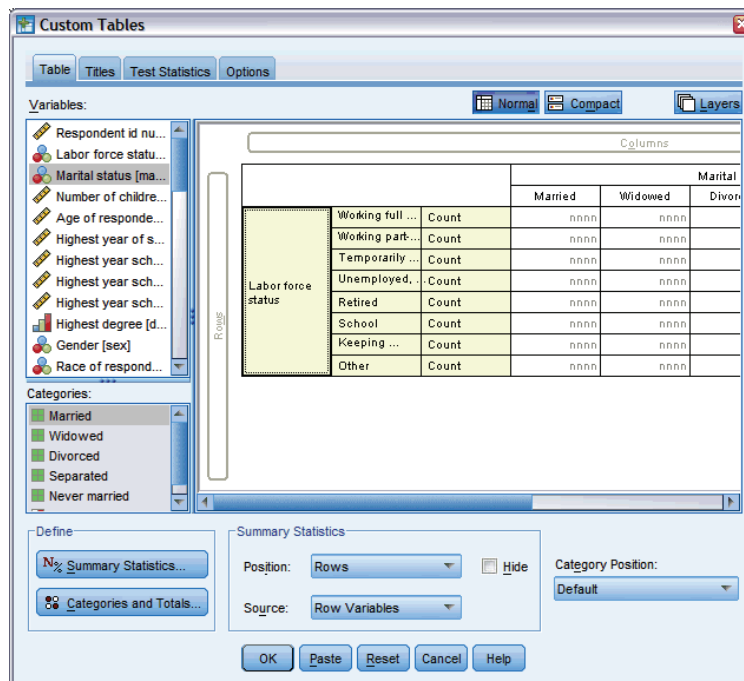
▶ Select *Labor force status* and click Summary Statistics in the Define group.

Figure 10-18
Summary Statistics dialog box
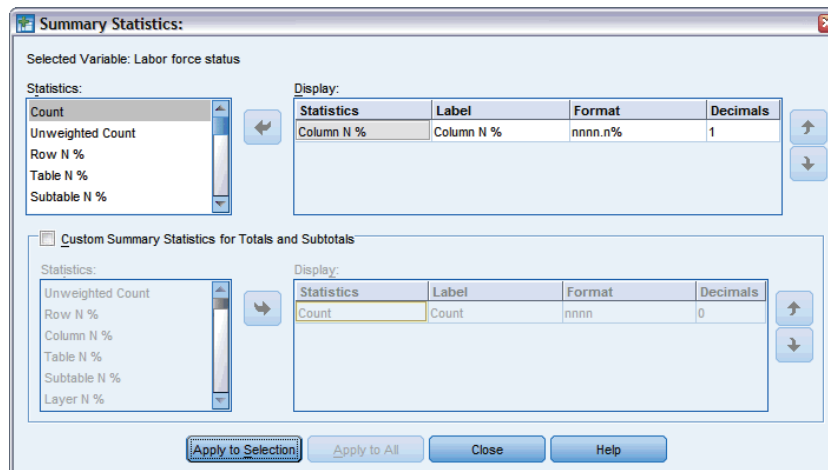
▶ Select Column N % in the Statistics list and add it to the Display list.

▶ Deselect Count from the Display list.

▶ Click Apply to Selection.

▶ In the Custom Tables dialog box, click the Test Statistics tab.

Figure 10-19
Test Statistics tab with Compare column proportions (z tests) selected

▶ Select Compare column proportions (z-tests).

▶ Click OK to create the table and obtain the column proportions tests.

Figure 10-20
Labor force status by Marital status

		Marital status				
		Married	Widowed	Divorced	Separated	Never married
		Column %	Column %	Column %	Column %	Column %
Labor force status	Working full time	57.8%	15.5%	66.1%	62.4%	59.1%
	Working part-time	10.3%	7.1%	7.8%	9.7%	15.4%
	Temporarily not working	1.7%	.7%	2.0%	1.1%	1.7%
	Unemployed, laid off	1.0%	1.1%	2.2%	.0%	4.8%
	Retired	12.5%	53.0%	11.9%	6.5%	2.6%
	School	.7%	.4%	1.6%	2.2%	9.0%
	Keeping house	14.9%	19.4%	5.6%	14.0%	5.3%
	Other	1.2%	2.8%	2.7%	4.3%	2.1%

This table is a crosstabulation of *Labor force status* by *Marital status*, with column proportions shown as the summary statistic.

Figure 10-21
Comparisons of column proportions

		Marital status				
		Married	Widowed	Divorced	Separated	Never married
		(A)	(B)	(C)	(D)	(E)
Labor force status	Working full time	B		A B	B	B
	Working part-time					A B C
	Temporarily not working					
	Unemployed, laid off				.	A B
	Retired	E	A C D E	E		
	School					A B C
	Keeping house	C E	C E		C E	
	Other					

The column proportions test table assigns a letter key to each category of the column variables. For *Marital status*, the category *Married* is assigned the letter A, *Widowed* is assigned the letter B, and so on, through the category *Never married*, which is assigned the letter E. For each pair of columns, the column proportions are compared using a *z* test. Seven sets of column proportions tests are performed, one for each level of *Labor force status*. Since there are five levels of *Marital status*, (5*4)/2 = 10 pairs of columns are compared in each set of tests, and Bonferroni adjustments are used to adjust the significance values. For each significant pair, the key of the smaller category is placed under the category with the larger proportion.

For the set of tests associated with *Working full time*, the B key appears in each of the other columns. Also, the A key appears in the C column. No other keys are reported in other columns. Thus, you can conclude that the proportion of divorced persons who are working full time is greater than the proportion of married persons working full time, which in turn is greater than the proportion of widowers working full time. The proportions of people who are separated or never married and working full time cannot be differentiated from people who are divorced or married and working full time, but these proportions are greater than the proportion of widowers working full time.

For the tests associated with *Working part time* or *School*, the A, B, and C keys appear in the E column. No other keys are reported in other columns. Thus, the proportions of people who have never been married and are in school or are working part time are greater than the proportions of married, widowed, or divorced people who are in school or working part time.

For the tests associated with *Temporarily not working* or with *Other* labor status, no other keys are reported in any columns. Thus, there is no discernible difference in the proportions of married, widowed, divorced, separated, or never-married people who are temporarily not working or are in an otherwise uncategorized employment situation.

The tests associated with *Retired* show that the proportion of widowers who are retired is greater than the proportions of all other marital categories who are retired. Moreover, the proportions of married or divorced people who are retired is greater than the proportion of never-married persons who are retired.

There are greater proportions of people married, widowed, or separated and keeping house than proportions of people divorced or never married and keeping house.

The proportion of people who have never been married and are *Unemployed, laid off* is higher than the proportions of people who are married or widowed and unemployed. Also, note that the *Separated* column is marked with a ".", which indicates that the observed proportion of separated people in the *Unemployed, laid off* row is either 0 or 1, and therefore no comparisons can be made using that column for unemployed respondents.

Significance Results in APA-style Notation

If you do not want the significance results in a separate table, you can choose to display them in the main table. Significance results are identified using an APA-style notation with subscript letters. Complete the previous steps for comparing column proportions, but make the following change on the Test Statistics tab:

▶ In the Identify Significant Differences area, select In the main table using APA-style subscripts.

Figure 10-22
Test Statistics tab with Compare column proportions (z tests) and In the main table using subscripts selected
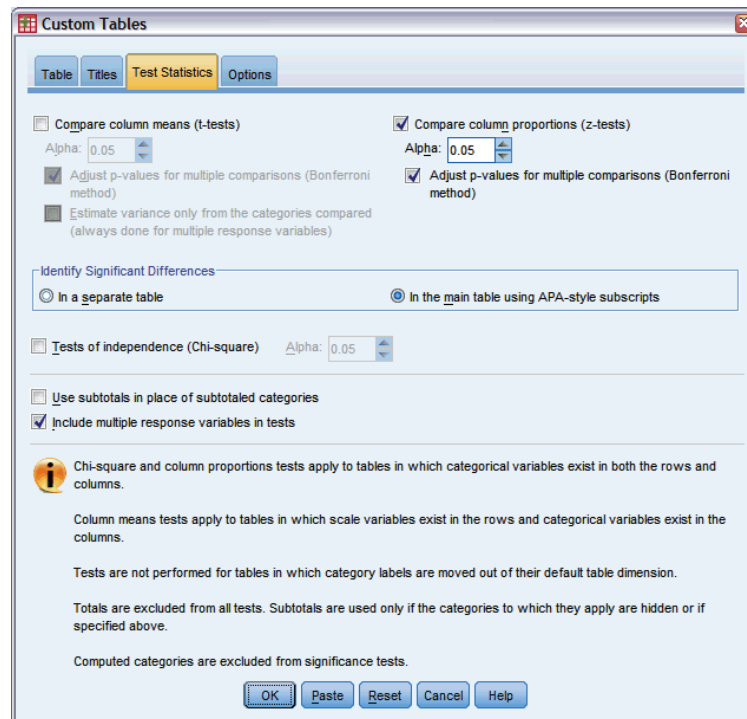
▶ Click OK to create the table and obtain the column means tests using APA-style notation.

Figure 10-23

Comparisons of column proportions using APA-style notation

		Marital status				
		Married	Widowed	Divorced	Separated	Never married
		Column N %	Column N %	Column N %	Column N %	Column N %
Labor force status	Working full time	57.8%$_a$	15.5%$_b$	66.1%$_c$	62.4%$_{a,c}$	59.1%$_{a,c}$
	Working part-time	10.3%$_a$	7.1%$_a$	7.8%$_a$	9.7%$_{a,b}$	15.4%$_b$
	Temporarily not working	1.7%$_a$	.7%$_a$	2.0%$_a$	1.1%$_a$	1.7%$_a$
	Unemployed, laid off	1.0%$_a$	1.1%$_a$	2.2%$_{a,b}$	.0%	4.8%$_b$
	Retired	12.5%$_a$	53.0%$_b$	11.9%$_a$	6.5%$_{a,c}$	2.6%$_c$
	School	.7%$_a$	.4%$_a$	1.6%$_a$	2.2%$_{a,b}$	9.0%$_b$
	Keeping house	14.9%$_a$	19.4%$_a$	5.6%$_b$	14.0%$_a$	5.3%$_b$
	Other	1.2%$_a$	2.8%$_a$	2.7%$_a$	4.3%$_a$	2.1%$_a$

The column proportions test table assigns a subscript letter to the categories of the column variable. For each pair of columns, the column proportions are compared using a *z* test. If a pair of values is significantly different, the values have *different* subscript letters assigned to them.

For the set of tests associated with *Working full time*, the Widowed category has a subscript letter not used in the other columns. Also, the Married category and Divorced category share the same subscript letters, as do the Separated and Never Married categories. Considering these subscript letters and the actual proportions shown in the table, you can make the same analysis as demonstrated in the previous example with separate tables. Thus, you can conclude that the proportion of divorced persons who are working full time is greater than the proportion of married persons working full time, which in turn is greater than the proportion of widowers working full time. The proportions of people who are separated or never married and working full time cannot be differentiated from people who are divorced or married and working full time, but these proportions are greater than the proportion of widowers working full time. The rest of the analysis from the previous example applies.

Effects of Nesting and Stacking on Column Proportions Tests

The rule for column proportions tests is as follows: a separate set of pairwise tests is performed for each innermost subtable. To see how nesting affects the tests, consider the previous example, but with *Labor force status* nested within levels of *Gender*.

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Drag and drop *Gender* from the variable list into the Rows area of the canvas pane.

▶ Click OK to create the table.

148

Chapter 10

Figure 10-24
Comparisons of column proportions

				Married (A)	Widowed (B)	Divorced (C)	Separated (D)	Never married (E)
Gender	Male	Labor force status	Working full time	B		B	B	B
			Working part-time					A
			Temporarily not working		.			
			Unemployed, laid off		.		.	A
			Retired	E	A C D E	E		
			School		.			A C
			Keeping house					
			Other				A	
	Female	Labor force status	Working full time	B		A B	B	B
			Working part-time	B				B
			Temporarily not working				.	
			Unemployed, laid off				.	A
			Retired	E	A C D E	E		
			School					A B C
			Keeping house	C E	C E		C	
			Other					

With *Labor force status* nested within levels of *Gender*, 14 sets of column proportions tests are performed—one for each level of *Labor force status* for each level of *Gender*. The same letter keys are assigned to the categories of *Marital status*.

There are a couple of things to note about the table results:

■ With more tests, there are more columns with zero column proportion. They are most common among separated respondents and widowed males.

■ The column differences previously seen among respondents *keeping house* seems to be entirely due to females.

