Introduction to IBM SPSS Statistics
Student Guide

Course Code: 0G507
ERC 1.0

Introduction to IBM SPSS Statistics

0G507

Published October 2010

Table of Contents

Lesson 1: Course Introduction

1.1 *Course Objectives*

After completing this lesson students will be able to:

- Perform useful analyses on their data using PASW® Statistics

To support the achievement of this primary objective, students will also be able to:

- Explain the use of PASW Statistics for basic data analysis
- Import data from different types of file formats
- Define, save and view variable properties
- Use the Data Editor to enter data values, work with Data Editor features and multiple Data Editor windows
- Summarize individual variables using tables and graphs
- Group values of variables using various methods
- Create new variables using the **Compute Variable** dialog
- Analyze relationships between categorical variables and between categorical and scale variables
- Select cases in a data file using various methods
- Present results with charts
- Use the Output Viewer and its editing features
- Use basic syntax to automate analyses
- Use the Toolbars
- Use all types of help available in PASW Statistics

1.2 *About SPSS, an IBM Company*

SPSS® Inc., an IBM® Company is a leading global provider of predictive analytics 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 solutions address interconnected business objectives across an entire organization by focusing on the convergence of analytics, IT architecture and business process. Commercial, government and academic customers worldwide rely on SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. SPSS was acquired by IBM in October 2009. For more information, visit http://www.spss.com.

1.3 *Supporting Materials*

Supporting Materials

The following materials are used in this course. All of the files can be found in the *C:/Train/StatisticsIntro* directory. All paths in the following lessons will be relative to this directory. If the course is being conducted in a non-SPSS, an IBM Company sponsored facility, the instructor will define the base directory for files.

- *Census_Small.sav* – a PASW Statistics data file. These data are a subset of demographic and attitudinal variables from a survey conducted in 2008 of a sample of the general population.
- *Census_Small.xls* – an Excel file in Excel 2003 format of the Census data.
- *Census_Small.mdb* – an Access file of the Census data.
- *Census_Small.txt* – a delimited text file of the Census data.

- *Employee data.sav* – a PASW Statistics data file. The data contains salary and job information on employees of a banking institution.
- *Employee data.xls* – an Excel file in Excel 2003 format of the Employee data.
- *Employee data.mdb* – an Access file of the Employee data.
- *Employee data.dat* – a delimited text file of the Employee data.

- *Customer satisfaction 2008.sav* – a PASW Statistics data file. The hypothetical data contains demographic and satisfaction variables for shoppers in retail stores in several cities collected in 2008.
- *Customer satisfaction 2009.xls* – an Excel file in Excel 2003 format of customer satisfaction data collected in 2009.
- *Customer satisfaction.dat* – a delimited text file of the Customer Satisfaction data.
- *Customer satisfaction 2009.sav* – a PASW Statistics data file. The hypothetical data contains demographic and satisfaction variables for shoppers in retail stores in several cities collected in 2009.
- *Sample_Viewer.spv* – a sample PASW Statistics Output file used to demonstrate managing and editing results.

1.4 *Course Introduction*

PASW Statistics provides many facilities for managing, modifying, and analyzing data. This course introduces the student to the steps of data analysis and how to use basic elements of PASW Statistics to perform those steps and answer basic questions from your data.

1.5 *Course Assumptions*

The following points are assumed going into this course. They are **not addressed** in the course.

- Attendees have access to PASW Statistics 18 with PASW® Statistics Base.
- Attendees have access to the supporting course files listed in the Supporting Materials section above.
- No prior experience with PASW Statistics 18 is required.
- No experience with statistical analysis is required although an introductory statistics course (or equivalent) would be helpful.
- Attendees will work independently with PASW Statistics 18 to gain proficiency with it.

Lesson 2: Introducing IBM SPSS Statistics

2.1 *Objectives*

After completing this lesson students will be able to:

- Explain the use of PASW Statistics for basic data analysis

To support the achievement of this primary objective, students will also be able to:

- Explain the basic steps of data analysis using PASW Statistics
- Describe the roles of the primary windows within PASW Statistics
- Describe the basic layout of PASW Statistics dialog boxes

2.2 *Introduction*

This lesson introduces the PASW Statistics environment and demonstrates a typical session. We use PASW Statistics to read a previously defined PASW Statistics data file and then produce a simple statistical summary and chart. You will learn the roles of the primary windows within PASW Statistics and see a few features that smooth the way when running analyses.

More detailed instruction about many of the topics in this lesson will follow in later lessons. Here, we give a basic framework for understanding and using PASW Statistics.

Business Context

PASW Statistics is software for data analysis. Typical results are reports, statistical tests and charts, using data such as respondents to a survey or customers of a bank as input. Some examples include:

- Monthly reports and charts to track patient statistics such as length of stay and reason for admissions to a hospital.
- Examine the relationship of demographic information and propensity to purchase a product.
- Analysis to determine what factors relate to satisfied customers.

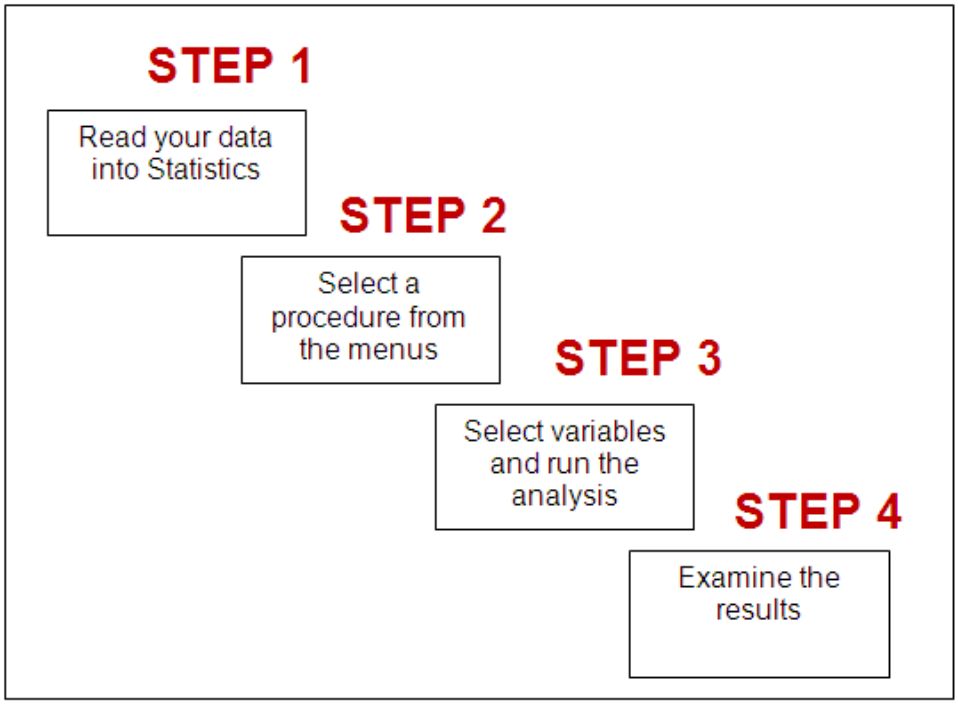
Supporting Materials

Census_Small.sav – a PASW Statistics data file. These data are a subset of demographic and attitudinal variables from a survey conducted in 2008 of a sample of the general population.

2.3 *Basic Steps of Data Analysis*

Given a dataset, many types of data analysis can be done. The analysis always requires the same steps: read your data, select a procedure from the menus, run a descriptive report, graph or statistical procedure and finally examine the results. When you run your analysis in PASW Statistics, this translates to the four basic steps shown in the figure below.

Figure 2.1 Basics Steps of Data Analysis



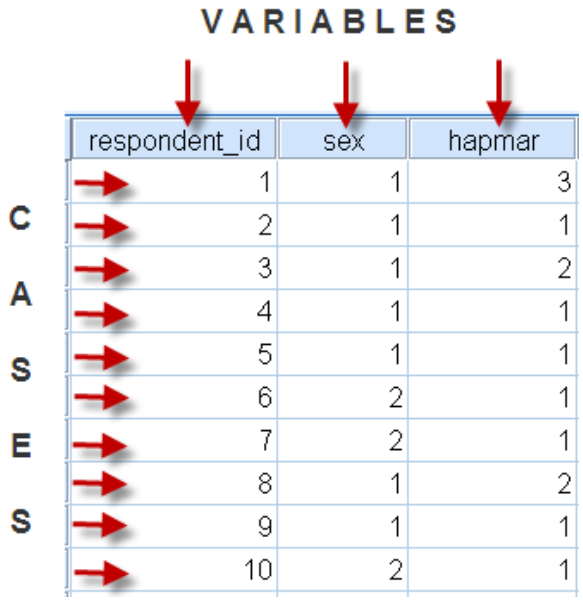
STEPS	ACTIONS
Read your data into PASW Statistics	Read the data into PASW Statistics by opening a data file or enter the data values directly into the PASW Statistics Data Editor.
Select a procedure from the menus	Select a procedure from the PASW Statistics **Analyze** or **Graphs** menus to calculate statistics, produce a report, create charts.
Select variables and run the analysis	In the procedure dialog box, select the variables from the data that will be used in the graph, report, or statistical procedure and run the procedure.
Examine the results	Examine and interpret the results.

Often times, data analysis requires an additional data transformation step to create new variables or modify values of an existing variable before you can get the desired report or graph. For example, you might want to report individuals in age groups rather than by actual years of age.

Data Analysis Example

For example, data are collected for a social survey. The figure below shows the structure of the data. Respondents who participated in the survey define the rows, information about the respondent representing the answers to the survey questions such as gender and happiness of marriage define the columns. Typically, PASW Statistics does not use the terms “rows” and “columns” for the data, but “cases” and “variables”. Cases represent the units of analysis and variables represent the items that have been measured.

Figure 2.2 Data Structure



In PASW Statistics, variable names are used **(sex, hapmar)** and data values are often entered using numeric values for the answers (e.g. 1 for a male respondent, 2 for a female respondent). You can define descriptive labels for the variables and values. These labels are displayed in PASW Statistics tables and charts,

One goal of analyzing this dataset might be to investigate how many people are happily married and if there are differences between men and women. Typical tables for this analysis are shown below. From these, we can conclude that most people are very happy in their marriage and that there are no big differences between men and women.

Figure 2.3 Tables reporting Happiness of Marriage

HAPPINESS OF MARRIAGE

		Frequency	Percent
Valid	VERY HAPPY	596	61.5
	PRETTY HAPPY	343	35.4
	NOT TOO HAPPY	30	3.1
	Total	969	100.0

		RESPONDENTS SEX		
		MALE	FEMALE	Total
		%	%	%
HAPPINESS OF MARRIAGE	VERY HAPPY	63.2%	59.9%	61.5%
	PRETTY HAPPY	34.0%	36.7%	35.4%
	NOT TOO HAPPY	2.8%	3.4%	3.1%
Total		100.0%	100.0%	100.0%

Note

The tables show the labels for the variables (HAPPINESS OF MARRIAGE, not *hapmar*; RESPONDENTS SEX, not *sex*) and displays the value labels and not the values (e.g. MALE and FEMALE and not values 1 and 2 for RESPONDENTS SEX)

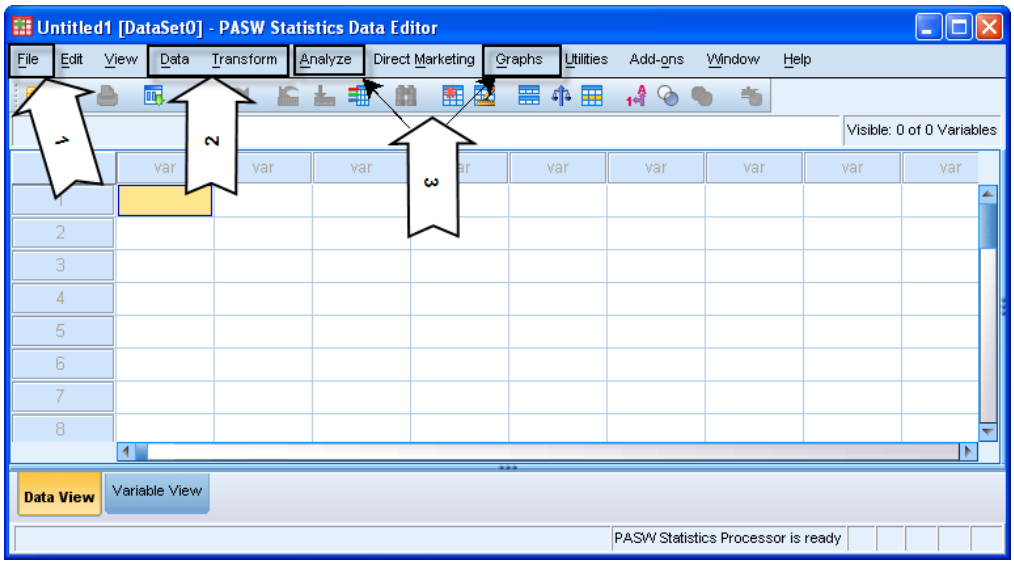
2.4 *Procedure: Running a Basic Analysis*

First, PASW Statistics must be started. Assuming a standard installation of PASW Statistics, this is done from the Windows Start menu:

Select **All Programs...SPSS Inc... PASW Statistics 18.0...PASW Statistics 18.0** (alternatively, click on the PASW Statistics 18.0 shortcut on your Desktop if you have one).

When you start PASW Statistics, a dialog box opens, prompting you to access data or run a tutorial. Click Cancel button to close this dialog box. One window, the Data Editor window is then displayed. The **Data Editor** window contains the data to be analyzed. At least one Data Editor window must be open in each PASW Statistics session. Other types of windows are opened as required during a session.

Figure 2.4 Data Editor Window (Data View)



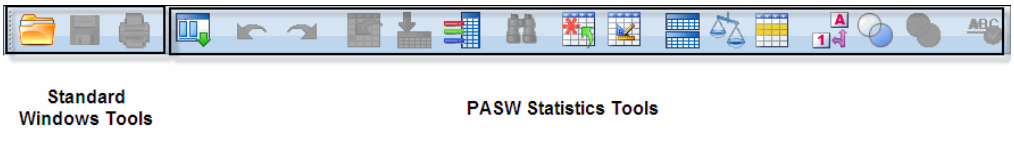
The Main Menu reflects the basic steps of data analysis.

1) File menu: Read Data
2) Data and Transform menus: Data Preparation (optional step)
3) Analyze and Graphs menus: Produce reports, statistics, and charts

The other menu items provide support and customizing functions, such as changing fonts in the data view (the View menu), performing standard Windows functions to cut, copy and paste (the Edit menu), improving efficiency (the Utilities menu) and asking for assistance (the Help menu).

You can also access many of the most commonly used menu features using the toolbar. The Data Editor toolbar contains common edit operations and operations associated with the Main Menu. You can place the cursor over any tool button and an explanation popup label will appear. The five tool buttons (Open, Save, Print, Undo, Redo) are common to most Windows applications and are common to all types of PASW Statistics windows. Other tool buttons are specific to PASW Statistics.

Figure 2.5 Data Editor Toolbar



Further Information

Place the mouse curser over a tool button to see a popup help tip.

Refer to *Menu and Toolbar Reference* appendix for explanations of each PASW Statistics Tool.

Read the Data

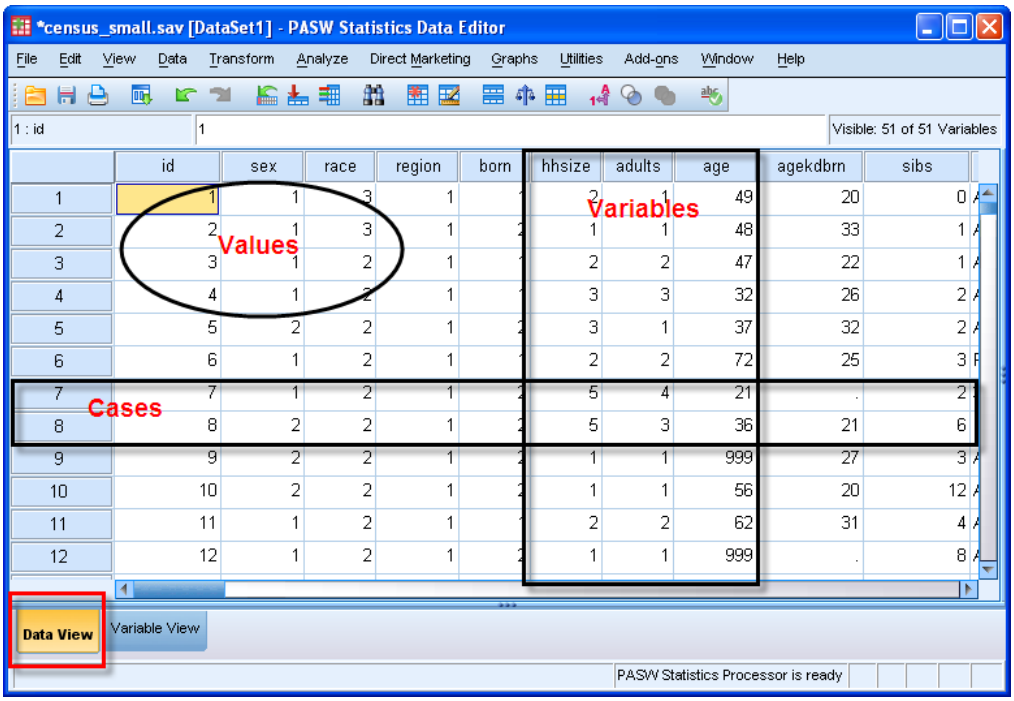
To begin the analysis, read the data. PASW Statistics data file is the data format saved by PASW Statistics. A PASW Statistics **data** file (file extension .sav) includes data descriptions, e.g. labels for the variables and labels for the values as well as the data values.

To open a PASW Statistics data file, from the main menu

Select **File...Open...Data** and select the data file.

After the data file is opened, the data are displayed in the Data Editor, Data View tab.

Figure 2.6 Data Displayed in Data Editor Data View



The rows contain the cases or units of analysis; in the columns, the variables; in the cells of the table, the data values (value for case x on variable y).

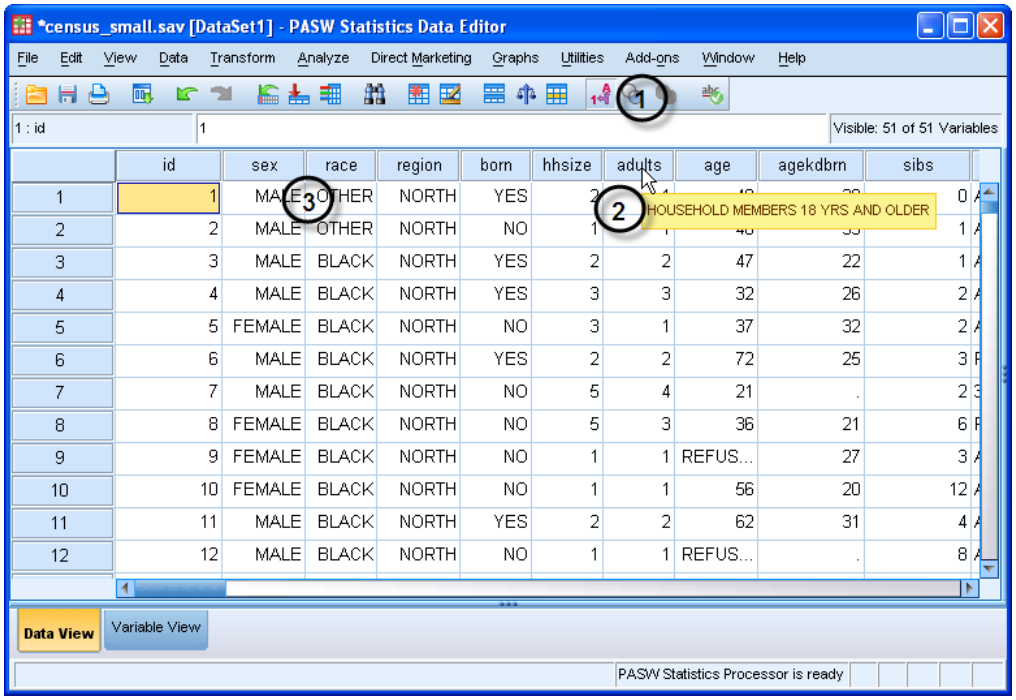
Data View Display Options

Typically, each variable and each value has a descriptive label. We can display these labels in the Data View:

- Click the  button in the Toolbar (or select **View…Value Labels** from the main menu) to display the value labels. Note, that the data are still numeric values, but displaying the value labels makes it easier to interpret the responses.

- Place the mouse cursor on a variable name (the column headings) to display a more descriptive variable label is displayed if one has been defined for that variable.

Figure 2.7 Data Editor Data View Display Options



1) Button  to toggle value labels on/off
2) Mouse curser to display Variable Label
3) Value Labels displayed in the cells of the Data View

All of the variable properties such as the variable and value labels are displayed in the Variable View tab of the Data Editor.

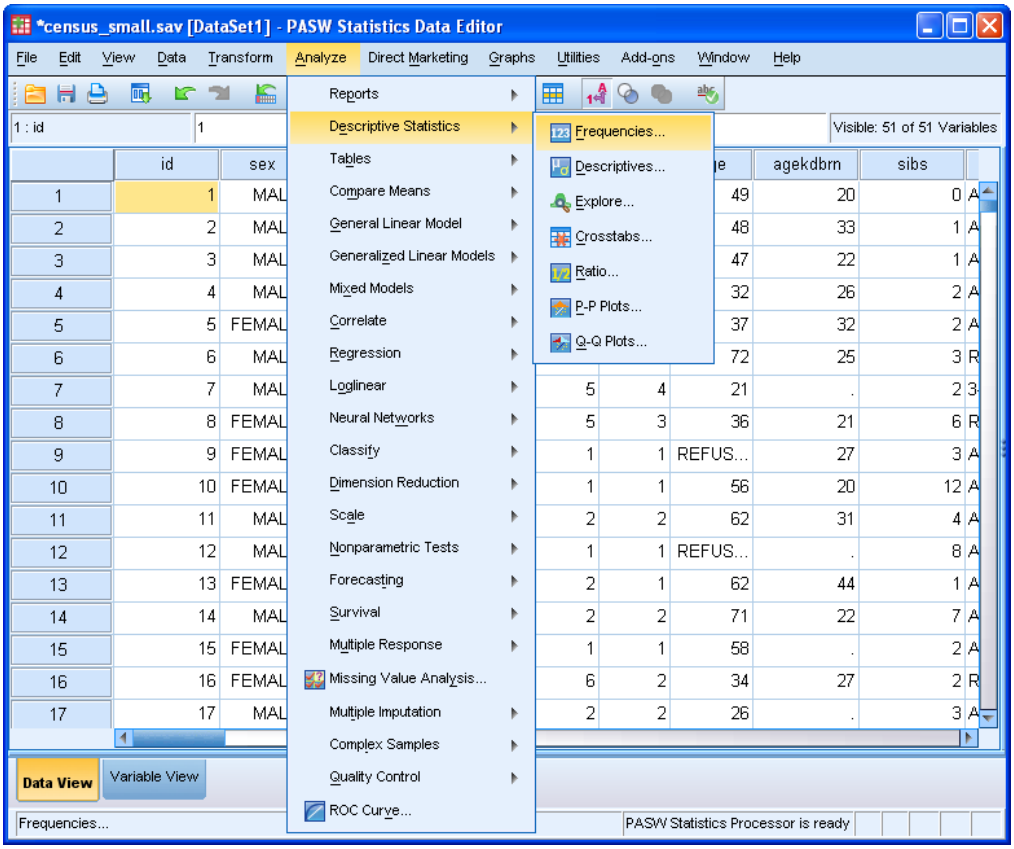
Further Information

Variable properties and how to assign them are discussed in the *Variable Properties* lesson.

Select a Procedure

The next step in our analysis is to select a procedure. The Analyze menu contains the list of reporting and statistical analysis categories. Most of the categories are followed by an arrow, which indicates that there are several analysis procedures available within the category; they will appear on a submenu when the category is selected.

Figure 2.8 Select a Procedure: Analyze Menu



To select a procedure, select Analyze, an analysis category, and then the procedure. This will open the procedure dialog box.

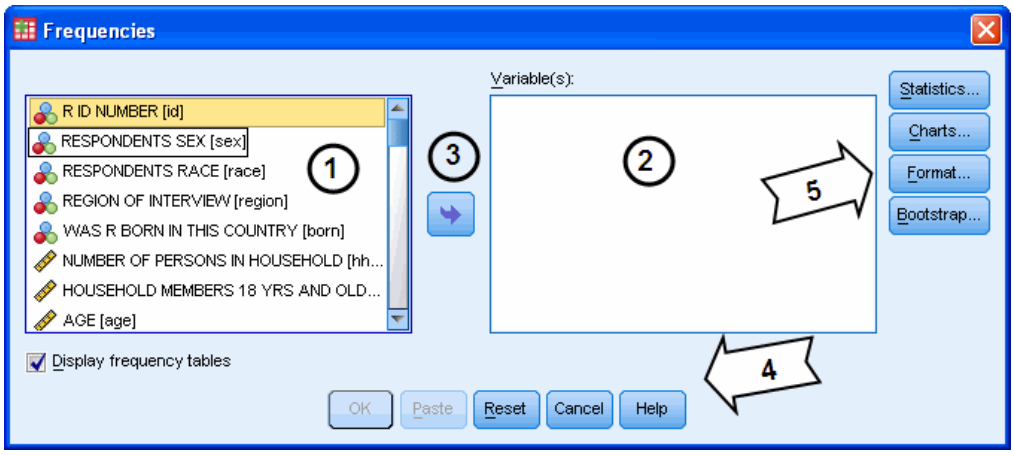
For example, to run the Frequencies procedure which produces simple frequency tables (tables of counts) and bar charts:

Select **Analyze…Descriptive Statistics…Frequencies**

Select Variables and Run Analysis

The variables are selected and the analysis is run from the procedure dialog box. Most procedure dialog boxes have the same basic components and contain a number of common features.