To see how stacking affects the tests:

▶ Open the table builder again (Analyze menu, Tables, Custom Tables).

▶ Drag and drop *Highest degree* from the variable list into the Rows area below *Gender*.

▶ Click OK to create the table.

Figure 10-25
Comparisons of column proportions

				Marital status				
				Married	Widowed	Divorced	Separated	Never married
				(A)	(B)	(C)	(D)	(E)
Gender	Male	Labor force status	Working full time	B		B	B	B
			Working part-time					A
			Temporarily not working		.			
			Unemployed, laid off		.		.	A
			Retired	E	A C D E	E		
			School		.			A C
			Keeping house					
			Other				A	
	Female	Labor force status	Working full time	B		A B	B	B
			Working part-time	B				B
			Temporarily not working				.	
			Unemployed, laid off				.	A
			Retired	E	A C D E	E		
			School					A B C
			Keeping house	C E	C E		C	
			Other					
Highest degree	LT High school				A C E			
	High school							
	Junior college			B		B		B
	Bachelor			B				B
	Graduate			B				

With *Highest degree* stacked with *Gender*, 19 sets of column means tests are performed—the 14 previously discussed plus one for each level of *Highest degree*. The same letter keys are assigned to the categories of *Marital status*.

There are a few things to note about the table results:

■ The test results for the 14 previously run sets of tests are the same.

■ People who have less than a high school degree are more common among widowers than among married, divorced, or never-married respondents.

■ People with some post-high school education tend to be more common among those people who are married, divorced, and never married than among widowers.

A Note on Weights and Multiple Response Sets

Case weights are always based on counts, not responses, even when one of the variables is a multiple response variable.

Multiple Response Sets

Custom Tables and the Chart Builder support a special kind of "variable" called a **multiple response set**. Multiple response sets aren't really "variables" in the normal sense. You can't see them in the Data Editor, and other procedures don't recognize them. Multiple response sets use multiple variables to record responses to questions where the respondent can give more than one answer. Multiple response sets are treated like categorical variables, and most of the things you can do with categorical variables, you can also do with multiple response sets.

Multiple response sets are constructed from multiple variables in the data file. A multiple response set is a special construct within a data file. You can define and save multiple response sets in IBM® SPSS® Statistics data files, but you cannot import or export multiple response sets from/to other file formats. You can copy multiple response sets from other SPSS Statistics data files using Copy Data Properties, which is accessed from the Data menu in the Data Editor window.

Sample Data File

The examples in this chapter use the data file *survey_sample.sav*. For more information, see the topic Sample Files in Appendix A on p. 178.

All examples provided here display variable labels in dialog boxes, sorted in alphabetical order. Variable list display properties are specified on the General tab in the Options dialog box (Edit menu, Options).

Counts, Responses, Percentages, and Totals

All of the summary statistics available for categorical variables are also available for multiple response sets. Some additional statistics are also available for multiple response sets.

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ Drag and drop *News sources* (this is the descriptive label for the multiple response set *$mltnews*) from the variable list into the Rows area of the canvas pane.

The icon next to the "variable" in the variable list identifies it as a multiple dichotomy set.

Figure 11-1
Multiple dichotomy set icon

Figure 11-2
Multiple dichotomy set displayed in table preview
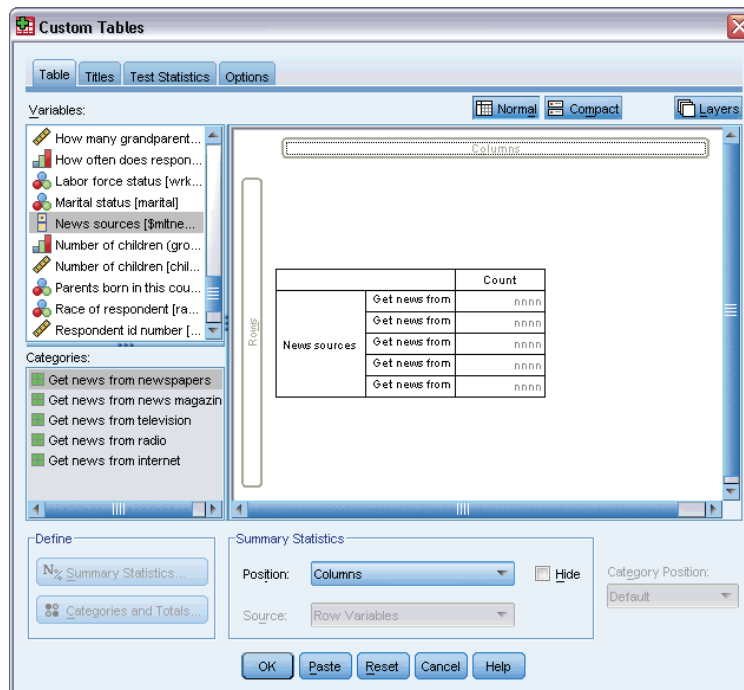
For a multiple dichotomy set, each "category" is, in fact, a separate variable, and the category labels are the variable labels (or variable names for variables without defined variable labels). In this example, the counts that will be displayed represent the number of cases with a *Yes* response for each variable in the set.

► Right-click *News sources* in the table preview on the canvas pane and select Categories and Totals from the pop-up context menu.

► Select (click) Total in the Categories and Totals dialog box, and then click Apply.

► Right-click *News sources* again and select Summary Statistics from the pop-up context menu.

► In the Summary Statistics dialog box, select Column N % in the Statistics list and click the arrow to add it to the Display list.

► Click Apply to Selection, and then click OK to create the table.

Figure 11-3
Multiple dichotomy counts and column percentages

		Count	Column N %
News sources	Get news from internet	867	41.7%
	Get news from radio	551	26.5%
	Get news from television	1077	51.8%
	Get news from news magazines	294	14.1%
	Get news from newspapers	805	38.7%
	Total	2081	100.0%

Totals That Don't Add Up

If you look at the numbers in the table, you may notice that there is a fairly large discrepancy between the "totals" and the values that are supposedly being totaled — specifically, the totals appear to be much lower than they should be. This is because the count for each "category" in the table is the number of cases with a value of 1 (a *Yes* response) for that variable, and the total number of *Yes* responses for all five variables in the multiple dichotomy set might easily exceed the total number of cases in the data file.

The total "count," however, is the total number of cases with a *Yes* response for at least one variable in the set, which can never exceed the total number of cases in the data file. In this example, the total count of 2,081 is almost 800 lower than the total number of cases in the data file. If none of these variables have missing values, this means that almost 800 survey respondents indicated that they don't get news from any of those sources. The total count is the base for the column percentages; so the column percentages in this example sum to more than the 100% displayed for the total column percentage.

Totals That Do Add Up

While "count" is typically a fairly unambiguous term, the above example demonstrates how it could be confusing in the context of totals for multiple response sets, for which *responses* is often the summary statistic you really want.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click *News sources* in the table preview on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select Responses in the Statistics list and click the arrow to add it to the Display list.

▶ Select Column Responses % in the Statistics list and click the arrow to add it to the Display list.

▶ Click Apply to Selection, and then click OK to create the table.

Figure 11-4
Multiple dichotomy responses and column response percentages

		Count	Column N %	Responses	Column Responses %
News sources	Get news from internet	867	41.7%	867	24.1%
	Get news from radio	551	26.5%	551	15.3%
	Get news from television	1077	51.8%	1077	30.0%
	Get news from news magazines	294	14.1%	294	8.2%
	Get news from newspapers	805	38.7%	805	22.4%
	Total	2081	100.0%	3594	100.0%

For each "category" in the multiple dichotomy set, *Responses* is identical to *Count*—and this will always be the case for multiple dichotomy sets. The totals, however, are very different. The total number of responses is 3,594—over 1,500 more than the total count and over 700 more than the total number of cases in the data file.

For percentages, the totals for *Column N %* and *Column Responses %* are both 100%—but the percentages for each category in the multiple dichotomy set are much lower for column response percentages. This is because the percentage base for column response percentages is the total

number of responses, which in this case is 3,594, resulting in much lower percentages than the column percentage base of 2,081.

Percentage Totals Greater Than 100%

Both column percentages and column response percentages yield total percentages of 100% even though, in our example, the individual values in the *Column N %* column clearly sum to greater than 100%. So, what if you want to show percentages based on total count rather than total responses but also want the "total" percentage to accurately reflect the sum of the individual category percentages?

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click *News sources* in the table preview on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select Column Responses % (Base: Count) in the Statistics list and click the arrow to add it to the Display list.

▶ Click Apply to Selection, and then click OK to create the table.

Figure 11-5
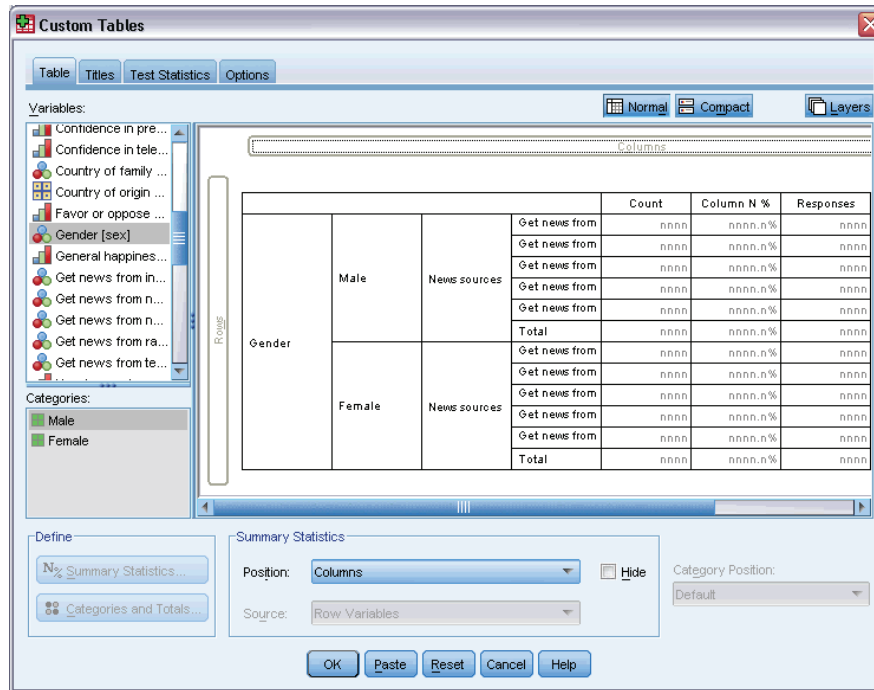*Column response percentages with count as the percentage base*

		Count	Column N %	Responses	Column Responses %	Column Responses % (Base: Count)
News sources	Get news from internet	867	41.7%	867	24.1%	41.7%
	Get news from radio	551	26.5%	551	15.3%	26.5%
	Get news from television	1077	51.8%	1077	30.0%	51.8%
	Get news from news magazines	294	14.1%	294	8.2%	14.1%
	Get news from newspapers	805	38.7%	805	22.4%	38.7%
	Total	2081	100.0%	3594	100.0%	172.7%

Using Multiple Response Sets with Other Variables

In general, you can use multiple response sets just like categorical variables. For example, you can crosstabulate a multiple response set with a categorical variable or nest a multiple response set within a categorical variable.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Drag and drop *Gender* from the variable list to the left side of the Rows area on the preview pane, nesting the multiple response set *News sources* within gender categories.

Figure 11-6
Table preview of nested multiple response set

▶ Right-click *Gender* in the table preview on the canvas pane and deselect (uncheck) Show Variable Label on the pop-up context menu.

▶ Do the same for *News sources*.

This will remove the columns with the variable labels from the table (since they aren't really necessary in this case).

▶ Click OK to create the table.