Figure 2.9 Select Variables: Dialog Box



Each procedure dialog box contains the components:

1) Source variables in the dataset
2) Target variables selected for analysis
3) Moves selected variable(s) to the target list (alternatively, use drag & drop to move variables)
4) Control buttons
5) Subdialog boxes for optional specifications

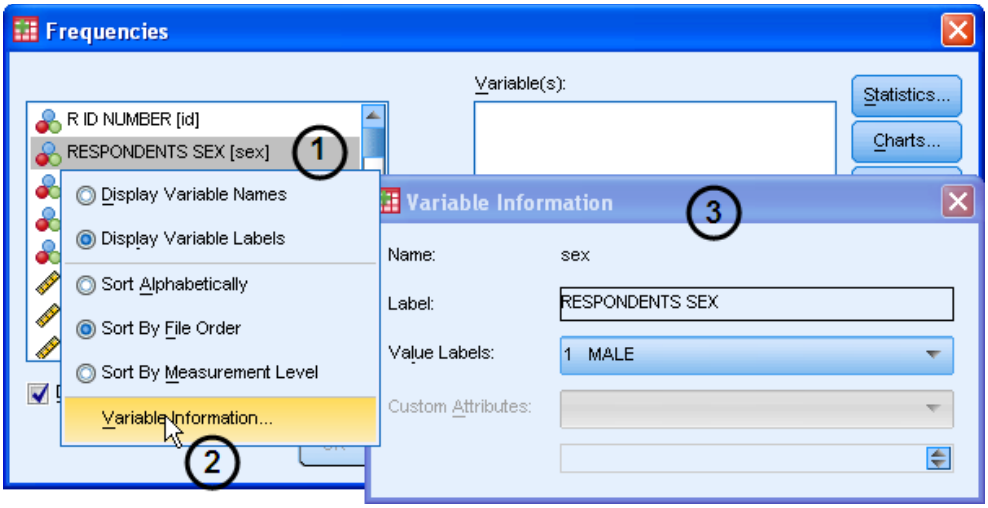
Tip

Any of the PASW Statistics dialog boxes can be resized to be larger, which allows the user to display a wider variable list.

In the source and target variable lists the variable label is shown, followed by the variable name in square brackets. If there were no variable labels, only the variable names would appear in the list box.

Use the right mouse key to display a variable description popup box with additional information on any variable in the source list.

Figure 2.10 Variable Information from Procedure Dialogs



1) Right-click a source variable.
2) Select Variable Information from the popup menu.
3) Displays Variable Information for the selected variable

The Right-click popup menu also allows the user to change the display (show variable label or not) and the order of variables (alphabetically or file order).

Tip

You can locate variables in the source list of most dialogs by typing the first letter of the label which will go to the first variable label that begins with that letter. Repeatedly typing the letter will move through the list in sequence to each variable label beginning with that letter.

Further Information

The icons that are displayed next to variables in the dialog box provide information about the variable type and measurement level. See the lesson *Variable Properties* for a discussion.

The control buttons that appear along the bottom of the dialog box instruct PASW Statistics to perform an action.

OK Runs the procedure. The OK button is disabled (grayed out) until the minimum specification is completed. For the Frequencies dialog, the minimum is at least one variable in the target variable list.

Reset Resets all specifications made in the dialog box and the associated sub-dialog boxes and keeps the dialog box open.

Cancel Cancels the selections and closes the dialog box without running the procedure.

Help Opens the PASW Statistics Help facility to the help relevant for this dialog box.

Paste Pastes the PASW Statistics syntax for the specification into the Syntax Editor window.

Further Information

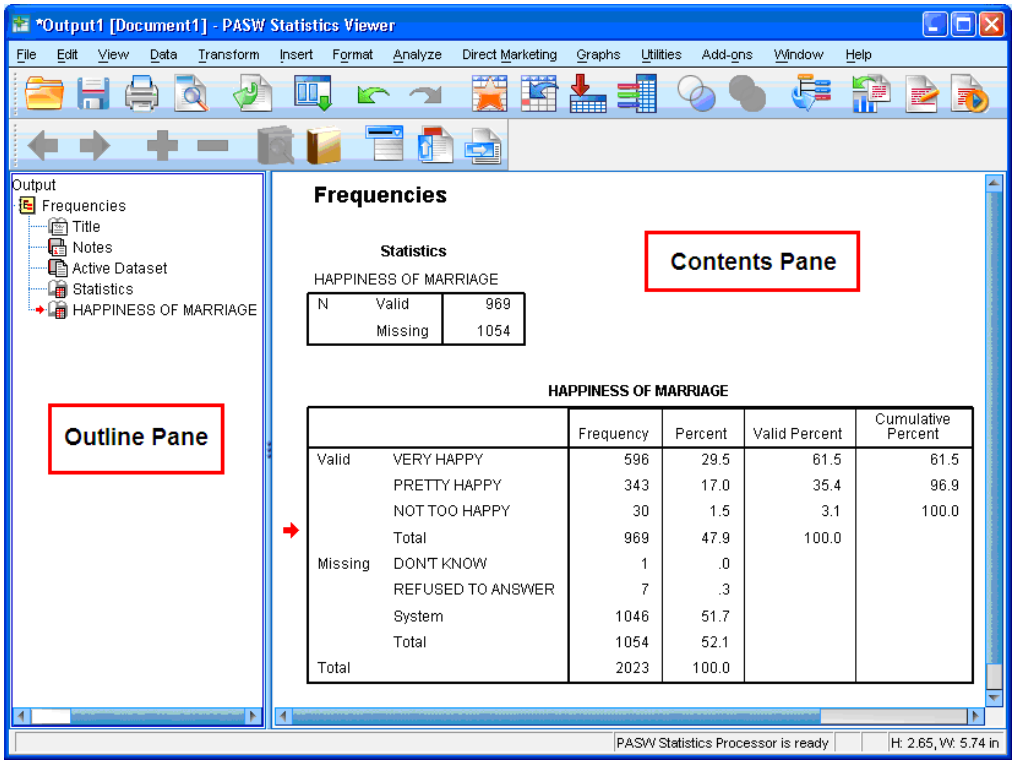
The Paste button is discussed in the *PASW Statistics Syntax* lesson.

Since PASW Statistics procedures provide a great deal of flexibility, often not all of the possible choices are in a single dialog box. The main dialog box contains the minimum information required to run the procedure. Additional optional specifications can be made in subdialog boxes. The subdialog boxes are accessed from control buttons located on the right side of the main dialog box. Buttons with an ellipsis (...) after the name indicate an optional subdialog box will be displayed. For example, there are three optional subdialog boxes for the main Frequencies dialog box: **Statistics…**, **Charts…**, and **Format…** Subdialog boxes do not show an OK button, but instead have a Continue button, to return to the main dialog box.

Examine the Results

When a procedure has been run, the results are displayed in a separate window, the **Output Viewer** window. The Viewer window has two sections or panes. The outline pane on the left side contains an outline or list of all items displayed in the contents pane on the right.

Figure 2.11 Output Viewer Results



You can quickly go to any item in the Viewer by selecting it in the outline pane on the left.

2.5 *Demonstration: Running a Basic Analysis*

In this demonstration we will open a file containing a subset of data collected for a general social survey, in 2008. We will run the Frequencies procedure on some variables of interest.

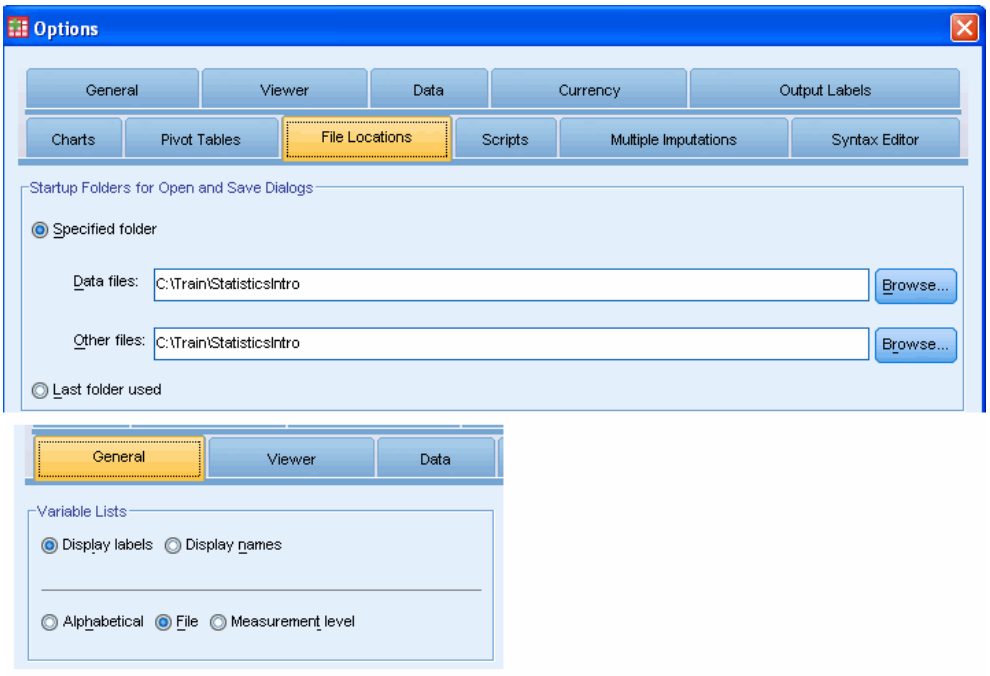
Supporting Materials	We use the *Census_small.sav* PASW Statistics data file for this demonstration All of the files can be found in the *C:/Train/StatisticsIntro* directory. If the course is being conducted in a non-SPSS, an IBM Company sponsored facility, the instructor will define the base directory for files.

Setting the Default Startup Folder and Variable Display in Dialog Boxes

The startup folder that will appear in all Open and Save dialog boxes can be set using the **File Locations** tab on the **Edit...Options** menu. Change the Startup Folders for the Data Files and Other Files to the course folder.

The settings for display of the variables in the source list of dialog boxes can be set in the **Edit...Options** menu on the **General** tab. In this course, the examples are shown with the variable labels displayed in file order. For a new user of PASW Statistics, this provides a more complete description of the variables. This is the default setting and can be confirmed before proceeding.

Figure 2.12 Startup Folder Location Setting in Edit Options Dialog



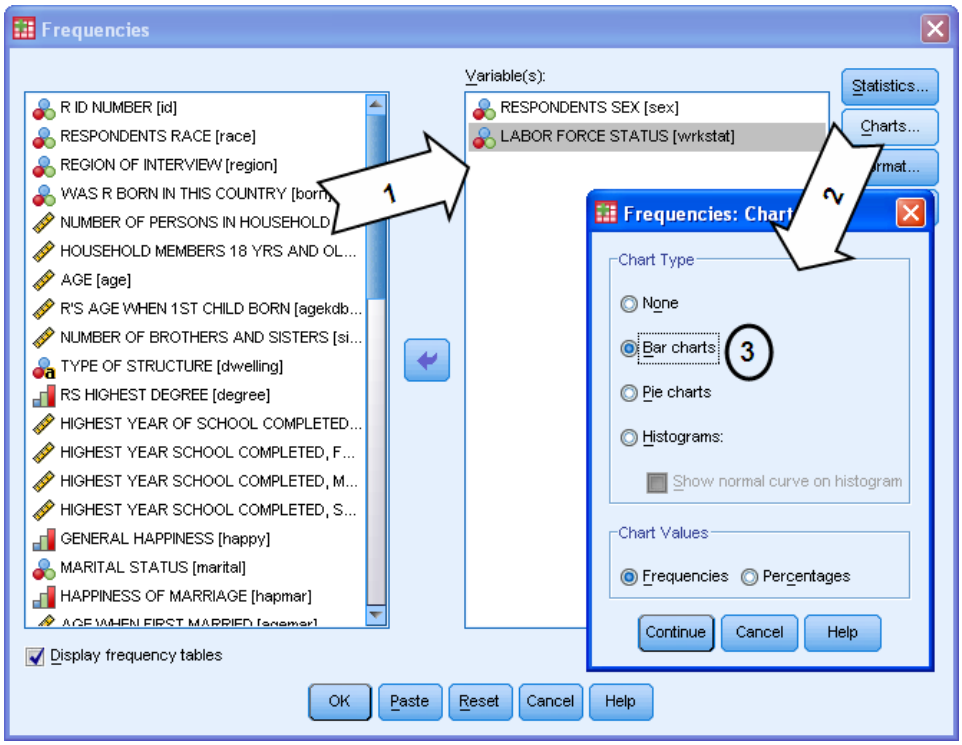
Start PASW Statistics if you haven't already and open the *census_small.sav* using the **File...Open...Data** menu or the tool .

If you are new to a dataset, it's always a good idea to get more information by looking at the variable labels and turning value labels on. This dataset contains a number of background variables, for example, *sex, race, region*; information about the respondent's, parents' and spouse's education and socio economic status (*educ, paeduc, maeduc, speduc, sei, pasei, masei* and *spsei*); and the respondent's opinion with respect to a number of issues such as trust in education, trust in financial system.

We will produce frequency tables (tables of counts) and bar charts for two variables in our data, *RESPONDENTS SEX* and *LABOR FORCE STATUS*. Charts can be an important resource in the understanding of your data and the presentation of results.

Select **Analyze…Descriptive Statistics…Frequencies**

Figure 2.13 Specifying Frequencies

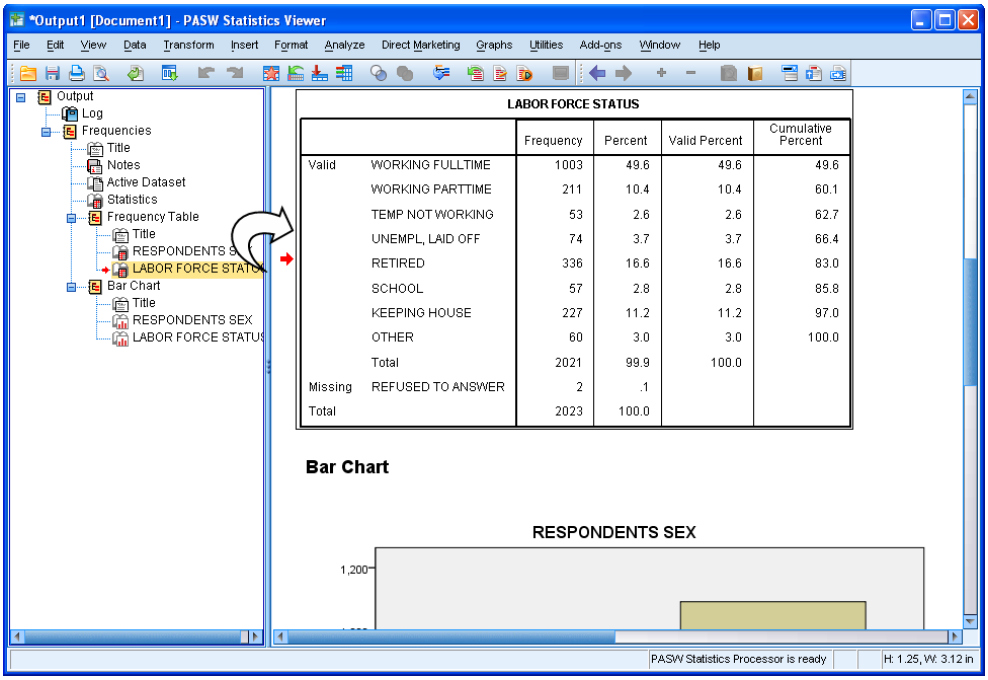


1) Place *RESPONDENTS SEX [sex]* and *LABOR FORCE STATUS [wrkstat]* in the target Variable(s) list.
2) Select **Chart** subdialog.
3) Select **Bar Charts**.