Figure 11-7
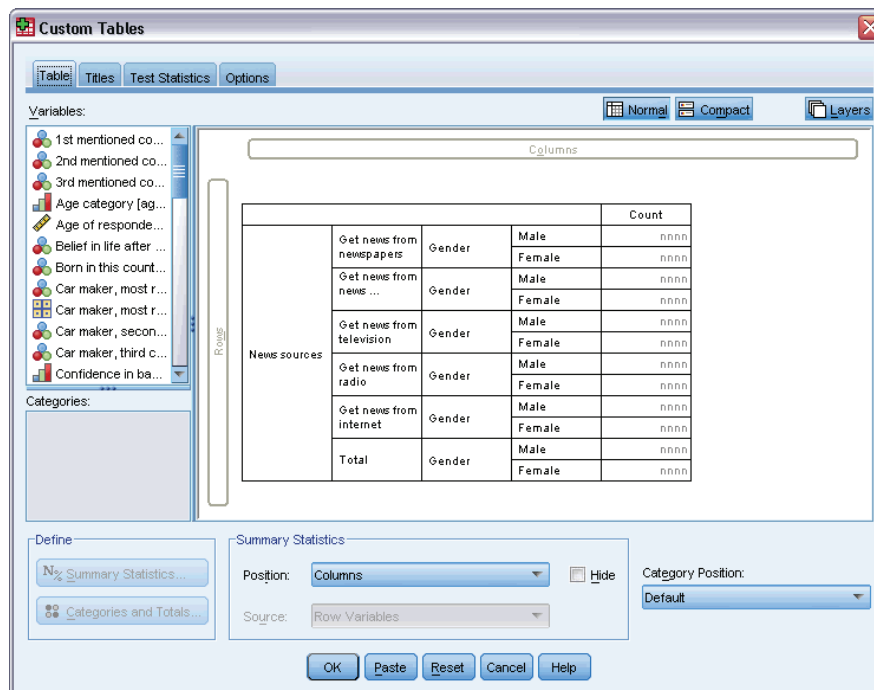*Multiple response set nested within a categorical variable*

		Count	Column N %	Responses	Column Responses %	Column Responses % (Base: Count)
Male	Get news from internet	359	40.1%	359	23.3%	40.1%
	Get news from radio	233	26.0%	233	15.1%	26.0%
	Get news from television	451	50.3%	451	29.3%	50.3%
	Get news from news magazines	121	13.5%	121	7.9%	13.5%
	Get news from newspapers	375	41.9%	375	24.4%	41.9%
	Total	896	100.0%	1539	100.0%	171.8%
Female	Get news from internet	508	42.9%	508	24.7%	42.9%
	Get news from radio	318	26.8%	318	15.5%	26.8%
	Get news from television	626	52.8%	626	30.5%	52.8%
	Get news from news magazines	173	14.6%	173	8.4%	14.6%
	Get news from newspapers	430	36.3%	430	20.9%	36.3%
	Total	1185	100.0%	2055	100.0%	173.4%

Statistics Source Variable and Available Summary Statistics

In the absence of a scale variable in a table, categorical variables and multiple response sets are treated the same way regarding the statistics source variable: The innermost nested variable in the statistics source dimension is the statistics source variable. Since there are some summary statistics that can be assigned only to multiple response sets, this means that the multiple response set must be the innermost nested variable in the statistics source dimension if you want any of the special multiple response summary statistics.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ In the table preview on the canvas pane, drag and drop *News sources* to the left of *Gender*, changing the nesting order.

Figure 11-8
Categorical variable nested within multiple response set

All of the special multiple response summary statistics—responses, column response percentages—are removed from the table preview because the categorical variable *Gender* is now the innermost nested variable and therefore the statistics source variable.

Luckily, the table builder "remembers" these settings. If you move *News sources* back to its previous position, nested within *Gender*, all of the response-related summary statistics are restored to the table preview.

Multiple Category Sets and Duplicate Responses

Multiple category sets provide one feature not available for multiple dichotomy sets: the ability to count duplicate responses. In many cases, duplicate responses in multiple category sets probably represent coding errors. For example, for a survey question such as "What three countries do you think make the best cars?" a response of *Sweden, Germany, and Sweden* probably isn't valid.

In other cases, however, duplicate responses may be perfectly valid. For example, if the question were "Where were your last three cars made?" a response of *Sweden, Germany, and Sweden* makes perfect sense.

Custom Tables provides a choice for duplicate responses in multiple category sets. By default, duplicate responses are not counted, but you can request that they be included.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Reset to clear any previous settings.

▶ Drag and drop *Car maker, most recent cars* from the variable list into the Rows area of the canvas pane.

The icon next to the "variable" in the variable list identifies it as a multiple category set.

Figure 11-9
Multiple category set icon

Figure 11-10
Multiple category set in table builder preview
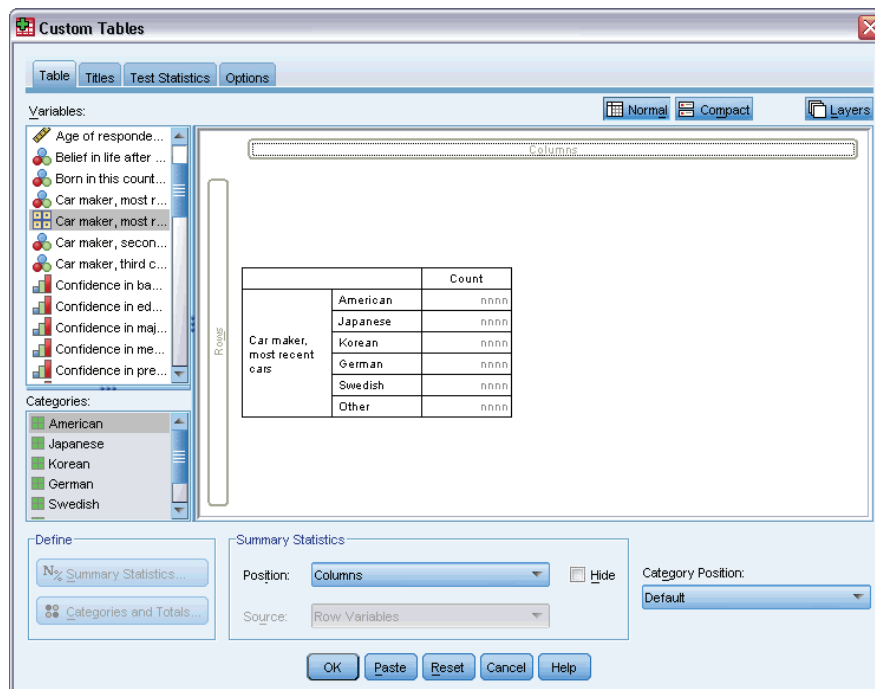
For multiple category sets, the categories displayed represent the common set of defined value labels for all of the variables in the set (whereas for multiple dichotomy sets, the "categories" are actually the variable labels for each variable in the set).

▶ Right-click *Car maker, most recent cars* in the table preview on the canvas pane and select Categories and Totals from the pop-up context menu.

▶ Select (click) Total in the Categories and Totals dialog box, and then click Apply.

▶ Right-click *Car maker, most recent cars* again and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select Responses in the Statistics list and click the arrow to add it to the Display list.

▶ Click Apply to Selection, and then click OK to create the table.

Figure 11-11
Multiple category set: Counts and responses without duplicates

		Count	Responses
Car maker, most recent cars	American	1938	1938
	Japanese	1327	1327
	Korean	695	695
	German	693	693
	Swedish	360	360
	Other	343	343
	Total	2832	5356

By default, duplicate responses are not counted; so in this table, the values for each category in the *Count* and *Responses* columns are identical. Only the totals differ.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click the Options tab.

▶ Click (check) Count duplicate responses for multiple category sets.

▶ Click OK to create the table.

Figure 11-12
Multiple category set with duplicate responses included

		Count	Responses
Car maker, most recent cars	American	1938	2797
	Japanese	1327	1717
	Korean	695	760
	German	693	754
	Swedish	360	383
	Other	343	359
	Total	2832	6770

In this table, there is quite a noticeable difference between the values in the *Count* and *Responses* columns, particularly for American cars, indicating that many respondents have owned multiple American cars.

Significance Testing with Multiple Response Sets

You can use multiple response sets in significance tests in essentially the same way you would use categorical variables.
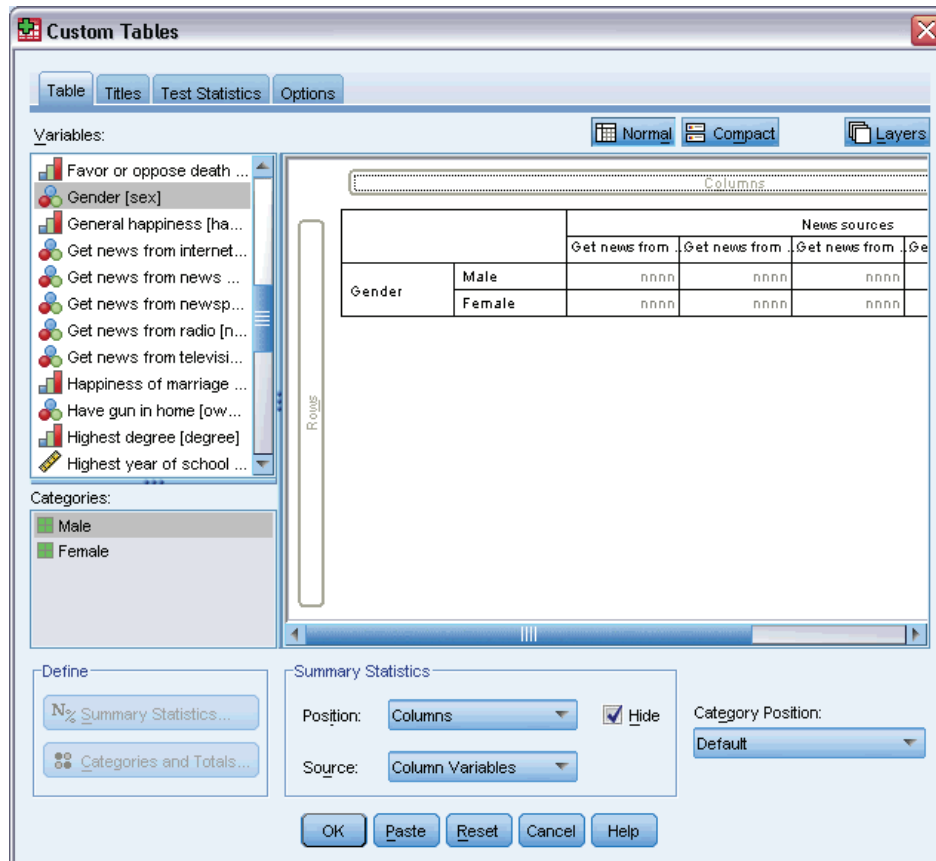
■ For tests of independence (chi-square) or comparing column proportions (z-tests), tests are performed on counts, and Count must be one of the summary statistics displayed in the table.

■ For multiple category sets, tests comparing column proportions or column means (t-tests) are not performed if Count duplicate responses for multiple category sets is selected on the Options tab.

Tests of Independence with Multiple Response Sets

This example creates a crosstabulation of a categorical variable and a multiple response set and performs a chi-square test of independence on the crosstabulation.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Reset to clear any previous settings.

▶ Drag and drop *News sources* (this is the descriptive label for the multiple dichotomy set *$mltnews*) from the variable list into the Columns area of the canvas pane.

▶ Drag and drop *Gender* from the variable list into the Rows area of the canvas pane.

Figure 11-13
News sources (multiple response set) and Gender selected as column and row variables

▶ Click the Test Statistics tab.

▶ Select (check) Tests of independence (chi-square).

▶ If it is not already selected, select Include multiple response variables in test.

Figure 11-14
Test Statistics tab
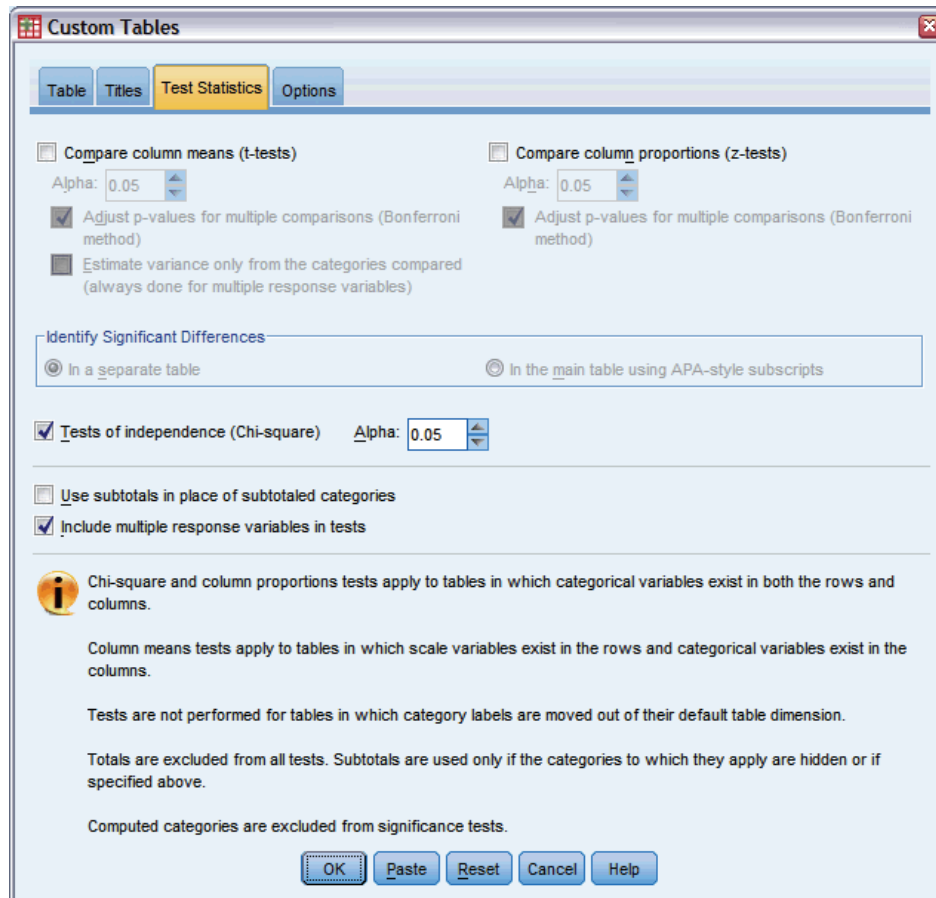
▶ Click OK to run the procedure.