Results for a Basic Analysis

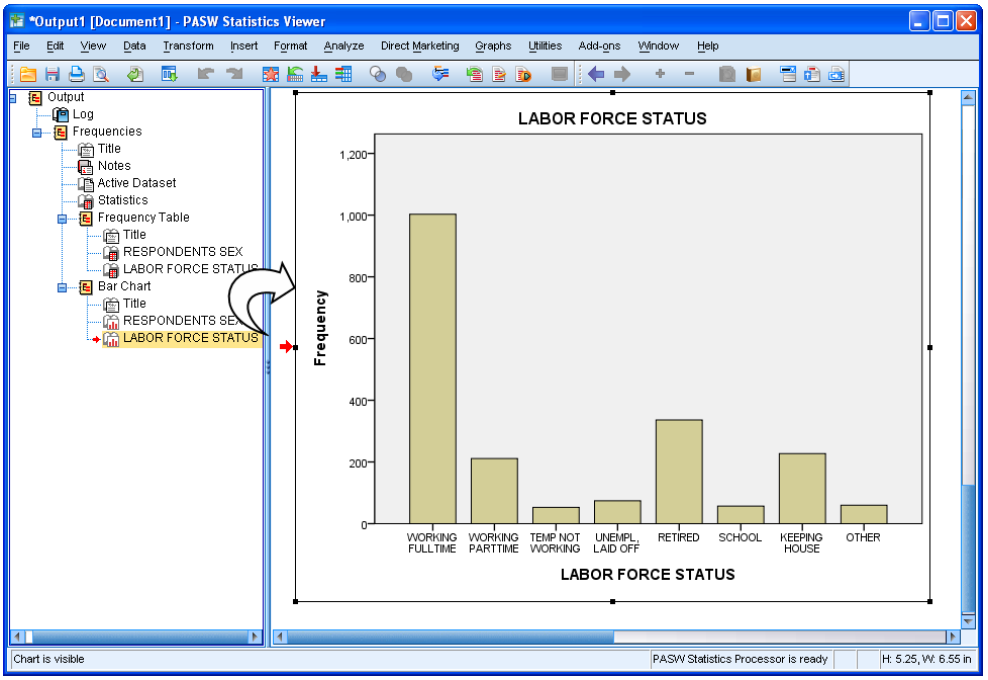
After clicking **OK** in the Frequencies dialog box, all of the results are displayed in the Output Viewer window. You can quickly go to any table or chart in the Viewer by selecting it in the outline pane on the left. We examine some results.

Figure 2.14 Reviewing Frequencies Results: Frequency Table



Select **LABOR FORCE STATUS** under Frequency Table in the Outline Pane to view the table for labor force status. To view the bar chart for this variable, select LABOR FORCE STATUS under Bar Chart in the Outline Pane.

Figure 2.15 Reviewing Frequencies Results: Bar Chart



Further Information

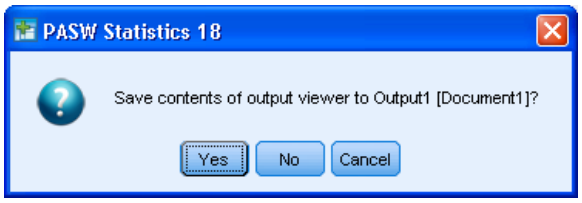
The charts and tables can be edited by double-clicking on them in the contents pane of the Viewer window, and the results can be copied and pasted into other applications. These topics are covered in the lesson *Editing and Creating Charts* and the lesson *PASW Statistics Viewer*.

After running this basic analysis and viewing the results, exit PASW Statistics.

Select **File…Exit**

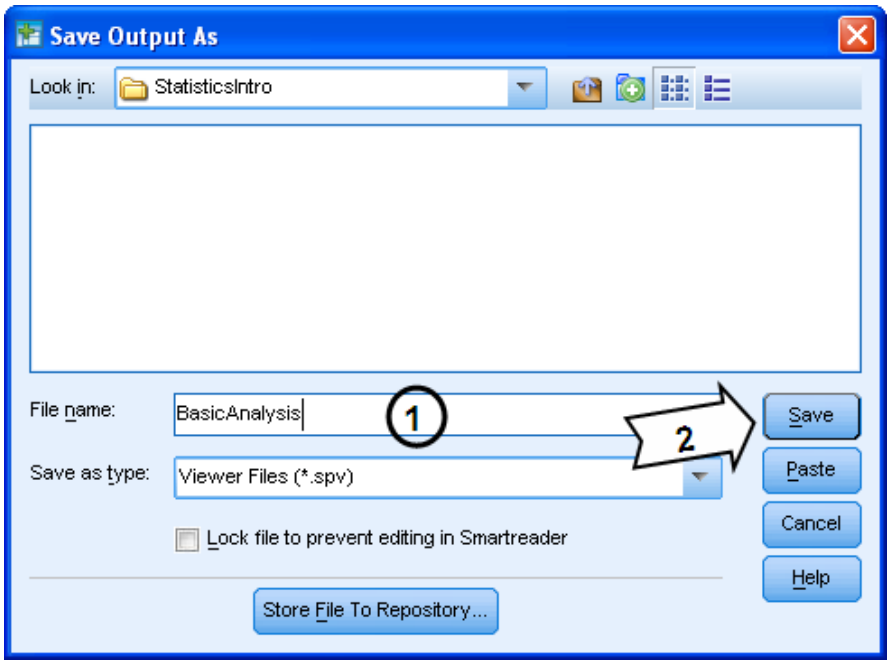
A message box will display warning that the output has not been saved.

Figure 2.16 Warning Alert to Save Results in Output Viewer



To save the output, select **Yes** and the **Save Output As** dialog will open where you save the results in a PASW Statistics output file.

Figure 2.17 Save Output Viewer Dialog Box



1) Type the File name for the output file. By default, the file is saved in the startup folder. To save in a different folder, select the new folder from the **Look In:** dropdown list.
2) Select **Save** button. The output will be saved in the specified folder.

The file extension for a Viewer file is **.spv.** This is the PASW Statistics format for a Viewer document containing the table and chart results**.**

Apply Your Knowledge

For each of the following statements, answer True or False.

1. "Cases" are the columns in the Data Editor window (Data View tab).

2. Value labels are descriptive labels for the variable names.

3. In PASW Statistics, data and output (e.g. a Frequencies table) are displayed in the same window.

4. .sav is the extension for a PASW Statistics data file.

5. .spv is the extension for a PASW Statistics Viewer file.

6. The Data Editor window Data View can display either values or value labels.

7. Turning value labels on/off in the Data Editor can be done variable-by-variable (so, for example, value labels for variable sex are displayed in the Data View, while value labels for region are not displayed in the Data View).

2.6 *Lesson Summary*

In this lesson we described the PASW Statistics environment and demonstrated running a basic analysis.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Explain the use of PASW Statistics for basic data analysis

And, they should also be able to:

- Explain the basic steps of data analysis using PASW Statistics
- Describe the roles of the primary windows within PASW Statistics
- Describe the basic layout of PASW Statistics dialog boxes

Lesson 3: Reading Data

3.1 *Objectives*

After completing this lesson students will be able to:

- Import data from different types of file formats

To support the achievement of this primary objective, students will be able to

- Describe choices on the File menu to read and save data files
- Read Excel files
- Read from an Access database
- Read delimited text files

3.2 *Introduction*

Getting data into PASW Statistics is the first step before any analysis can be done. If the data are available in a PASW Statistics data file (file extension .sav), the lesson *Introducing PASW Statistics* showed how to open this file type. However, different file types such as Excel, Access, and text formats can be read. This lesson discusses how to open files of these types.

Supporting Materials

The *Census_Small* data in several file formats are used in this lesson. These data are a subset of demographic and attitudinal variables from a survey conducted in 2008 of a sample of the general population.

- *Census_Small.xls* – Excel 2003 format file
- *Census_Small.mdb* – Access 2003 database file
- *Census_Small.txt* – Comma delimited text file

3.3 *The File Menu*

All types of files are read from the File menu. Three selections are relevant to reading data.

Menu entry	Files of types
Open…Data	PASW Statistics; Lotus; Excel; dBase, SAS, Stata
Open Database…New Query	Any database for which an ODBC driver is available
Read Text Data	Delimited (e.g. comma or tab) text data; fixed text data

Saving data is also done from the File menu.

Menu entry	Files of types
Save, Save as	PASW Statistics; Lotus; Excel; dBase, text (delimited, fixed)
Export to Database	Any database for which an ODBC driver is available

In this lesson we will focus on reading data files from common file types. After you have imported a data file, it is always a good idea to save it into a PASW Statistics data file.

3.4 *Reading Excel and Other Spreadsheet Files*

PASW Statistics 18 reads Excel files as of Excel 2.1 through Excel 2007. If the file originates from an Excel version earlier than Excel 95 or from another spreadsheet product, the dialog box will not present choices related to features not available in these earlier file versions. For purposes of our discussion, we will consider Excel 95 or later files.

Further Information

Refer to PASW Statistics Help for considerations if reading files originating with versions of Excel earlier than Excel 95.

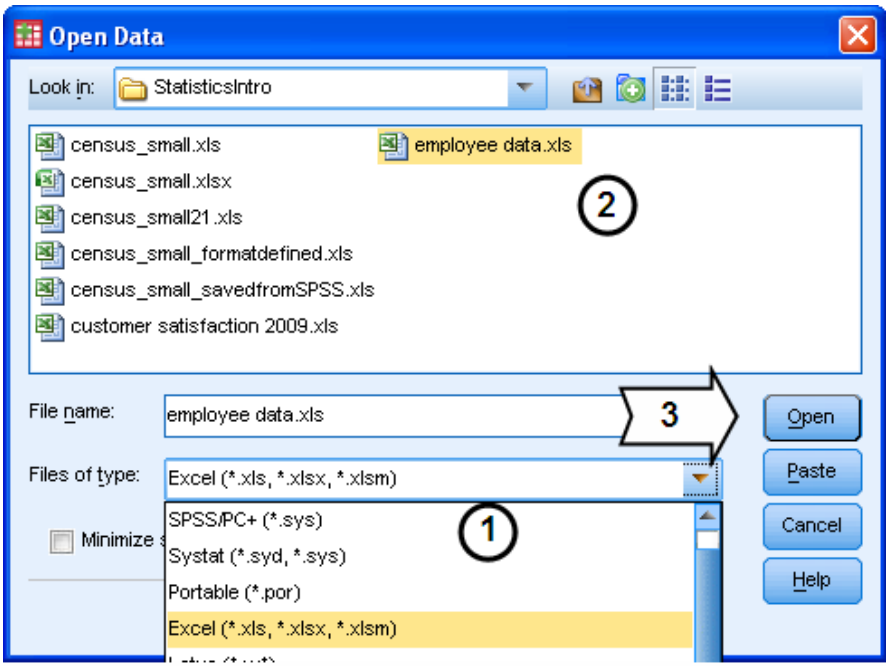
Requirements when reading an Excel file are:

- PASW Statistics assumes that all information is data values.
- You must specify in PASW Statistics if the first row in the worksheet contains variable names. If other cells contain information other than data values, they must be deleted from the spreadsheet before importing into PASW Statistics or excluded by specifying a range of cells in PASW Statistics.
- Each column of the spreadsheet is read as a PASW Statistics variable.
- PASW Statistics attempts to classify the data type (e.g.: whether the columns contain string or numeric data). For Excel 95 or later files, the data type and width for each variable are determined by the data type and width in the Excel file. If the column contains more than one data type (for example, date and numeric), the data type is set to string, and all values are read as valid string values. For files from earlier versions of Excel, the data type is determined by examining the first data values. Therefore the first row of data values should contain values indicative of the type of data.
- PASW Statistics reads only the values in the cells of the spreadsheet. Formulas in the spreadsheet will be computed and these computed values are imported into PASW Statistics. The formulas and other spreadsheet characteristics associated with the cells will not be imported into PASW Statistics.
- When reading the Excel file into PASW Statistics, the file must not be open in Excel.

3.5 *Procedure: Reading Excel and Other Spreadsheet Files*

An Excel file is opened from the **File…Open...Data** menu.