Figure 11-15
Chi-square results

		News sources
Gender	Chi-square	10.266
	df	5
	Sig.	.068

The significance level of 0.068 for the chi-square test indicates that males and females probably do not differ significantly in their choices of news sources (assuming you use a significance value of 0.05 or lower as your criterion for determining statistical significance).

Comparing Column Means with Multiple Response Sets

This example calculates means of a scale variable within categories defined by a multiple response set and compares each category mean to every other category mean for significant differences.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Reset to clear any previous settings.

▶ Drag and drop *News sources* (this is the descriptive label for the multiple dichotomy set *$mltnews*) from the variable list into the Columns area of the canvas pane.

▶ Drag and drop *Age of respondent* into the Rows area of the canvas pane.

Figure 11-16
Comparing mean age across multiple response categories
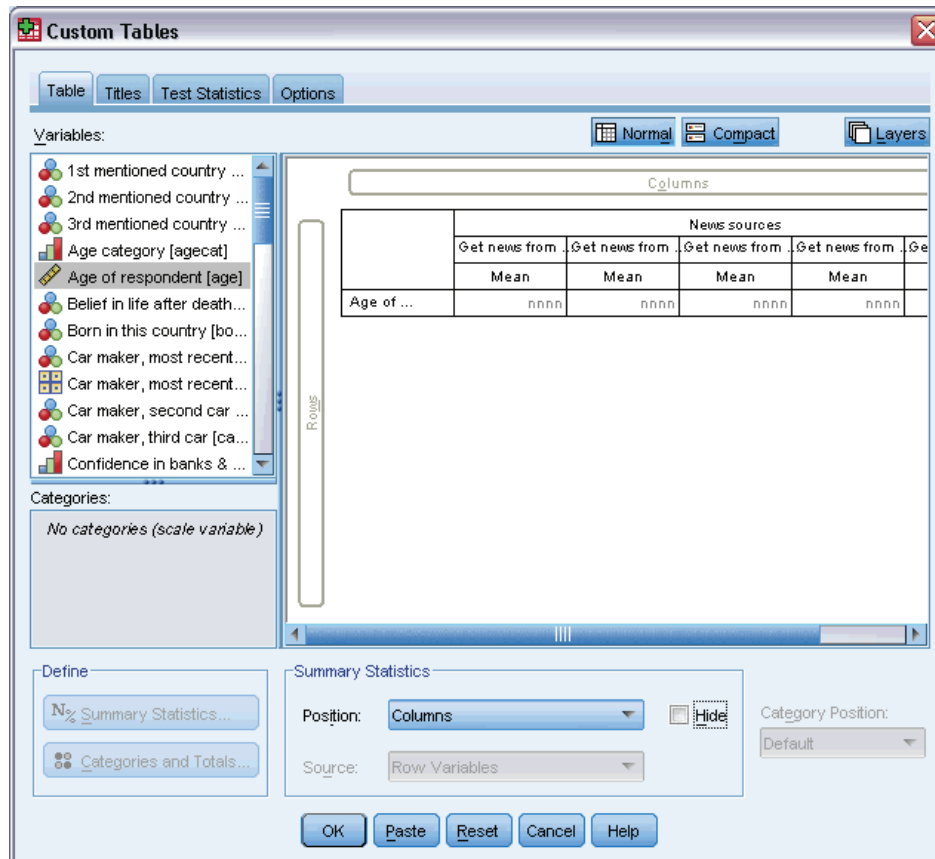
▶ Click the Test Statistics tab.

▶ Select (check) Compare Column Means (t-tests).

▶ If it is not already selected, select Include multiple response variables in test.

▶ Click OK to run the procedure.

Figure 11-17
Significance test results

	News sources				
	Get news from newspapers	Get news from news magazines	Get news from television	Get news from radio	Get news from internet
	Mean	Mean	Mean	Mean	Mean
Age of respondent	52	40	48	40	40

Comparisons of Column Means

	News sources				
	Get news from newspapers	Get news from news magazines	Get news from television	Get news from radio	Get news from internet
	(A)	(B)	(C)	(D)	(E)
Age of respondent	B C D E		B D E		

Results are based on two-sided tests assuming equal variances with significance level 0.05. For each significant pair, the key of the smaller category appears under the category with larger mean.

- Each category of the multiple response set is identified by a letter (A, B, C, D, E), and for each category for which the mean of another category is both lower and differs significantly from the mean of that category, the letter representing the category with the lower mean is displayed.

- *Get news from newspapers* (A) has the highest mean age , and all other category means differ significantly from it.

- *Get news from television* (C) has the next highest mean age, and all remaining category means (B, D, and E) differ significantly from it. (C also differs significantly from A, as previously indicated.)

- The mean ages for *Get news from magazine* (B), *Get news from radio* (D), and *Get news from internet* (E) do not differ significantly from each other.

Missing Values

Many data files contain a certain amount of missing data. A wide variety of factors can result in missing data. For example, survey respondents may not answer every question, certain variables may not be applicable to some cases, and coding errors may result in some values being thrown out.

There are two kinds of missing values in IBM® SPSS® Statistics:

- **User-missing.** Values defined as containing missing data. Value labels can be assigned to these values to identify why the data are missing (such as a code of 99 and a value label of *Not Applicable* for pregnancy in males).

- **System-missing.** If no value is present for a numeric variable, it is assigned the system-missing value. This is indicated by a period in the Data View of the Data Editor.

There are a number of facilities that can help to compensate for the effects of missing data and even analyze patterns in missing data. This chapter, however, has a much simpler goal: to describe how Custom Tables handles missing data and how missing data affect the computation of summary statistics.

Sample Data File

The examples in this chapter use the data file *missing_values.sav*. For more information, see the topic Sample Files in Appendix A on p. 178. This is a very simple, completely artificial data file, with only one variable and ten cases, designed to illustrate basic concepts about missing values.

Tables without Missing Values

By default, user-missing categories are not displayed in custom tables (and system-missing values are never displayed).

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ In the table builder, drag and drop *Variable with missing values* (the only variable in the file) from the variable list into the Rows area of the canvas pane.

▶ Right-click the variable on the canvas pane and select Categories and Totals from the pop-up context menu.

▶ Click (check) Total in the Categories and Totals dialog box, and then click Apply.

▶ Right-click *Variable with missing values* in the table preview on the canvas pane again and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, select Column N % in the Statistics list and click the arrow to add it to the Display list.

▶ Click Apply to Selection.

Figure 12-1
Table preview without missing values
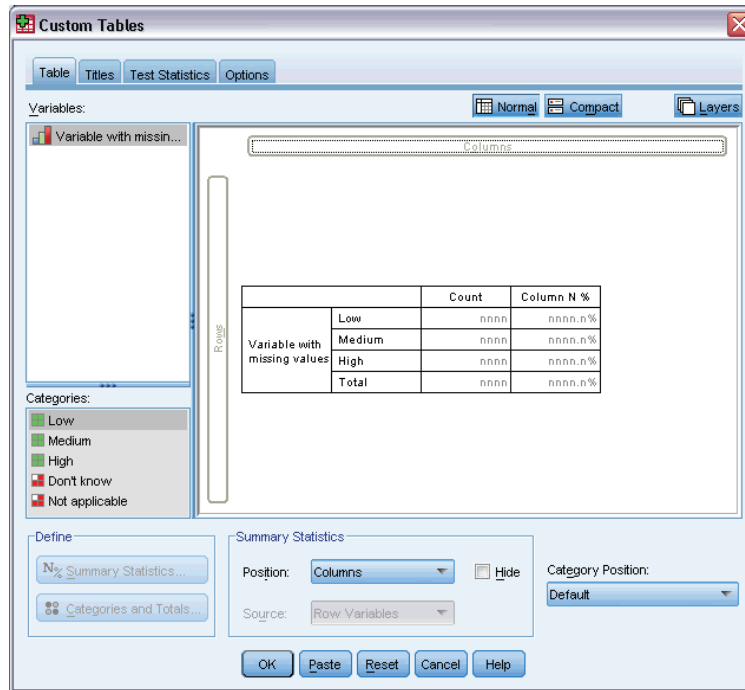
You may notice a slight discrepancy between the categories displayed in the table preview on the canvas pane and the categories displayed in the Categories list (below the variable list on the left side of the table builder). The Categories list contains a category labeled *Missing Values* that isn't included in the table preview because missing value categories are excluded by default. Since "values" is plural in the label, this indicates that the variable has two or more user-missing categories.

▶ Click OK to create the table.

Figure 12-2
Table without missing values

		Count	Column N %
Variable with missing values	Low	2	28.6%
	Medium	3	42.9%
	High	2	28.6%
	Total	7	100.0%

Everything in this table is perfectly fine. The category values add up to the totals, and the percentages accurately reflect the values you'd get using the total count as the percentage base (for example, 3/7= 0.429, or 42.9%). The total count, however, is not the total number of cases in the data file; it's the total number of cases with **non-missing** values, or cases that don't have user-missing or system-missing values for that variable.

Including Missing Values in Tables

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click *Variable with missing values* in the table preview on the canvas pane and select Categories and Totals from the pop-up context menu.

▶ Click (check) Missing Values in the Categories and Totals dialog box, and then click Apply.

Figure 12-3
Table preview with missing values category displayed
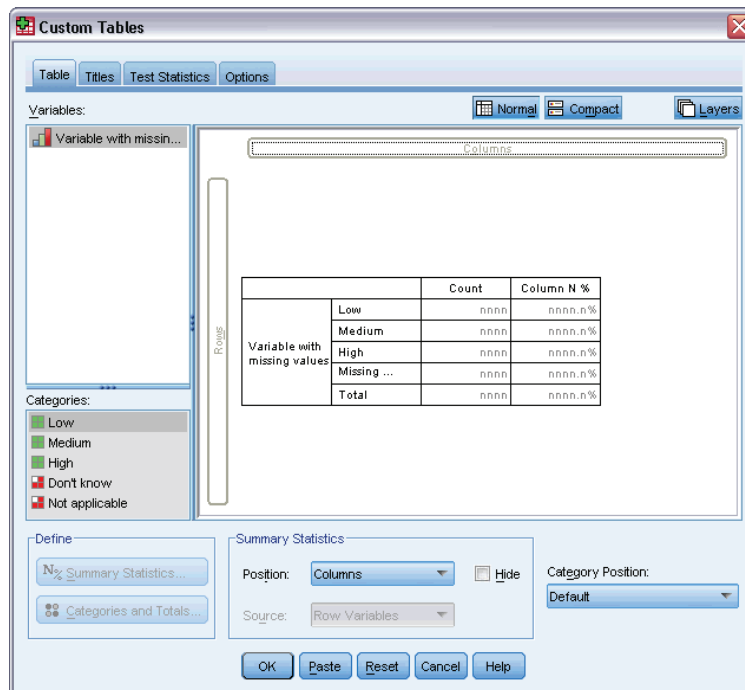
Now the table preview includes a *Missing Values* category. Although the table preview displays only one category for missing values, all user-missing categories will be displayed in the table.

▶ Right-click *Variable with missing values* in the table preview on the canvas pane again and select Summary Statistics from the pop-up context menu.

▶ In the Summary Statistics dialog box, click (check) Custom Summary Statistics for Totals and Subtotals.

▶ Select Valid N in the custom summary Statistics list and click the arrow to add it to the Display list.

▶ Do the same for Total N.

▶ Click Apply to Selection, and then click OK in the table builder to create the table.

Figure 12-4
Table with missing values

		Count	Column N %	Valid N	Total N
Variable with missing values	Low	2	22.2%		
	Medium	3	33.3%		
	High	2	22.2%		
	Don't know	1	11.1%		
	Not applicable	1	11.1%		
	Total	9	100.0%	7	10

The two defined user-missing categories—*Don't know* and *Not applicable*—are now displayed in the table, and the total count is now 9 instead of 7, reflecting the addition of the two cases with user-missing values (one in each user-missing category). The column percentages are also different now, because they are based on the number of non-missing and user-missing values. Only system-missing values are not included in the percentage calculation.

Valid N shows the total number of non-missing cases (7), and *Total N* shows the total number of cases, including both user-missing and system-missing. The total number of cases is 10, one more than the count of non-missing and user-missing values displayed as the total in the *Count* column. This is because there's one case with a system-missing value.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Right-click *Variable with missing values* in the table preview on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ Select Column Valid N % in the top Statistics list (not the custom summaries for totals and subtotals) and click the arrow to add it to the Display list.

▶ Do the same for Column Total N %.

▶ You can also add them both to the list of custom summary statistics for totals and subtotals.

▶ Click Apply to Selection, and then click OK to create the table.

Figure 12-5
Table with missing values and valid and total percentages

		Count	Column N %	Column Valid N %	Column Total N %	Valid N	Total N
Variable with missing values	Low	2	22.2%	28.6%	20.0%		
	Medium	3	33.3%	42.9%	30.0%		
	High	2	22.2%	28.6%	20.0%		
	Don't know	1	11.1%	.0%	10.0%		
	Not applicable	1	11.1%	.0%	10.0%		
	Total	9	100.0%	100.0%	100.0%	7	10

■ *Column N %* is the percentage in each category based on the number of non-missing and user-missing values (since user-missing values have been explicitly included in the table).

■ *Column Valid N %* is the percentage in each category based on only the valid, non-missing cases. These values are the same as the column percentages were in the original table that did not include user-missing values.

■ *Column Total N %* is the percentage in each category based on all cases, including both user-missing and system-missing. If you add up the individual category percentages in this category, you'll see that they add up to only 90%, because one case out of the total of 10 cases (10%) has the system-missing value. Although this case is included in the base for the

percentage calculations, no category is provided in the table for cases with system-missing values.

Formatting and Customizing Tables

Custom Tables provides the ability to control a number of table-formatting properties as part of the table-building process, including:

■ Display format and labels for summary statistics

■ Minimum and maximum data column width

■ Text or value displayed in empty cells

These settings persist within the table builder interface (until you change them, reset the table builder settings, or open a different data file), enabling you to create multiple tables with the same formatting properties without manually editing the tables after creating them. You can also save these formatting settings, along with all of the other table parameters, using the Paste button in the table builder interface to paste command syntax into a syntax window, which you can then save as a file.

You can also change many formatting properties of tables after they have been created, using all of the formatting capabilities available in the Viewer for pivot tables. This chapter, however, focuses on controlling table formatting properties before the table is created. For more information on pivot tables, use the Index tab in the Help system and type pivot tables as the keyword.

Sample Data File

The examples in this chapter use the data file *survey_sample.sav*. For more information, see the topic Sample Files in Appendix A on p. 178.

All examples provided here display variable labels in dialog boxes, sorted in alphabetical order. Variable list display properties are set on the General tab in the Options dialog box (Edit menu, Options).

Summary Statistics Display Format

Custom Tables attempts to apply relatively intelligent default formats to summary statistics, but there will probably be times when you want to override these defaults.

▶ From the menus, choose:
Analyze > Tables > Custom Tables...

▶ In the table builder, drag and drop *Age category* from the variable list into the Rows area on the canvas pane.

▶ Drag and drop *Confidence in television* below *Age category* in the Rows area, stacking the two variables in the row dimension.

▶ Right-click *Age category* in the table preview on the canvas pane and select Select All Row Variables from the pop-up context menu.

▶ Right-click *Age category* again and select Categories and Totals from the pop-up context menu.

▶ In the Categories and Totals dialog box, select (check) Total and then click Apply.

▶ Right-click either variable in the table preview on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ Select Column N % in the Statistics list and click the arrow key to add it to the Display list.

▶ Select (check) Custom Summary Statistics for Totals and Subtotals.

▶ In the Statistics list for custom summary statistics, select Column N % and click the arrow to add it to the Display list.

▶ Do the same for Mean.

▶ Then click Apply to All.

Figure 13-1
Default display formats in table preview
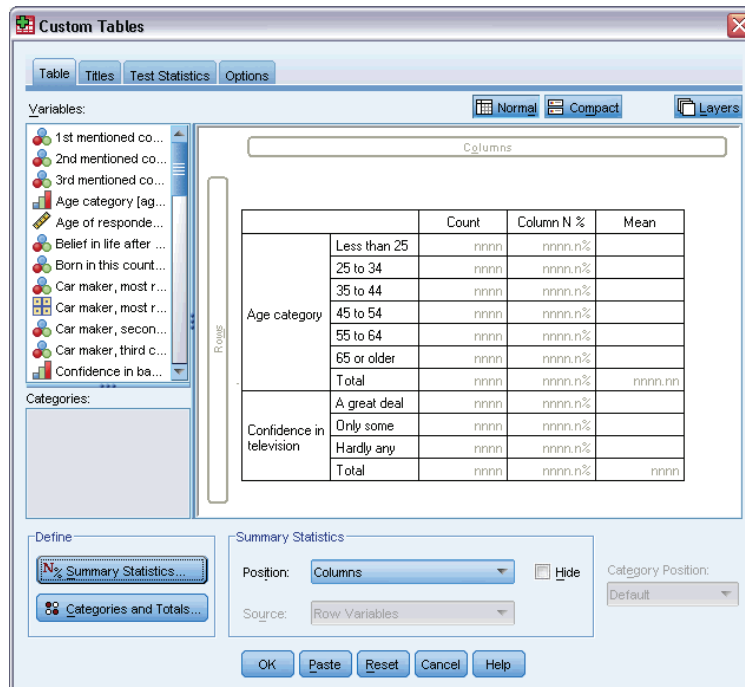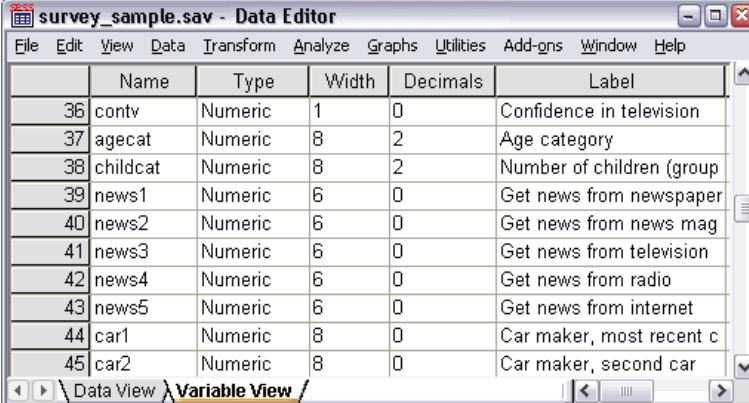
The placeholder values in the table preview reflect the default format for each summary statistic.

■ For counts, the default display format is nnnn—integer values with no decimal places.

■ For percentages, the default display format is nnnn.n%—numbers with a single decimal place and a percentage sign after the value.

■ For the mean, the default display format is *different* for the two variables.

For summary statistics that aren't some form of count (including Valid N and Total N) or percentage, the default display format is the display format defined for the variable in the Data Editor. If you look at the variables in Variable View in the Data Editor, you will see that *Age category* (variable *agecat*) is defined as having two decimal positions, while *Confidence in television* (variable *contv*) is defined as having zero decimal positions.

Figure 13-2
Variable View in the Data Editor

This is one of those cases where the default format probably isn't the format you want, since it would probably be better if both mean values displayed the same number of decimals.

▶ Right-click either variable in the table preview on the canvas pane and select Summary Statistics from the pop-up context menu.

171

Formatting and Customizing Tables

Figure 13-3
Summary Statistics dialog box
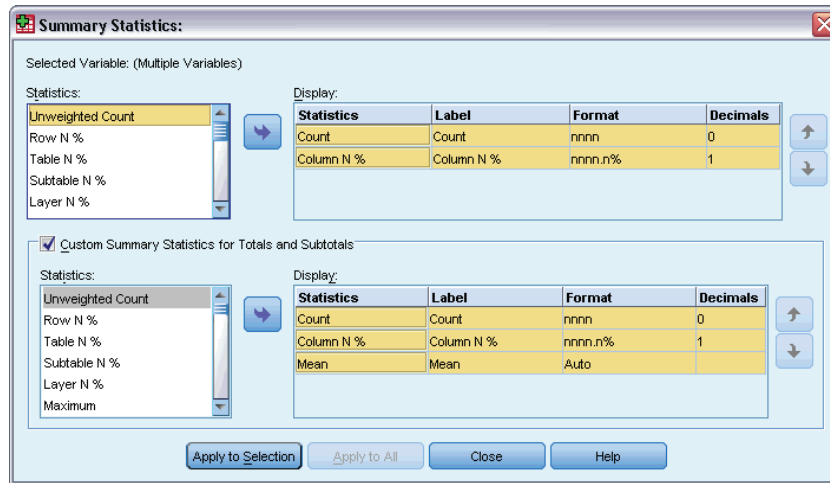
For the mean, the Format cell in the Display list indicates that the format is *Auto*, which means that the defined display format for the variable will be used, and the Decimals cell is disabled. In order to specify the number of decimals, you first need to select a different format.

► In the custom summary statistics Display list, click the Format cell for the mean, and select nnnn from the drop-down list of formats.

► In the Decimals cell, enter a value of 1.

► Then click Apply to All to apply this setting to both variables.

Figure 13-4
Table preview with user-specified summary statistics display formats
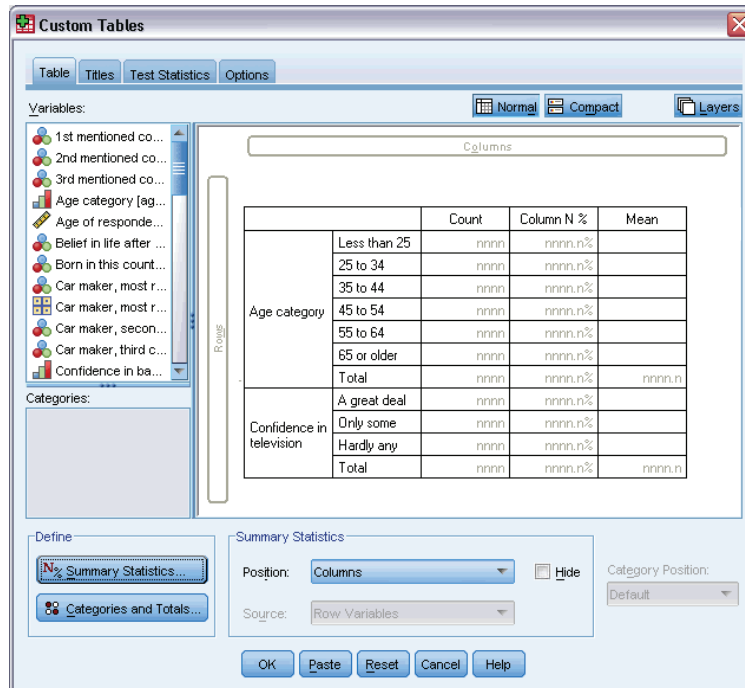
Now the table preview indicates that both mean values will be displayed with one decimal position. (You could go ahead and create this table now—but you might find the "mean" value for *Age category* a little difficult to interpret, since the actual numeric codes for this variable range only from 1 to 6.)

Display Labels for Summary Statistics

In addition to the display formats for summary statistics, you can also control the descriptive labels for each summary statistic.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click Reset to clear any previous settings in the table builder.

▶ In the table builder, drag and drop *Age category* from the variable list into the Rows area on the canvas pane.

▶ Drag and drop *How get paid last week* from the variable list into the Columns area on the canvas pane.

▶ Right-click *Age category* in the table preview on the canvas pane and select Summary Statistics from the pop-up context menu.

▶ Select Column N % in the Statistics list and click the arrow key to add it to the Display list.

▶ Double-click anywhere in the word *Column* in the Label cell in the Display list to edit the contents of the cell. Delete the word *Column* from the label, changing the label to simply *%.*

▶ Edit the Label cell for *Count* in the same way, changing the label to simply *N.*

While we're here, let's change the format of the Column N % statistic to remove the unnecessary percentage sign (since the column label indicates that the column contains percentages).