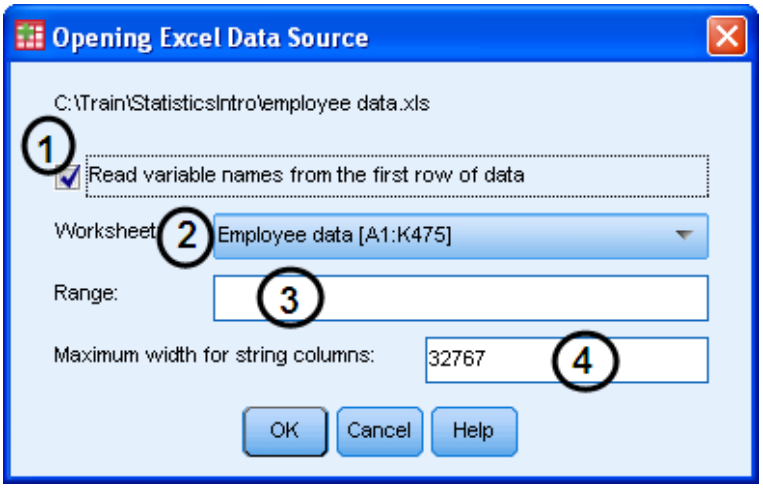
Figure 3.1 Open Data (Excel file) Dialog Box



1) Select Files of type: **Excel (*.xls, *.xlsx, *.xlsm).**
2) Select the file. Browse to the folder in the Look in: dropdown list if necessary.
3) Select the **Open** button.

This opens a dialog box which specifies how the Excel file should be read.

Figure 3.2 Opening Excel Data Source Dialog Box



1) Check on if first row in the worksheet contains text for variable names rather than data values.
2) Select worksheet (available for Excel files later than Excel 95).
3) Specify Range using the Excel range for cells.
4) Specify maximum width for string (alphanumeric) columns.

If the column headings in the worksheet do not conform to the PASW Statistics variable-naming rules, they are converted into valid variable names and the original column headings are saved as variable labels.

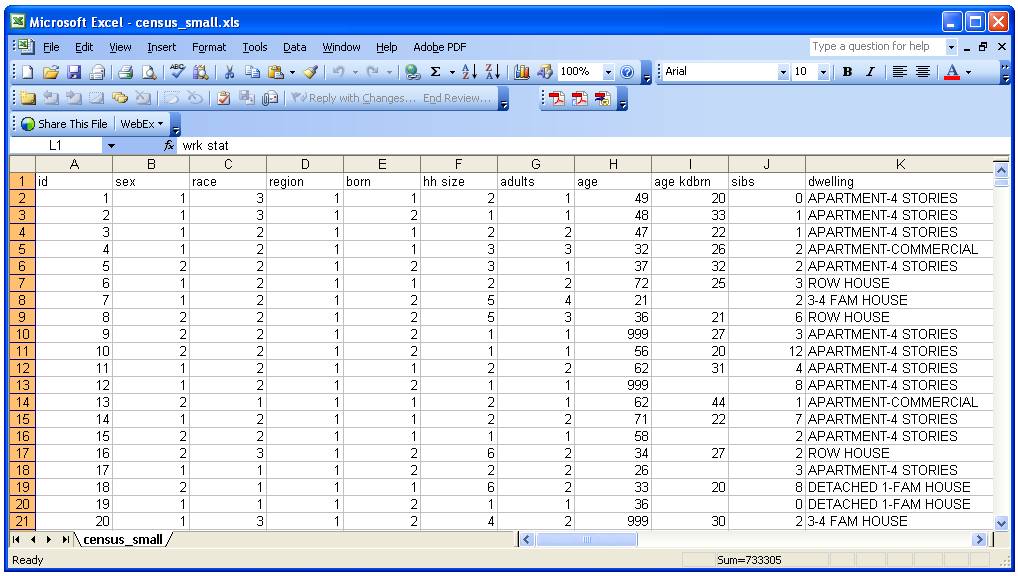
Tip

If you want to import only a portion of the worksheet, specify the range of cells to be imported in the Range text box, e.g. A1:E481. Leave the Range textbox blank for automatic selection of rows and columns.

3.6 *Demonstration: Reading an Excel File*

Below is an example of an Excel 2003 workbook (file *census_small.xls*), with worksheet *census_small* selected. Note that this Excel file contains variable names in the first row and the first observation of data is in row 2. Some variable names contain blanks (e.g. *hh size*). Blanks in variable names are not allowed in PASW Statistics, so PASW Statistics will convert these names to valid PASW Statistics variable names. Finally, note that the field *dwelling* contains string and not numeric values.

Figure 3.3 Excel File Displayed in Excel 2003



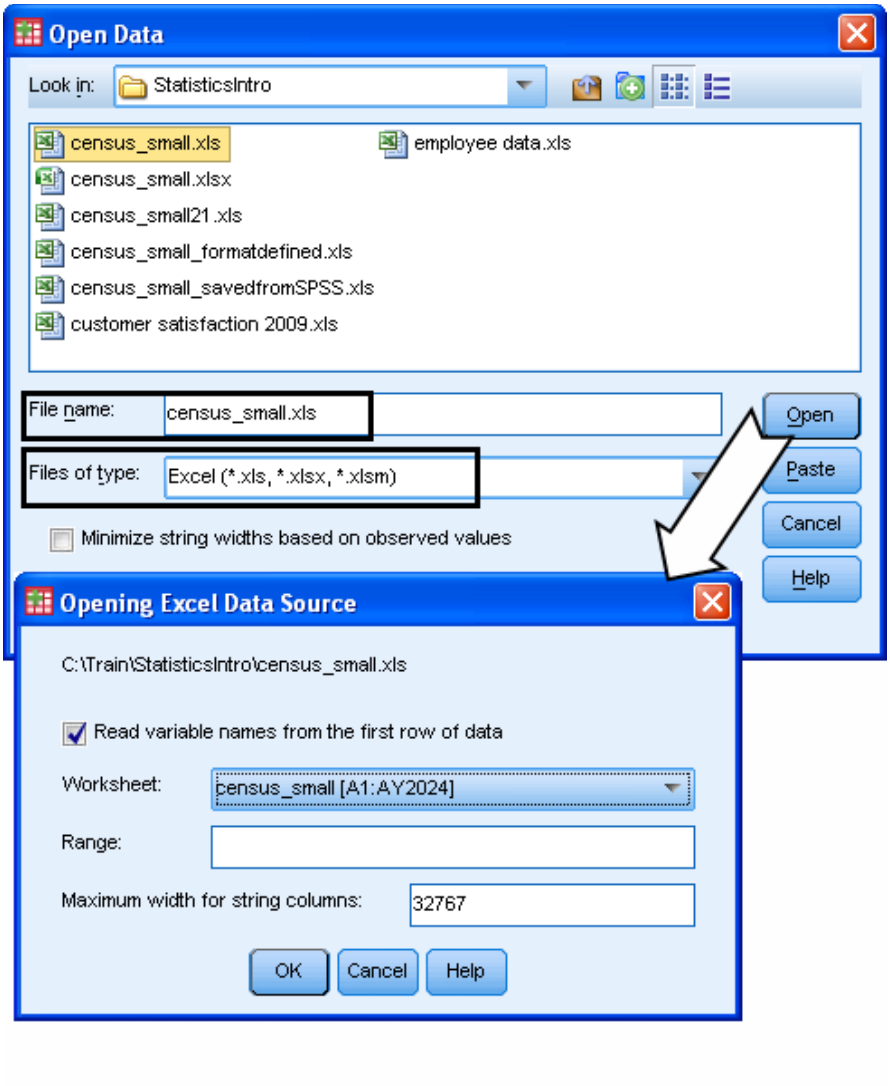
	A	B	C	D	E	F	G	H	I	J	K
1	id	sex	race	region	born	hh size	adults	age	age kdbrn	sibs	dwelling
2	1	1	3	1	1	2	1	49	20	0	APARTMENT-4 STORIES
3	2	1	3	1	2	1	1	48	33	1	APARTMENT-4 STORIES
4	3	1	2	1	1	2	2	47	22	1	APARTMENT-4 STORIES
5	4	1	2	1	1	3	3	32	26	2	APARTMENT-COMMERCIAL
6	5	2	2	1	2	3	1	37	32	2	APARTMENT-4 STORIES
7	6	1	2	1	1	2	2	72	25	3	ROW HOUSE
8	7	1	2	1	2	5	4	21		2	3-4 FAM HOUSE
9	8	2	2	1	2	5	3	36	21	6	ROW HOUSE
10	9	2	2	1	2	1	1	999	27	3	APARTMENT-4 STORIES
11	10	2	2	1	2	1	1	56	20	12	APARTMENT-4 STORIES
12	11	1	2	1	1	2	2	62	31	4	APARTMENT-4 STORIES
13	12	1	2	1	2	1	1	999		8	APARTMENT-4 STORIES
14	13	2	1	1	1	2	1	62	44	1	APARTMENT-COMMERCIAL
15	14	1	2	1	1	2	2	71	22	7	APARTMENT-4 STORIES
16	15	2	2	1	1	1	1	58		2	APARTMENT-4 STORIES
17	16	2	3	1	2	6	2	34	27	2	ROW HOUSE
18	17	1	1	1	2	2	2	26		3	APARTMENT-4 STORIES
19	18	2	1	1	1	6	2	33	20	8	DETACHED 1-FAM HOUSE
20	19	1	1	1	2	1	1	36		0	DETACHED 1-FAM HOUSE
21	20	1	3	1	2	4	2	999	30	2	3-4 FAM HOUSE

We will demonstrate how to open this file. The considerations and the operations for reading earlier or later versions of Excel files are the same as demonstrated here.

Detailed Steps for Reading an Excel File

1) From the **File** menu, select **Open** and **Data**.
2) Select Files of type: **Excel** and the *census_small.xls* file.
3) Accept all defaults in the **Opening Excel Data Source** dialog box.

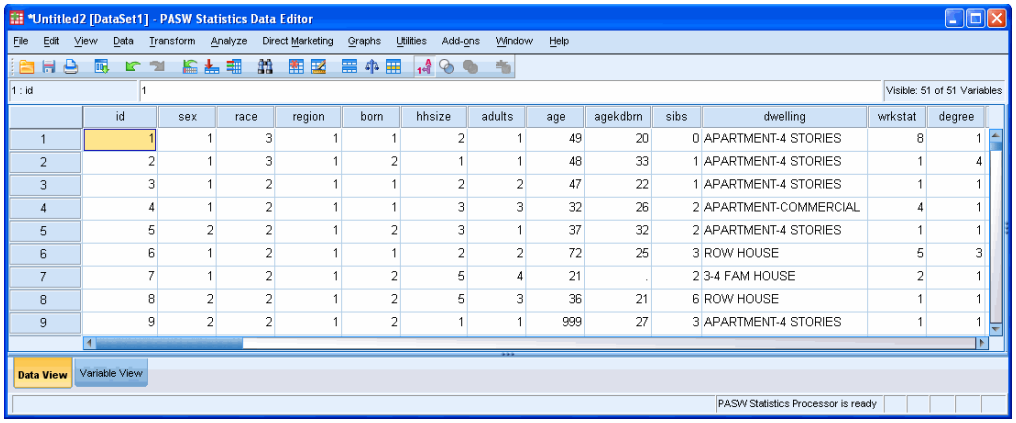
Figure 3.4 Open *census_small.xls* File



Results: Reading an Excel File

By default, a new Data Editor window with title "*Untitled 2[DataSet1]" is opened, displaying the data imported from the Excel file (the title can be different, if you have other datasets open when you import the Excel file).

Figure 3.5 Imported Excel File *census_sample.xls*



	id	sex	race	region	born	hhsize	adults	age	agekdbrn	sibs	dwelling	wrkstat	degree
1	1	1	3	1	1	2	1	49	20	0	APARTMENT-4 STORIES	8	1
2	2	1	3	1	2	1	1	48	33	1	APARTMENT-4 STORIES	1	4
3	3	1	2	1	1	2	2	47	22	1	APARTMENT-4 STORIES	1	1
4	4	1	2	1	1	3	3	32	26	2	APARTMENT-COMMERCIAL	4	1
5	5	2	2	1	2	3	1	37	32	2	APARTMENT-4 STORIES	1	1
6	6	1	2	1	1	2	2	72	25	3	ROW HOUSE	5	3
7	7	1	2	1	2	5	4	21	.	2	3-4 FAM HOUSE	2	1
8	8	2	2	1	2	5	3	36	21	6	ROW HOUSE	1	1
9	9	2	2	1	2	1	1	999	27	3	APARTMENT-4 STORIES	1	1

The variable names are read from the first row of the Excel spreadsheet. However, the blank spaces in the variable names, as in *hh size*, have been removed since blanks are not allowed in PASW Statistics variable names. Note that the string values of *dwelling* are imported without any problems.

Apply Your Knowledge: Reading an Excel File

For each of the following two statements, answer True or False.

1. An Excel 2007 file can be opened in PASW Statistics 18.

2. Lotus files can be read via **File…Open…Data**.

3. Which of the following statements is true? If a column name in the spreadsheet violates PASW Statistics variable naming rules:
    a. The import is aborted
    b. The column name will be replaced by **spreadsheetvar_unknown**
    c. The column name will be converted into a valid variable name
    d. The column will not be imported into PASW Statistics

3.7 *Reading Data from Databases*

Data from database sources such as Access, DB2, Oracle and other relational databases can be imported into PASW Statistics using the Open Database Wizard. Any database that uses ODBC (Open Database Connectivity) drivers can be read directly by PASW Statistics after the drivers are installed. ODBC drivers for many database formats are supplied with the PASW Statistics installation or can be downloaded from the SPSS, an IBM company website. Additional drivers can be obtained from third-party vendors.

Because a table in a database has the same structure as a PASW Statistics data file, there are no specific issues when reading data from a database. Field names in the database table which are not supported by PASW Statistics will be converted into valid PASW Statistics variable names.

One of the most common database applications, Microsoft Access, is discussed in this lesson.

Tip

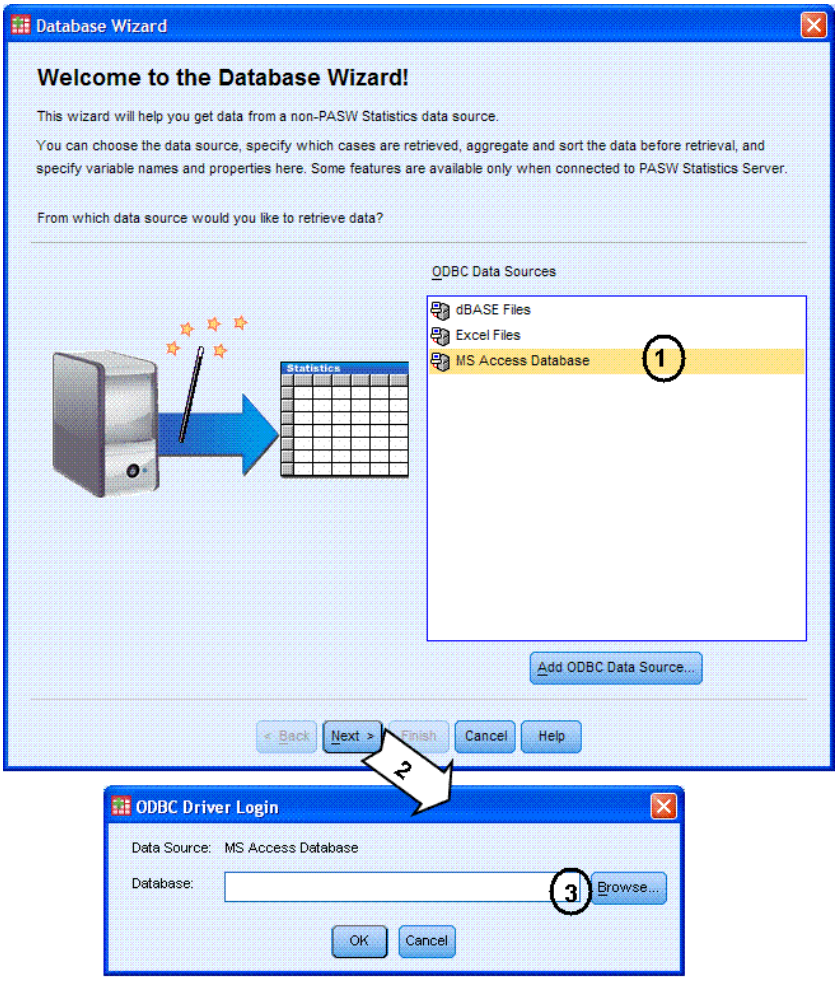
An Excel data file can be opened as an ODBC data source. If the Excel file contains date fields, this is often the preferred method rather than reading from the **File...Open...Data** menu.

3.8 *Procedure: Reading Data from Database*

The procedure for opening a table from a database is invoked from the **File…Open Database…New Query** menu.

The initial dialog box welcomes you to the Database Wizard. The ODBC data source specifying the type of database must be selected here (in this case: MS Access Database) and then the Database Wizard is entered. Move through the steps of the Wizard by selecting the **Next** button. In the second dialog box the database is selected.

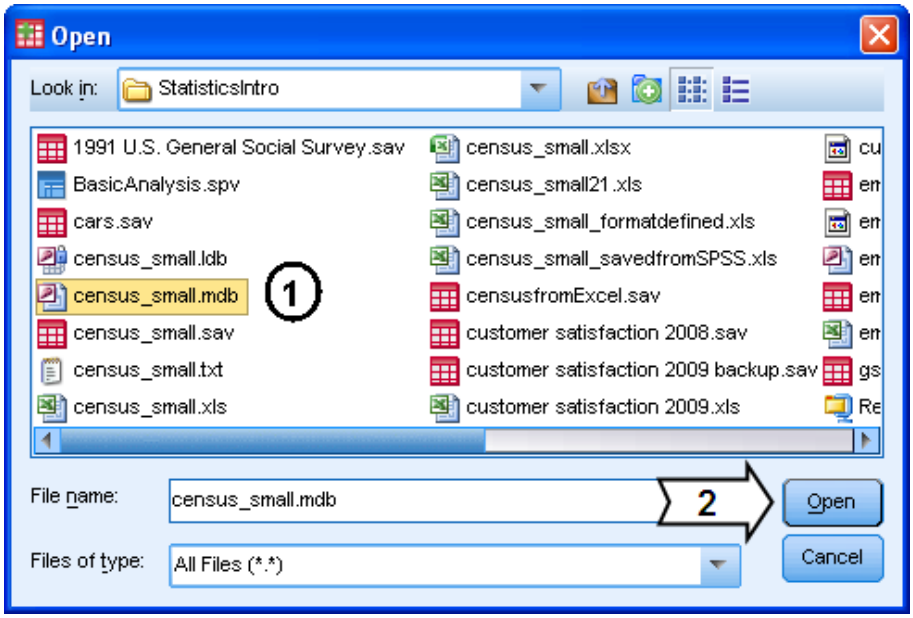
Figure 3.6 Database Wizard Welcome Dialog Box



1) Select the ODBC data source

2) Select **Next** button which opens the ODBC Driver Login
3) Select the **Browse** button to locate the database file. Some databases such as Oracle or SQL Server will require a user login in order to access the database.

Figure 3.7 Open Data (Database) Dialog Box



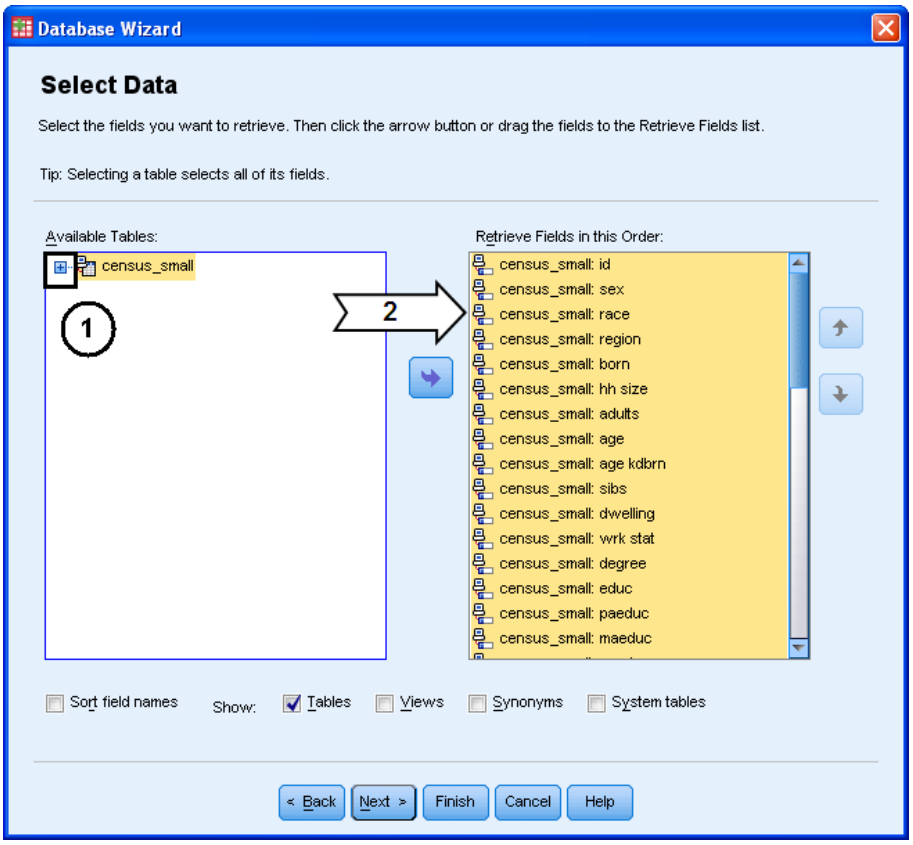
1) Select the Database file.
2) Select **Open** button and then **OK** button to move to the next Database Wizard step.

Note

If MS Access Database is not listed as an ODBC data source, you need to either download the Microsoft Access ODBC driver from the Microsoft website or install the SPSS Inc. Data Access Pack.

In the next step, you can specify the tables and fields that you want to import.

Figure 3.8 Database Wizard: Select Data Step



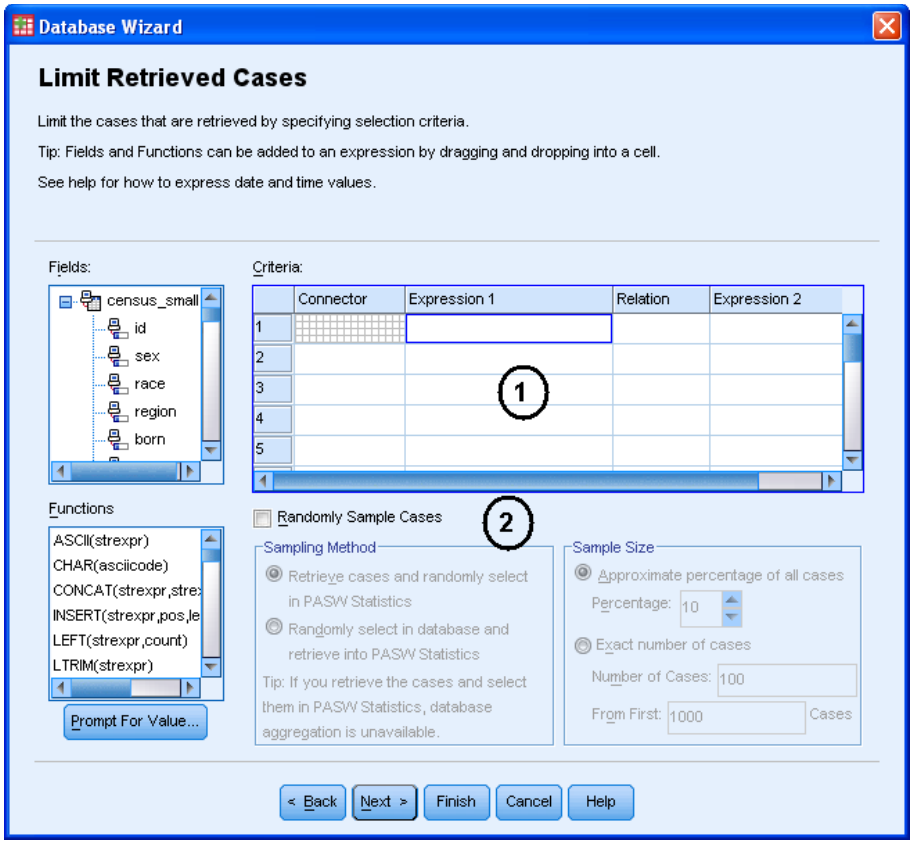
1) To see all of the fields in a table, you can click on the plus sign in front of the table name.
2) Select table and fields to import. You can select specific fields to import as variables or you can import the entire table.

Note

If you retrieve fields from multiple tables, the next step will ask to specify the relationship between the tables. Specify the key field(s) on each table whose values are used to link the records.

In the next step the number of cases that are imported can be limited. For large data sources, you may want to limit the number of cases to a small, representative sample to reduce the processing time. Or, you may want to select a subset of cases based on some logical criteria; for example, only customers who purchased after a specified date. This step is optional and if nothing is specified, all cases in the selected table will be imported.

Figure 3.9 Database Wizard: Limit Retrieved Cases Step



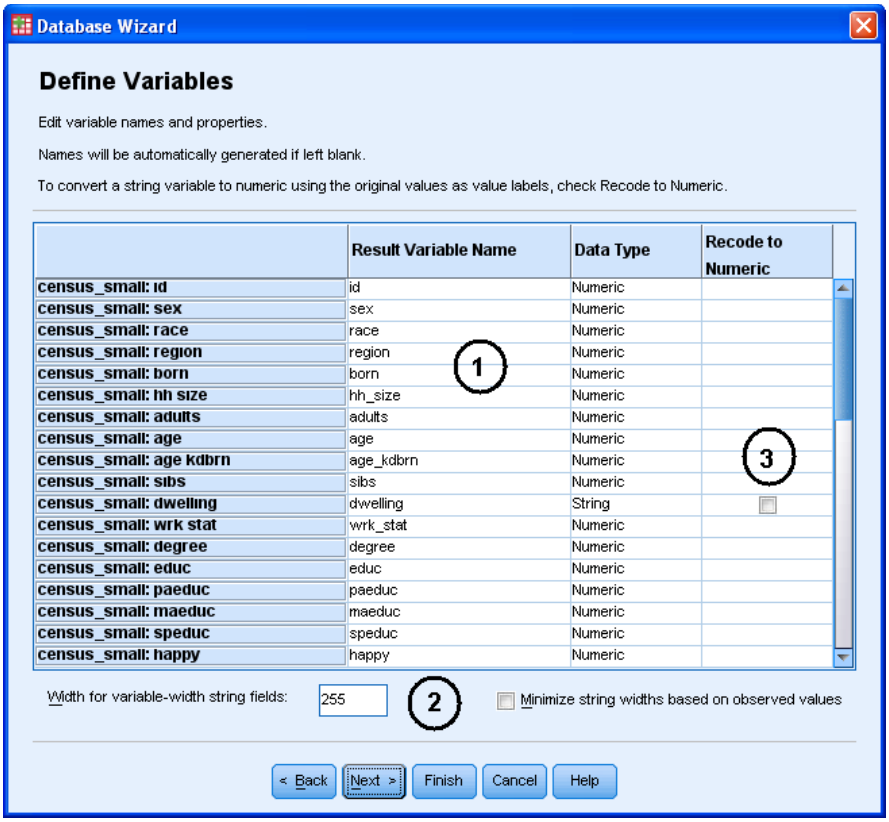
1) Build SQL criteria to specify a subset of cases to import (for example, males older than 30 years).
2) Select *Randomly Sample Cases* to select a random sample of cases from the data source.