▶ Click the Format cell for *Column N %* and select nnnn.n from the drop-down list of formats.

Figure 13-5
Summary Statistics dialog box with modified labels and formats
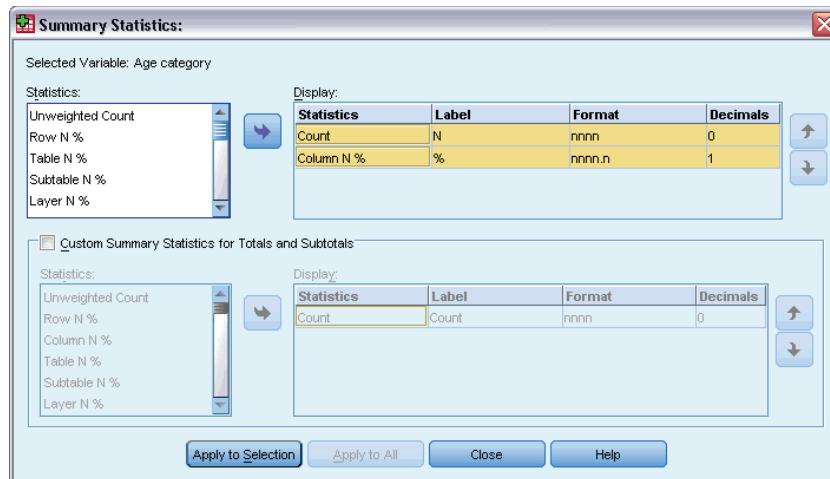
▶ Then click Apply to Selection.

Figure 13-6
Table preview with modified summary statistics labels
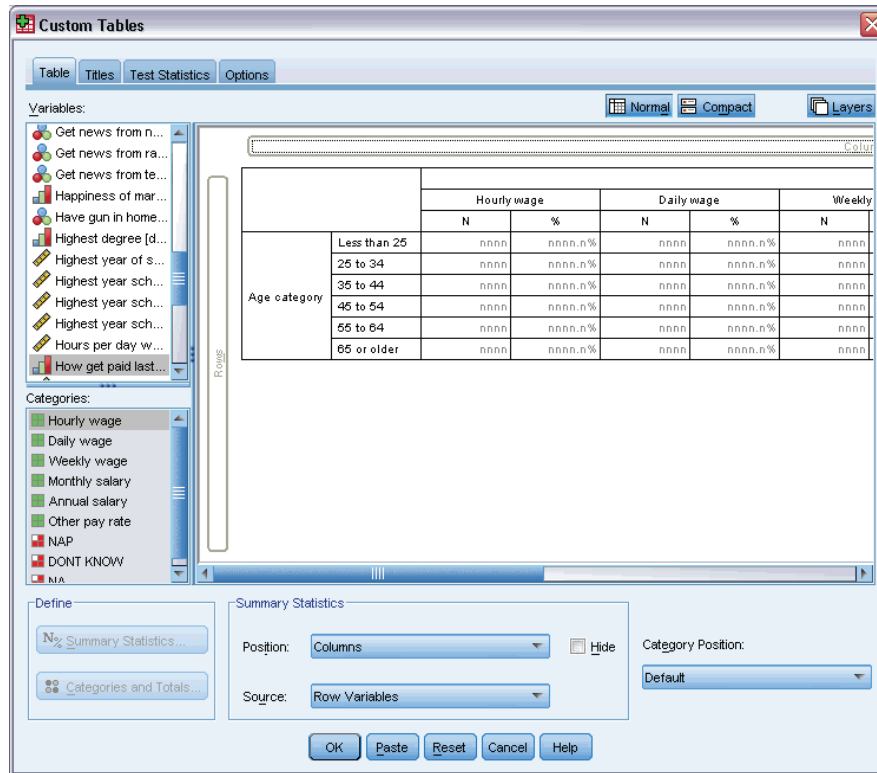
The table preview displays the modified display format and the modified labels.

▶ Click OK to create the table.

Figure 13-7
Table with modified summary statistics labels

		How get paid last week										
		Hourly wage		Daily wage		Weekly wage		Monthly salary		Annual salary		Other pay rate	
		N	%	N	%	N	%	N	%	N	%	N	%
Age category	Less than 25	91	14.0	0	.0	12	9.7	3	2.0	7	3.1	14	7.7
	25 to 34	175	26.9	5	29.4	33	26.6	37	24.8	63	28.0	31	17.1
	35 to 44	185	28.5	5	29.4	42	33.9	45	30.2	66	29.3	61	33.7
	45 to 54	124	19.1	5	29.4	25	20.2	38	25.5	58	25.8	41	22.7
	55 to 64	52	8.0	0	.0	10	8.1	23	15.4	29	12.9	19	10.5
	65 or older	23	3.5	2	11.8	2	1.6	3	2.0	2	.9	15	8.3

Column Width

You may have noticed that the table in the above example is rather wide. One solution to this problem would be to simply swap the row and column variables. Another solution is to make the columns narrower, since they seem to be much wider than necessary. (In fact, the reason we shortened the summary statistics labels was so that we could make the columns narrower.)

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click the Options tab.

▶ In the Width for Data Columns group, select Custom.

▶ For the Maximum, type 36. (Make sure that the Units setting is Points.)

Figure 13-8
Custom Tables dialog box, Options tab
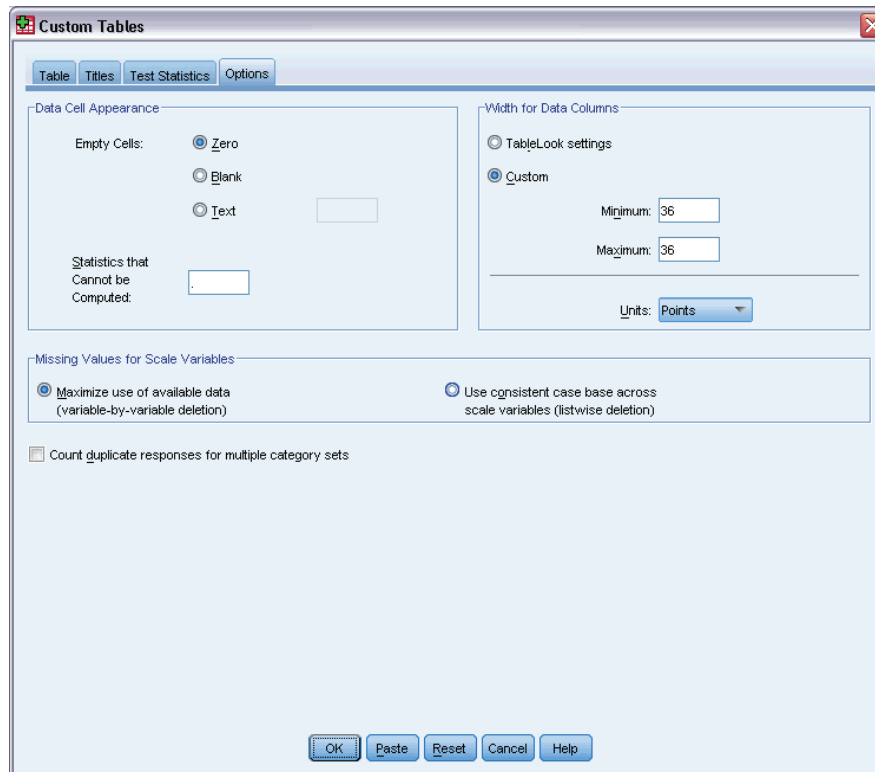
▶ Click OK to create the table.

Figure 13-9
Table with reduced column widths

		How get paid last week											
		Hourly wage		Daily wage		Weekly wage		Monthly salary		Annual salary		Other pay rate	
		N	%	N	%	N	%	N	%	N	%	N	%
Age category	Less than 25	91	14.0	0	.0	12	9.7	3	2.0	7	3.1	14	7.7
	25 to 34	175	26.9	5	29.4	33	26.6	37	24.8	63	28.0	31	17.1
	35 to 44	185	28.5	5	29.4	42	33.9	45	30.2	66	29.3	61	33.7
	45 to 54	124	19.1	5	29.4	25	20.2	38	25.5	58	25.8	41	22.7
	55 to 64	52	8.0	0	.0	10	8.1	23	15.4	29	12.9	19	10.5
	65 or older	23	3.5	2	11.8	2	1.6	3	2.0	2	.9	15	8.3

Now the table is much more compact.

Display Value for Empty Cells

By default, a 0 is displayed in empty cells (cells that contain no cases). You can instead display nothing in these cells (leave them blank) or specify a text string to display in empty cells.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Click the Options tab.

▶ In the Data Cell Appearance group, for Empty Cells select Text and type None.

▶ Click OK to create the table.

Figure 13-10
Table with "None" displayed in empty cells

		How get paid last week											
		Hourly wage		Daily wage		Weekly wage		Monthly salary		Annual salary		Other pay rate	
		N	%	N	%	N	%	N	%	N	%	N	%
Age category	Less than 25	91	14.0	None	None	12	9.7	3	2.0	7	3.1	14	7.7
	25 to 34	175	26.9	5	29.4	33	26.6	37	24.8	63	28.0	31	17.1
	35 to 44	185	28.5	5	29.4	42	33.9	45	30.2	66	29.3	61	33.7
	45 to 54	124	19.1	5	29.4	25	20.2	38	25.5	58	25.8	41	22.7
	55 to 64	52	8.0	None	None	10	8.1	23	15.4	29	12.9	19	10.5
	65 or older	23	3.5	2	11.8	2	1.6	3	2.0	2	.9	15	8.3

Now the four empty cells in the table display the text *None* instead of a value of 0.

Display Value for Missing Statistics

If a statistic cannot be computed, the default display value is a period (.), which is the symbol used to indicate the system-missing value. This is different from an "empty" cell, and therefore the display value for missing statistics is controlled separately from the display value for cells that contain no cases.

▶ Open the table builder (Analyze menu, Tables, Custom Tables).

▶ Drag and drop *Hours per day watching TV* from the variable list to the top of the Columns area on the canvas, above *How get paid last week*.

Since *Hours per day watching TV* is a scale variable, it automatically becomes the statistics source variable and the summary statistic changes to the mean.

▶ Right-click *Hours per day watching TV* in the table preview in the canvas pane and select Summary Statistics from the pop-up context menu.

▶ Select Valid N in the Statistics list and click the arrow key to add it to the Display list.
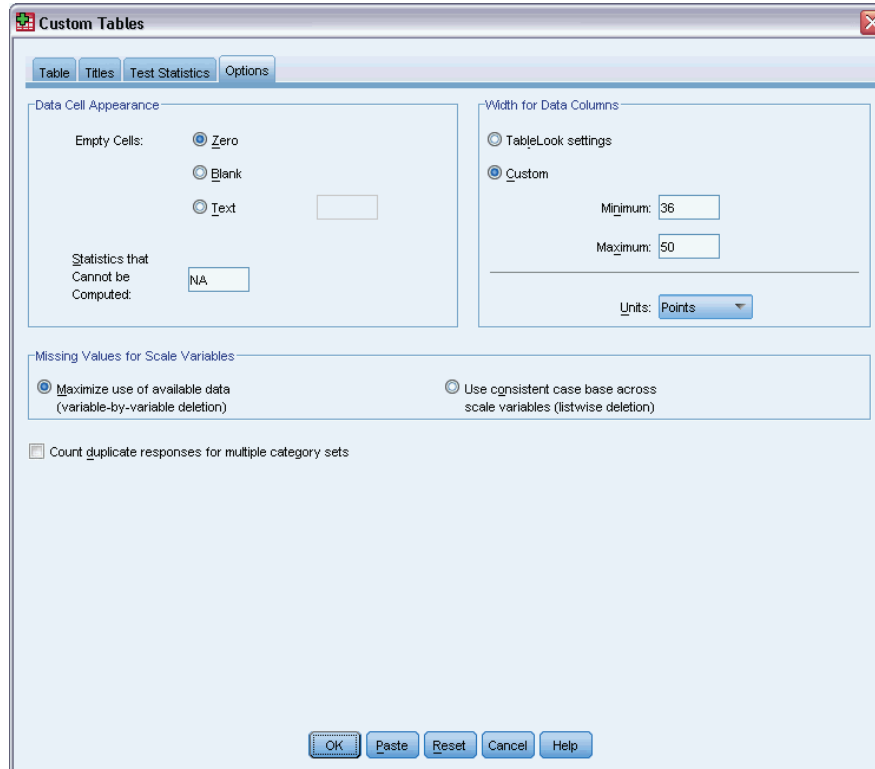
Figure 13-11
Summary Statistics dialog box for scale variables
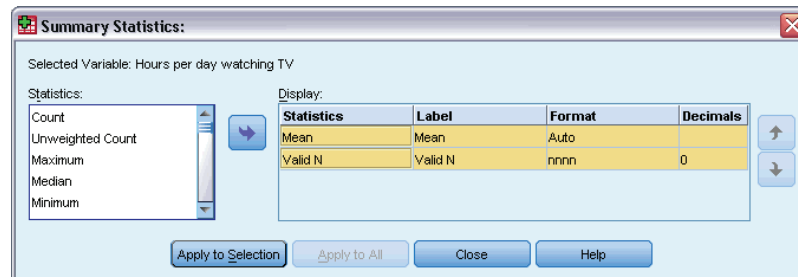
▶ Click Apply to Selection.