Note

For those using PASW Statistics Server: If you are in distributed mode, connected to a remote server with PASW Statistics Server, the next step provides the facility to aggregate the data before importing into PASW Statistics. The facility to aggregate data is also available in PASW Statistics, but pre-aggregating may save time for large data sources.

In the next step, variable names can be edited, length of string variables can be set or string variables converted to numeric variables.

Figure 3.10 Database Wizard: Define Variables Step



1) Edit the variable names in the *Result Variable Name* column. Field names in the database table are automatically converted to PASW Statistics variable names. If necessary, the field names are converted to valid variable names, and the original field names are preserved as variable labels.
2) Width of string variables can be specified. By default the width of a string variable is set to 255. Or, select *Minimize string widths based on observed values* to base the width of all string variables on the observed data values.
3) String (alphanumeric) fields can be recoded to numeric variables. This option converts string variables to numeric variables with integer values and retains the original string values as the value labels for the new variable.

Tip

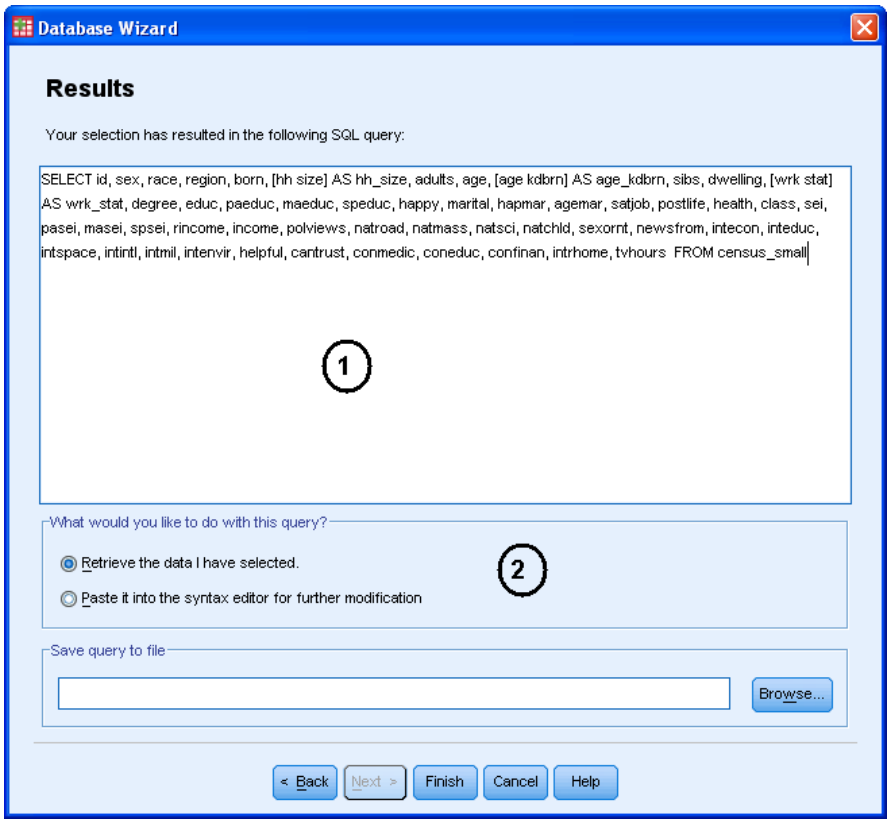
The recode of string variables to numeric variables is actually done in PASW Statistics, so you can wait to do the conversion after you have imported the data. See the *Creating Categorical Variables* lesson, topic *Automatic Recode* for a discussion.

Note

For those using PASW Statistics Server: If you are in distributed mode, connected to a remote server with PASW Statistics Server, the next step provides the facility to sort the data before reading it into PASW Statistics. You can also sort data after reading it into PASW Statistics, but presorting may save time for large data sources.

In the next step the SQL statement created from your selections in the Database Wizard appears. This statement can be executed now, pasted as PASW Statistics syntax, or saved as a query to a file for later use or modification. You will find these two latter options useful if you need to perform the database retrieval routinely every month, for example. See the lesson *PASW Statistics Syntax* for a discussion of syntax.

Figure 3.11 Database Wizard: Results Step

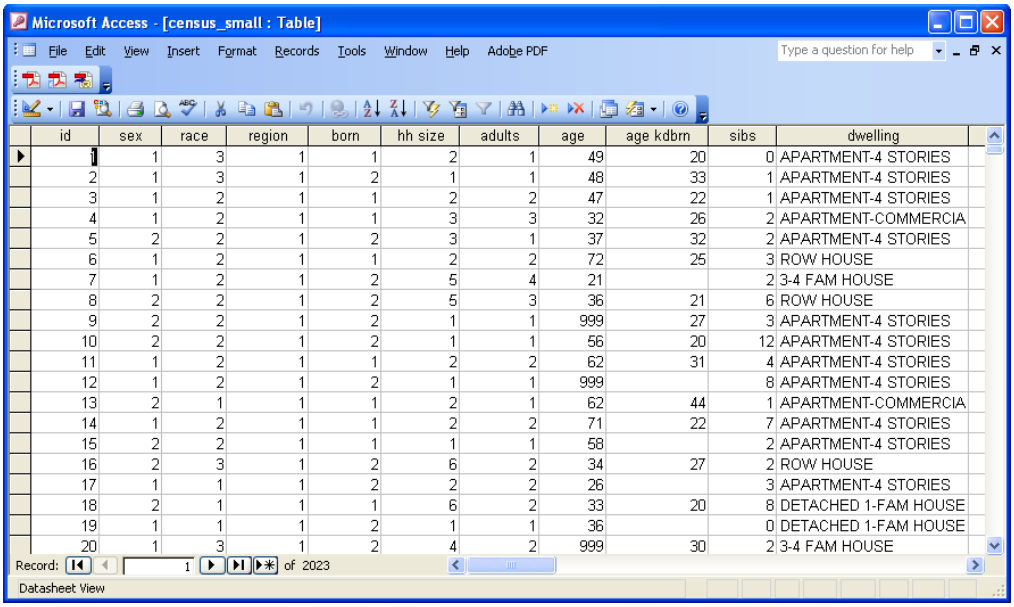


1) The SQL statement created from the selections in the Database Wizard is displayed.
2) Select to retrieve the data now, save as a query file, or paste as PASW Statistics syntax. The later two options are useful if you need to perform the database retrieval routinely, for example every month.

3.9 *Demonstration: Reading Data from Access Database*

We will read the table *census_small* from the *census.mdb* Access database. The figure below shows the data in Access. In this example, we will retrieve all fields from the table and all records.

Figure 3.12 *Census_Small.mdb* Database in Access



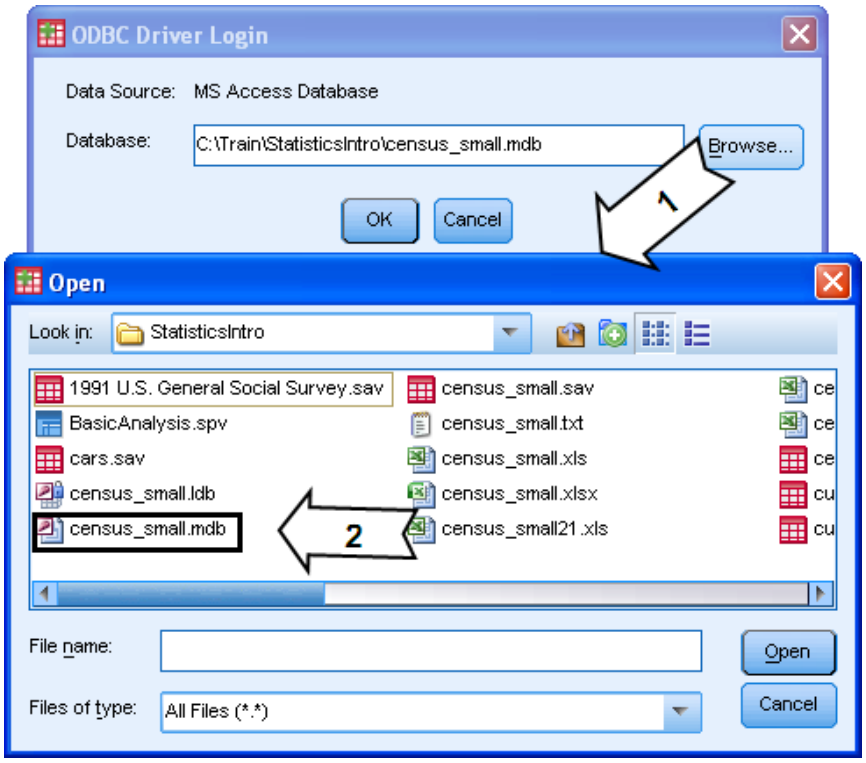
id	sex	race	region	born	hh size	adults	age	age kdbrn	sibs	dwelling
1	1	3	1	1	2	1	49	20	0	APARTMENT-4 STORIES
2	1	3	1	2	1	1	48	33	1	APARTMENT-4 STORIES
3	1	2	1	1	2	2	47	22	1	APARTMENT-4 STORIES
4	1	2	1	1	3	3	32	26	2	APARTMENT-COMMERCIA
5	2	2	1	2	3	1	37	32	2	APARTMENT-4 STORIES
6	1	2	1	1	2	2	72	25	3	ROW HOUSE
7	1	2	1	2	5	4	21		2	3-4 FAM HOUSE
8	2	2	1	2	5	3	36	21	6	ROW HOUSE
9	2	2	1	2	1	1	999	27	3	APARTMENT-4 STORIES
10	2	2	1	2	1	1	56	20	12	APARTMENT-4 STORIES
11	1	2	1	1	2	2	62	31	4	APARTMENT-4 STORIES
12	1	2	1	2	1	1	999		8	APARTMENT-4 STORIES
13	2	1	1	1	2	1	62	44	1	APARTMENT-COMMERCIA
14	1	2	1	1	2	2	71	22	7	APARTMENT-4 STORIES
15	2	2	1	1	1	1	58		2	APARTMENT-4 STORIES
16	2	3	1	2	6	2	34	27	2	ROW HOUSE
17	1	1	1	2	2	2	26		3	APARTMENT-4 STORIES
18	2	1	1	1	6	2	33	20	8	DETACHED 1-FAM HOUSE
19	1	1	1	2	1	1	36		0	DETACHED 1-FAM HOUSE
20	1	3	1	2	4	2	999	30	2	3-4 FAM HOUSE

Detailed Steps for Reading Data from Access Database

To import the Access database into PASW Statistics:

1) From the **File** menu, select **Open Database** and **New Query**
2) Select **MS Access Database** as the ODBC Data Source and **Next** button to open the ODBC Drive Login dialog box.
3) Select *census_small.mdb* database
4) Place all fields (or the entire table) in the *Retrieve Fields in this Order* box

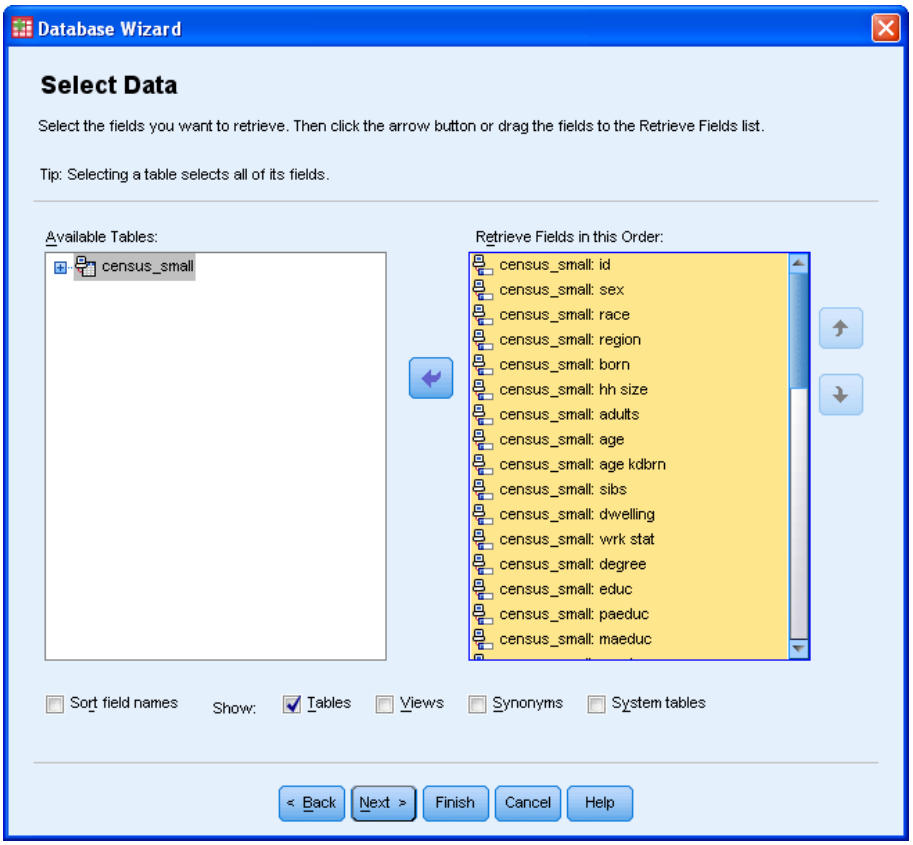
Figure 3.13 Select the *census_small.mdb* Access Database



1) Select **Browse** button
2) Select *census_small.mdb*

Next, all tables contained in the database are listed; here the database file has only one table: *census_small*.