▶ Click the Options tab.

▶ In the text field for Statistics that Cannot be Computed, type **NA**.

Figure 13-12
Changing the display value for statistics that cannot be computed

▶ Click OK to create the table.

Figure 13-13
Table with "NA" displayed for missing statistics

		Hours per day watching TV											
		How get paid last week											
		Hourly wage		Daily wage		Weekly wage		Monthly salary		Annual salary		Other pay rate	
		Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N	Mean	Valid N
Age category	Less than 25	3	71	NA	None	3	10	2	3	2	6	2	8
	25 to 34	3	134	5	2	2	30	2	29	2	52	2	22
	35 to 44	3	136	2	5	3	30	2	34	2	47	3	46
	45 to 54	2	90	2	4	2	22	2	36	2	45	2	34
	55 to 64	3	40	NA	None	3	7	2	15	2	23	3	15
	65 or older	3	18	2	2	1	1	NA	0	1	2	3	11

The text *NA* is displayed for the mean in three cells in the table. In each case, the corresponding *Valid N* value explains why: There are no cases with which to compute the mean.

You may, however, notice what appears to be a slight discrepancy—one of those three Valid N values is displayed as a 0, rather than the label *None* that is supposed to be displayed in cells with no cases. This is because although there are no valid cases to use to compute the mean, the category isn't really empty. If you go back to the original table with just the two categorical variables, you will see that there are, in fact, three cases in this crosstabulated category. There are no valid cases, however, because all three have missing values for the scale variable *Hours per day watching TV*.

Sample Files

The sample files installed with the product can be found in the *Samples* subdirectory of the installation directory. There is a separate folder within the Samples subdirectory for each of the following languages: English, French, German, Italian, Japanese, Korean, Polish, Russian, Simplified Chinese, Spanish, and Traditional Chinese.

Not all sample files are available in all languages. If a sample file is not available in a language, that language folder contains an English version of the sample file.

Descriptions

Following are brief descriptions of the sample files used in various examples throughout the documentation.

■ **accidents.sav.** This is a hypothetical data file that concerns an insurance company that is studying age and gender risk factors for automobile accidents in a given region. Each case corresponds to a cross-classification of age category and gender.

■ **adl.sav.** This is a hypothetical data file that concerns efforts to determine the benefits of a proposed type of therapy for stroke patients. Physicians randomly assigned female stroke patients to one of two groups. The first received the standard physical therapy, and the second received an additional emotional therapy. Three months following the treatments, each patient's abilities to perform common activities of daily life were scored as ordinal variables.

■ **advert.sav.** This is a hypothetical data file that concerns a retailer's efforts to examine the relationship between money spent on advertising and the resulting sales. To this end, they have collected past sales figures and the associated advertising costs..

■ **aflatoxin.sav.** This is a hypothetical data file that concerns the testing of corn crops for aflatoxin, a poison whose concentration varies widely between and within crop yields. A grain processor has received 16 samples from each of 8 crop yields and measured the alfatoxin levels in parts per billion (PPB).

■ **anorectic.sav.** While working toward a standardized symptomatology of anorectic/bulimic behavior, researchers made a study of 55 adolescents with known eating disorders. Each patient was seen four times over four years, for a total of 220 observations. At each observation, the patients were scored for each of 16 symptoms. Symptom scores are missing for patient 71 at time 2, patient 76 at time 2, and patient 47 at time 3, leaving 217 valid observations.

■ **bankloan.sav.** This is a hypothetical data file that concerns a bank's efforts to reduce the rate of loan defaults. The file contains financial and demographic information on 850 past and prospective customers. The first 700 cases are customers who were previously given

loans. The last 150 cases are prospective customers that the bank needs to classify as good or bad credit risks.

- **bankloan_binning.sav.** This is a hypothetical data file containing financial and demographic information on 5,000 past customers.

- **behavior.sav.** In a classic example , 52 students were asked to rate the combinations of 15 situations and 15 behaviors on a 10-point scale ranging from 0="extremely appropriate" to 9="extremely inappropriate." Averaged over individuals, the values are taken as dissimilarities.

- **behavior_ini.sav.** This data file contains an initial configuration for a two-dimensional solution for *behavior.sav*.

- **brakes.sav.** This is a hypothetical data file that concerns quality control at a factory that produces disc brakes for high-performance automobiles. The data file contains diameter measurements of 16 discs from each of 8 production machines. The target diameter for the brakes is 322 millimeters.

- **breakfast.sav.** In a classic study , 21 Wharton School MBA students and their spouses were asked to rank 15 breakfast items in order of preference with 1="most preferred" to 15="least preferred." Their preferences were recorded under six different scenarios, from "Overall preference" to "Snack, with beverage only."

- **breakfast-overall.sav.** This data file contains the breakfast item preferences for the first scenario, "Overall preference," only.

- **broadband_1.sav.** This is a hypothetical data file containing the number of subscribers, by region, to a national broadband service. The data file contains monthly subscriber numbers for 85 regions over a four-year period.

- **broadband_2.sav.** This data file is identical to *broadband_1.sav* but contains data for three additional months.

- **car_insurance_claims.sav.** A dataset presented and analyzed elsewhere concerns damage claims for cars. The average claim amount can be modeled as having a gamma distribution, using an inverse link function to relate the mean of the dependent variable to a linear combination of the policyholder age, vehicle type, and vehicle age. The number of claims filed can be used as a scaling weight.

- **car_sales.sav.** This data file contains hypothetical sales estimates, list prices, and physical specifications for various makes and models of vehicles. The list prices and physical specifications were obtained alternately from *edmunds.com* and manufacturer sites.

- **car_sales_uprepared.sav.** This is a modified version of *car_sales.sav* that does not include any transformed versions of the fields.

- **carpet.sav.** In a popular example , a company interested in marketing a new carpet cleaner wants to examine the influence of five factors on consumer preference—package design, brand name, price, a *Good Housekeeping* seal, and a money-back guarantee. There are three factor levels for package design, each one differing in the location of the applicator brush; three brand names (*K2R*, *Glory*, and *Bissell*); three price levels; and two levels (either no or yes) for each of the last two factors. Ten consumers rank 22 profiles defined by these factors. The variable *Preference* contains the rank of the average rankings for each profile. Low rankings correspond to high preference. This variable reflects an overall measure of preference for each profile.

- **carpet_prefs.sav.** This data file is based on the same example as described for *carpet.sav*, but it contains the actual rankings collected from each of the 10 consumers. The consumers were asked to rank the 22 product profiles from the most to the least preferred. The variables *PREF1* through *PREF22* contain the identifiers of the associated profiles, as defined in *carpet_plan.sav*.

- **catalog.sav.** This data file contains hypothetical monthly sales figures for three products sold by a catalog company. Data for five possible predictor variables are also included.

- **catalog_seasfac.sav.** This data file is the same as *catalog.sav* except for the addition of a set of seasonal factors calculated from the Seasonal Decomposition procedure along with the accompanying date variables.

- **cellular.sav.** This is a hypothetical data file that concerns a cellular phone company's efforts to reduce churn. Churn propensity scores are applied to accounts, ranging from 0 to 100. Accounts scoring 50 or above may be looking to change providers.

- **ceramics.sav.** This is a hypothetical data file that concerns a manufacturer's efforts to determine whether a new premium alloy has a greater heat resistance than a standard alloy. Each case represents a separate test of one of the alloys; the heat at which the bearing failed is recorded.

- **cereal.sav.** This is a hypothetical data file that concerns a poll of 880 people about their breakfast preferences, also noting their age, gender, marital status, and whether or not they have an active lifestyle (based on whether they exercise at least twice a week). Each case represents a separate respondent.

- **clothing_defects.sav.** This is a hypothetical data file that concerns the quality control process at a clothing factory. From each lot produced at the factory, the inspectors take a sample of clothes and count the number of clothes that are unacceptable.

- **coffee.sav.** This data file pertains to perceived images of six iced-coffee brands . For each of 23 iced-coffee image attributes, people selected all brands that were described by the attribute. The six brands are denoted AA, BB, CC, DD, EE, and FF to preserve confidentiality.

- **contacts.sav.** This is a hypothetical data file that concerns the contact lists for a group of corporate computer sales representatives. Each contact is categorized by the department of the company in which they work and their company ranks. Also recorded are the amount of the last sale made, the time since the last sale, and the size of the contact's company.

- **creditpromo.sav.** This is a hypothetical data file that concerns a department store's efforts to evaluate the effectiveness of a recent credit card promotion. To this end, 500 cardholders were randomly selected. Half received an ad promoting a reduced interest rate on purchases made over the next three months. Half received a standard seasonal ad.

- **customer_dbase.sav.** This is a hypothetical data file that concerns a company's efforts to use the information in its data warehouse to make special offers to customers who are most likely to reply. A subset of the customer base was selected at random and given the special offers, and their responses were recorded.

- **customer_information.sav.** A hypothetical data file containing customer mailing information, such as name and address.

- **customer_subset.sav.** A subset of 80 cases from *customer_dbase.sav*.

- **debate.sav.** This is a hypothetical data file that concerns paired responses to a survey from attendees of a political debate before and after the debate. Each case corresponds to a separate respondent.

- **debate_aggregate.sav.** This is a hypothetical data file that aggregates the responses in *debate.sav*. Each case corresponds to a cross-classification of preference before and after the debate.

- **demo.sav.** This is a hypothetical data file that concerns a purchased customer database, for the purpose of mailing monthly offers. Whether or not the customer responded to the offer is recorded, along with various demographic information.

- **demo_cs_1.sav.** This is a hypothetical data file that concerns the first step of a company's efforts to compile a database of survey information. Each case corresponds to a different city, and the region, province, district, and city identification are recorded.

- **demo_cs_2.sav.** This is a hypothetical data file that concerns the second step of a company's efforts to compile a database of survey information. Each case corresponds to a different household unit from cities selected in the first step, and the region, province, district, city, subdivision, and unit identification are recorded. The sampling information from the first two stages of the design is also included.

- **demo_cs.sav.** This is a hypothetical data file that contains survey information collected using a complex sampling design. Each case corresponds to a different household unit, and various demographic and sampling information is recorded.

- **dmdata.sav.** This is a hypothetical data file that contains demographic and purchasing information for a direct marketing company. *dmdata2.sav* contains information for a subset of contacts that received a test mailing, and *dmdata3.sav* contains information on the remaining contacts who did not receive the test mailing.

- **dietstudy.sav.** This hypothetical data file contains the results of a study of the "Stillman diet" . Each case corresponds to a separate subject and records his or her pre- and post-diet weights in pounds and triglyceride levels in mg/100 ml.

- **dvdplayer.sav.** This is a hypothetical data file that concerns the development of a new DVD player. Using a prototype, the marketing team has collected focus group data. Each case corresponds to a separate surveyed user and records some demographic information about them and their responses to questions about the prototype.

- **german_credit.sav.** This data file is taken from the "German credit" dataset in the Repository of Machine Learning Databases at the University of California, Irvine.

- **grocery_1month.sav.** This hypothetical data file is the *grocery_coupons.sav* data file with the weekly purchases "rolled-up" so that each case corresponds to a separate customer. Some of the variables that changed weekly disappear as a result, and the amount spent recorded is now the sum of the amounts spent during the four weeks of the study.

- **grocery_coupons.sav.** This is a hypothetical data file that contains survey data collected by a grocery store chain interested in the purchasing habits of their customers. Each customer is followed for four weeks, and each case corresponds to a separate customer-week and records information about where and how the customer shops, including how much was spent on groceries during that week.

- **guttman.sav.** Bell presented a table to illustrate possible social groups. Guttman used a portion of this table, in which five variables describing such things as social interaction, feelings of belonging to a group, physical proximity of members, and formality of the relationship were crossed with seven theoretical social groups, including crowds (for example, people at a football game), audiences (for example, people at a theater or classroom lecture), public (for example, newspaper or television audiences), mobs (like a crowd but with much more intense interaction), primary groups (intimate), secondary groups (voluntary), and the modern community (loose confederation resulting from close physical proximity and a need for specialized services).

- **health_funding.sav.** This is a hypothetical data file that contains data on health care funding (amount per 100 population), disease rates (rate per 10,000 population), and visits to health care providers (rate per 10,000 population). Each case represents a different city.

- **hivassay.sav.** This is a hypothetical data file that concerns the efforts of a pharmaceutical lab to develop a rapid assay for detecting HIV infection. The results of the assay are eight deepening shades of red, with deeper shades indicating greater likelihood of infection. A laboratory trial was conducted on 2,000 blood samples, half of which were infected with HIV and half of which were clean.

- **hourlywagedata.sav.** This is a hypothetical data file that concerns the hourly wages of nurses from office and hospital positions and with varying levels of experience.

- **insurance_claims.sav.** This is a hypothetical data file that concerns an insurance company that wants to build a model for flagging suspicious, potentially fraudulent claims. Each case represents a separate claim.

- **insure.sav.** This is a hypothetical data file that concerns an insurance company that is studying the risk factors that indicate whether a client will have to make a claim on a 10-year term life insurance contract. Each case in the data file represents a pair of contracts, one of which recorded a claim and the other didn't, matched on age and gender.

- **judges.sav.** This is a hypothetical data file that concerns the scores given by trained judges (plus one enthusiast) to 300 gymnastics performances. Each row represents a separate performance; the judges viewed the same performances.

- **kinship_dat.sav.** Rosenberg and Kim set out to analyze 15 kinship terms (aunt, brother, cousin, daughter, father, granddaughter, grandfather, grandmother, grandson, mother, nephew, niece, sister, son, uncle). They asked four groups of college students (two female, two male) to sort these terms on the basis of similarities. Two groups (one female, one male) were asked to sort twice, with the second sorting based on a different criterion from the first sort. Thus, a total of six "sources" were obtained. Each source corresponds to a 15×15 proximity matrix, whose cells are equal to the number of people in a source minus the number of times the objects were partitioned together in that source.

- **kinship_ini.sav.** This data file contains an initial configuration for a three-dimensional solution for *kinship_dat.sav*.

- **kinship_var.sav.** This data file contains independent variables *gender*, *gener*(ation), and *degree* (of separation) that can be used to interpret the dimensions of a solution for *kinship_dat.sav*. Specifically, they can be used to restrict the space of the solution to a linear combination of these variables.

- **marketvalues.sav.** This data file concerns home sales in a new housing development in Algonquin, Ill., during the years from 1999–2000. These sales are a matter of public record.

- **nhis2000_subset.sav.** The National Health Interview Survey (NHIS) is a large, population-based survey of the U.S. civilian population. Interviews are carried out face-to-face in a nationally representative sample of households. Demographic information and observations about health behaviors and status are obtained for members of each household. This data file contains a subset of information from the 2000 survey. National Center for Health Statistics. National Health Interview Survey, 2000. Public-use data file and documentation. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/.* Accessed 2003.

- **ozone.sav.** The data include 330 observations on six meteorological variables for predicting ozone concentration from the remaining variables. Previous researchers , , among others found nonlinearities among these variables, which hinder standard regression approaches.

- **pain_medication.sav.** This hypothetical data file contains the results of a clinical trial for anti-inflammatory medication for treating chronic arthritic pain. Of particular interest is the time it takes for the drug to take effect and how it compares to an existing medication.

- **patient_los.sav.** This hypothetical data file contains the treatment records of patients who were admitted to the hospital for suspected myocardial infarction (MI, or "heart attack"). Each case corresponds to a separate patient and records many variables related to their hospital stay.

- **patlos_sample.sav.** This hypothetical data file contains the treatment records of a sample of patients who received thrombolytics during treatment for myocardial infarction (MI, or "heart attack"). Each case corresponds to a separate patient and records many variables related to their hospital stay.

- **poll_cs.sav.** This is a hypothetical data file that concerns pollsters' efforts to determine the level of public support for a bill before the legislature. The cases correspond to registered voters. Each case records the county, township, and neighborhood in which the voter lives.

- **poll_cs_sample.sav.** This hypothetical data file contains a sample of the voters listed in *poll_cs.sav*. The sample was taken according to the design specified in the *poll.csplan* plan file, and this data file records the inclusion probabilities and sample weights. Note, however, that because the sampling plan makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*poll_jointprob.sav*). The additional variables corresponding to voter demographics and their opinion on the proposed bill were collected and added the data file after the sample as taken.

- **property_assess.sav.** This is a hypothetical data file that concerns a county assessor's efforts to keep property value assessments up to date on limited resources. The cases correspond to properties sold in the county in the past year. Each case in the data file records the township in which the property lies, the assessor who last visited the property, the time since that assessment, the valuation made at that time, and the sale value of the property.

- **property_assess_cs.sav.** This is a hypothetical data file that concerns a state assessor's efforts to keep property value assessments up to date on limited resources. The cases correspond to properties in the state. Each case in the data file records the county, township, and neighborhood in which the property lies, the time since the last assessment, and the valuation made at that time.

- **property_assess_cs_sample.sav.** This hypothetical data file contains a sample of the properties listed in *property_assess_cs.sav*. The sample was taken according to the design specified in the *property_assess.csplan* plan file, and this data file records the inclusion probabilities and sample weights. The additional variable *Current value* was collected and added to the data file after the sample was taken.

- **recidivism.sav.** This is a hypothetical data file that concerns a government law enforcement agency's efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender and records their demographic information, some details of their first crime, and then the time until their second arrest, if it occurred within two years of the first arrest.

- **recidivism_cs_sample.sav.** This is a hypothetical data file that concerns a government law enforcement agency's efforts to understand recidivism rates in their area of jurisdiction. Each case corresponds to a previous offender, released from their first arrest during the month of June, 2003, and records their demographic information, some details of their first crime, and the data of their second arrest, if it occurred by the end of June, 2006. Offenders were selected from sampled departments according to the sampling plan specified in *recidivism_cs.csplan*; because it makes use of a probability-proportional-to-size (PPS) method, there is also a file containing the joint selection probabilities (*recidivism_cs_jointprob.sav*).

- **rfm_transactions.sav.** A hypothetical data file containing purchase transaction data, including date of purchase, item(s) purchased, and monetary amount of each transaction.

- **salesperformance.sav.** This is a hypothetical data file that concerns the evaluation of two new sales training courses. Sixty employees, divided into three groups, all receive standard training. In addition, group 2 gets technical training; group 3, a hands-on tutorial. Each employee was tested at the end of the training course and their score recorded. Each case in the data file represents a separate trainee and records the group to which they were assigned and the score they received on the exam.

- **satisf.sav.** This is a hypothetical data file that concerns a satisfaction survey conducted by a retail company at 4 store locations. 582 customers were surveyed in all, and each case represents the responses from a single customer.

- **screws.sav.** This data file contains information on the characteristics of screws, bolts, nuts, and tacks .

- **shampoo_ph.sav.** This is a hypothetical data file that concerns the quality control at a factory for hair products. At regular time intervals, six separate output batches are measured and their pH recorded. The target range is 4.5–5.5.

- **ships.sav.** A dataset presented and analyzed elsewhere that concerns damage to cargo ships caused by waves. The incident counts can be modeled as occurring at a Poisson rate given the ship type, construction period, and service period. The aggregate months of service for each cell of the table formed by the cross-classification of factors provides values for the exposure to risk.

- **site.sav.** This is a hypothetical data file that concerns a company's efforts to choose new sites for their expanding business. They have hired two consultants to separately evaluate the sites, who, in addition to an extended report, summarized each site as a "good," "fair," or "poor" prospect.

- **smokers.sav.** This data file is abstracted from the 1998 National Household Survey of Drug Abuse and is a probability sample of American households. (*http://dx.doi.org/10.3886/ICPSR02934*) Thus, the first step in an analysis of this data file should be to weight the data to reflect population trends.

- **stroke_clean.sav.** This hypothetical data file contains the state of a medical database after it has been cleaned using procedures in the Data Preparation option.

■ **stroke_invalid.sav.** This hypothetical data file contains the initial state of a medical database and contains several data entry errors.

■ **stroke_survival.** This hypothetical data file concerns survival times for patients exiting a rehabilitation program post-ischemic stroke face a number of challenges. Post-stroke, the occurrence of myocardial infarction, ischemic stroke, or hemorrhagic stroke is noted and the time of the event recorded. The sample is left-truncated because it only includes patients who survived through the end of the rehabilitation program administered post-stroke.

■ **stroke_valid.sav.** This hypothetical data file contains the state of a medical database after the values have been checked using the Validate Data procedure. It still contains potentially anomalous cases.

■ **survey_sample.sav.** This data file contains survey data, including demographic data and various attitude measures. It is based on a subset of variables from the 1998 NORC General Social Survey, although some data values have been modified and additional fictitious variables have been added for demonstration purposes.

■ **telco.sav.** This is a hypothetical data file that concerns a telecommunications company's efforts to reduce churn in their customer base. Each case corresponds to a separate customer and records various demographic and service usage information.

■ **telco_extra.sav.** This data file is similar to the *telco.sav* data file, but the "tenure" and log-transformed customer spending variables have been removed and replaced by standardized log-transformed customer spending variables.

■ **telco_missing.sav.** This data file is a subset of the *telco.sav* data file, but some of the demographic data values have been replaced with missing values.

■ **testmarket.sav.** This hypothetical data file concerns a fast food chain's plans to add a new item to its menu. There are three possible campaigns for promoting the new product, so the new item is introduced at locations in several randomly selected markets. A different promotion is used at each location, and the weekly sales of the new item are recorded for the first four weeks. Each case corresponds to a separate location-week.

■ **testmarket_1month.sav.** This hypothetical data file is the *testmarket.sav* data file with the weekly sales "rolled-up" so that each case corresponds to a separate location. Some of the variables that changed weekly disappear as a result, and the sales recorded is now the sum of the sales during the four weeks of the study.

■ **tree_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.

■ **tree_credit.sav.** This is a hypothetical data file containing demographic and bank loan history data.

■ **tree_missing_data.sav** This is a hypothetical data file containing demographic and bank loan history data with a large number of missing values.

■ **tree_score_car.sav.** This is a hypothetical data file containing demographic and vehicle purchase price data.

■ **tree_textdata.sav.** A simple data file with only two variables intended primarily to show the default state of variables prior to assignment of measurement level and value labels.

- **tv-survey.sav.** This is a hypothetical data file that concerns a survey conducted by a TV studio that is considering whether to extend the run of a successful program. 906 respondents were asked whether they would watch the program under various conditions. Each row represents a separate respondent; each column is a separate condition.

- **ulcer_recurrence.sav.** This file contains partial information from a study designed to compare the efficacy of two therapies for preventing the recurrence of ulcers. It provides a good example of interval-censored data and has been presented and analyzed elsewhere .

- **ulcer_recurrence_recoded.sav.** This file reorganizes the information in *ulcer_recurrence.sav* to allow you model the event probability for each interval of the study rather than simply the end-of-study event probability. It has been presented and analyzed elsewhere .

- **verd1985.sav.** This data file concerns a survey . The responses of 15 subjects to 8 variables were recorded. The variables of interest are divided into three sets. Set 1 includes *age* and *marital*, set 2 includes *pet* and *news*, and set 3 includes *music* and *live*. *Pet* is scaled as multiple nominal and *age* is scaled as ordinal; all of the other variables are scaled as single nominal.

- **virus.sav.** This is a hypothetical data file that concerns the efforts of an Internet service provider (ISP) to determine the effects of a virus on its networks. They have tracked the (approximate) percentage of infected e-mail traffic on its networks over time, from the moment of discovery until the threat was contained.

- **wheeze_steubenville.sav.** This is a subset from a longitudinal study of the health effects of air pollution on children . The data contain repeated binary measures of the wheezing status for children from Steubenville, Ohio, at ages 7, 8, 9 and 10 years, along with a fixed recording of whether or not the mother was a smoker during the first year of the study.

- **workprog.sav.** This is a hypothetical data file that concerns a government works program that tries to place disadvantaged people into better jobs. A sample of potential program participants were followed, some of whom were randomly selected for enrollment in the program, while others were not. Each case represents a separate program participant.

Notices

Licensed Materials – Property of SPSS Inc., an IBM Company. © Copyright SPSS Inc. 1989, 2010.

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing,

using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Index