Figure 3.14 Select Table and Fields



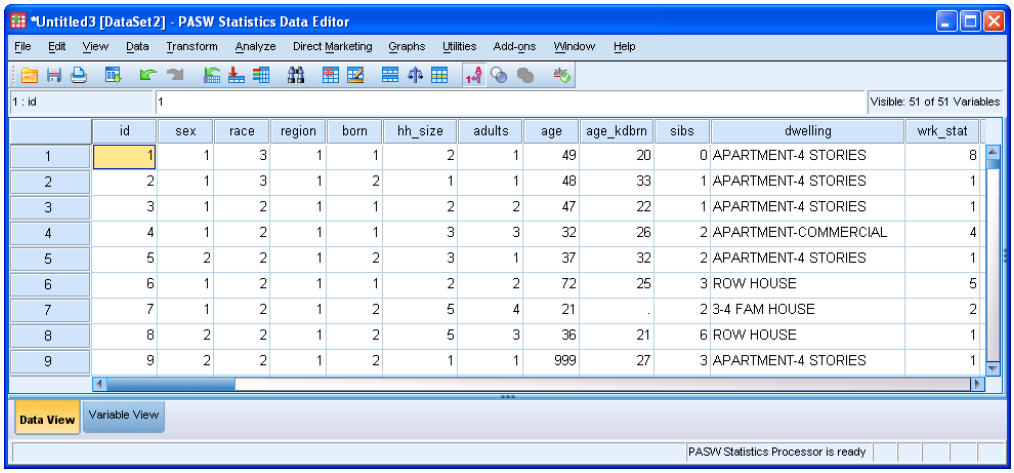
In this example we will retrieve all fields. Select the table and drag to the *Retrieve Fields* box.

At this point we can step through the Database Wizard, or we can decide to finish. Because we do not want to limit the number of cases retrieved, and so on, we just select the **Finish** button.

Results: Reading Data from Access

By default, a new Data Editor window is opened, containing the data imported from the database.

Figure 3.15 Imported Access Database



	id	sex	race	region	born	hh_size	adults	age	age_kdbrn	sibs	dwelling	wrk_stat
1	1	1	3	1	1	2	1	49	20	0	APARTMENT-4 STORIES	8
2	2	1	3	1	2	1	1	48	33	1	APARTMENT-4 STORIES	1
3	3	1	2	1	1	2	2	47	22	1	APARTMENT-4 STORIES	1
4	4	1	2	1	1	3	3	32	26	2	APARTMENT-COMMERCIAL	4
5	5	2	2	1	2	3	1	37	32	2	APARTMENT-4 STORIES	1
6	6	1	2	1	1	2	2	72	25	3	ROW HOUSE	5
7	7	1	2	1	2	5	4	21	.	2	3-4 FAM HOUSE	2
8	8	2	2	1	2	5	3	36	21	6	ROW HOUSE	1
9	9	2	2	1	2	1	1	999	27	3	APARTMENT-4 STORIES	1

Field names that did not conform to PASW Statistics variable-naming rules are converted into valid variable names and the original column headings are saved as variable labels.

Apply Your Knowledge: Reading Data from Access

For each of the following statements, answer True or False.

1. A table from an Access database can be opened via **File…Open…Data** menu and then select Files of Type: Access.

2. A table stored in a DB2 database can be imported into PASW Statistics 18 if there is an appropriate ODBC driver for this database installed.

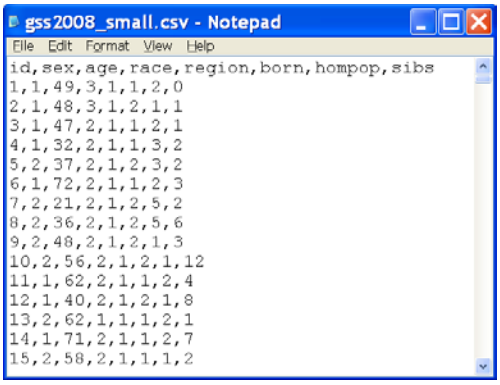
3. Multiple tables from the same database can be imported into PASW Statistics.

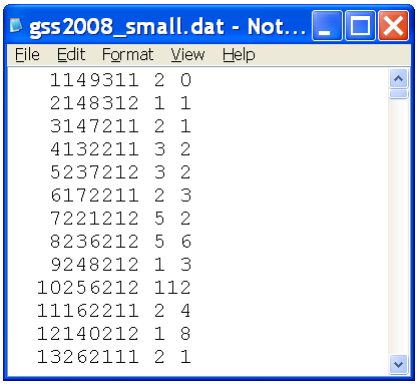
3.10 *Reading a Text File*

Text or ASCII files are another common source of data. Many spreadsheet programs and databases can save their contents in text file formats. Text files can be either Fixed width format or Delimited format. Fixed width files refer to rows of data with each variable's value in the same column position in all rows. Delimited files refer to rows of data that use a standard character to separate each variable. Commonly used delimiter characters are tabs or commas.

The figure below shows a comma delimited text file (left) and a fixed width text file (right). Notice that the comma separated file has variable names in the first row.

Figure 3.16 Text File Formats: Comma Delimited File and Fixed Width File



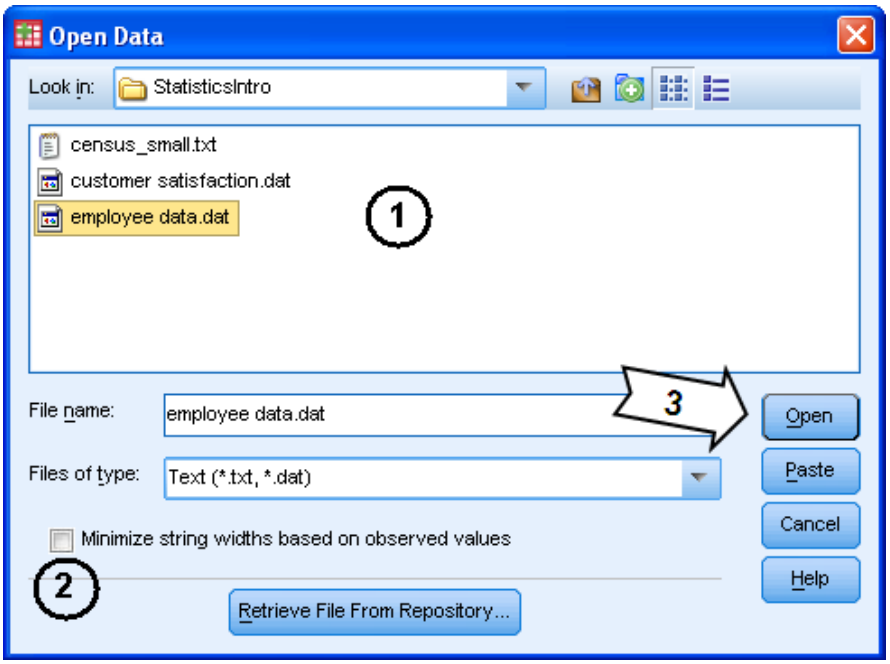


3.11 *Procedure: Reading a Delimited Text File*

In this lesson we will discuss how to open a delimited text file.

The procedure for opening a text file is invoked from the **File…Read Text Data** menu.

Figure 3.17 Open Data (Text File) Dialog Box



1) Select the text file.
2) Optionally, select *Minimize string widths based on observed values* to set the width of all string variables based on the observed values.
3) Select **Open** button to open the Text Import Wizard.

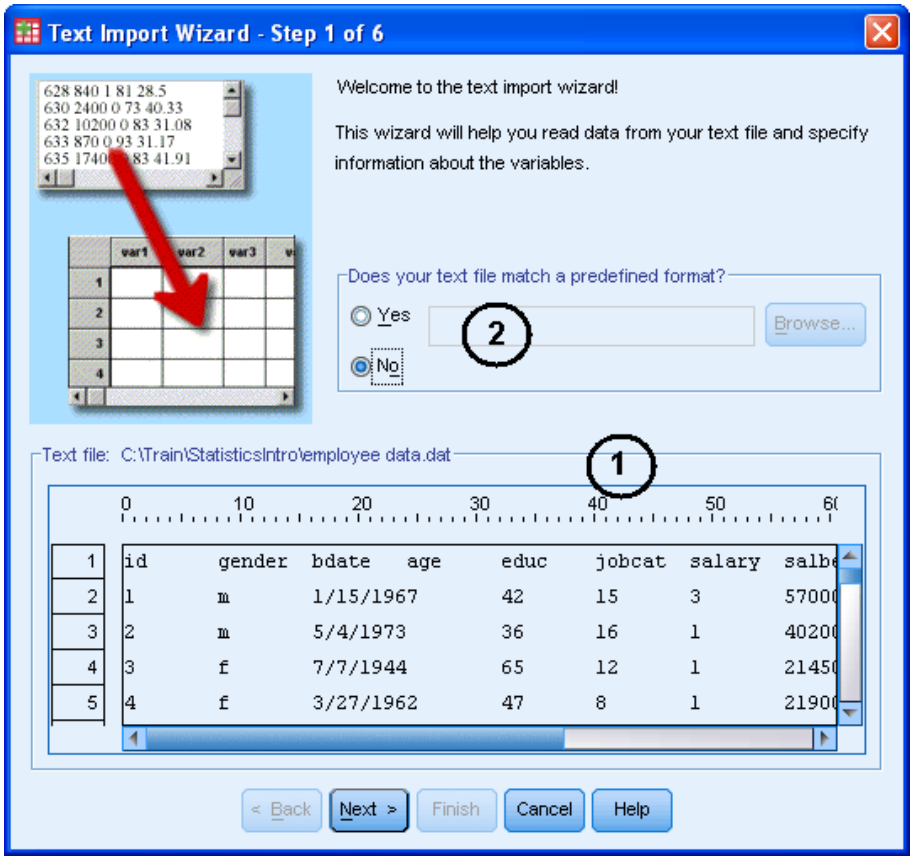
Further

If the text file has a file extension other than *.txt* or *.dat*, select **All Files (*.*)** in the Files of type: dropdown list. This is necessary if you use the *.csv* file extension for a comma delimited file such as those saved from Excel.

The Text Import Wizard guides you through the process of defining how the specified text file should be read.

In the first step, the first few rows of data in the text file are previewed. You can also specify to use a previously saved format. This is a useful feature if you periodically receive text data in the same format. Examples are an ongoing customer or employee survey in which data are collected every quarter, or any type of sales or transaction data collected monthly.

Figure 3.18 Text Import Wizard – Step 1 of 6

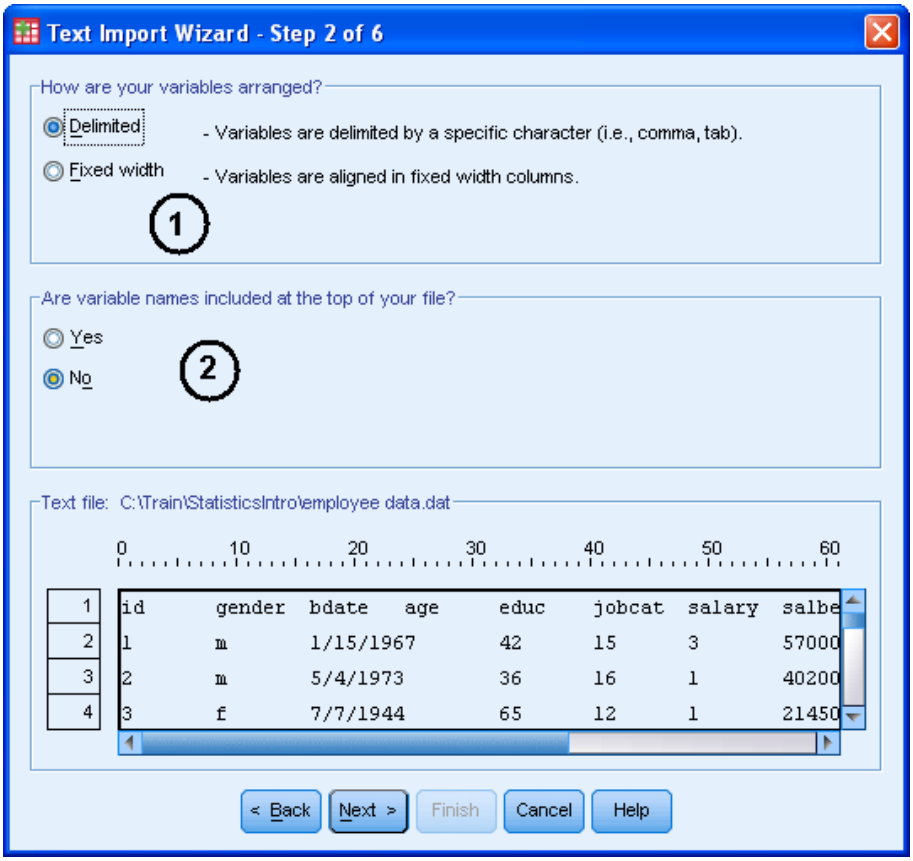


1) The Data Preview area confirms that name of the text file and shows the first few rows of the file.
2) Select whether to use a predefined format (Yes/No). If a predefined format is used, select the **Browse** button to specify the file containing the predefined format.

Move through the steps of the Wizard by selecting the **Next** button.

In the second step, select the format of the text file and indicate whether variable names are recorded in the first row of the file. Delimited files commonly have variable names recorded in the first row of the file.

Figure 3.19 Text Import Wizard – Step 2 of 6

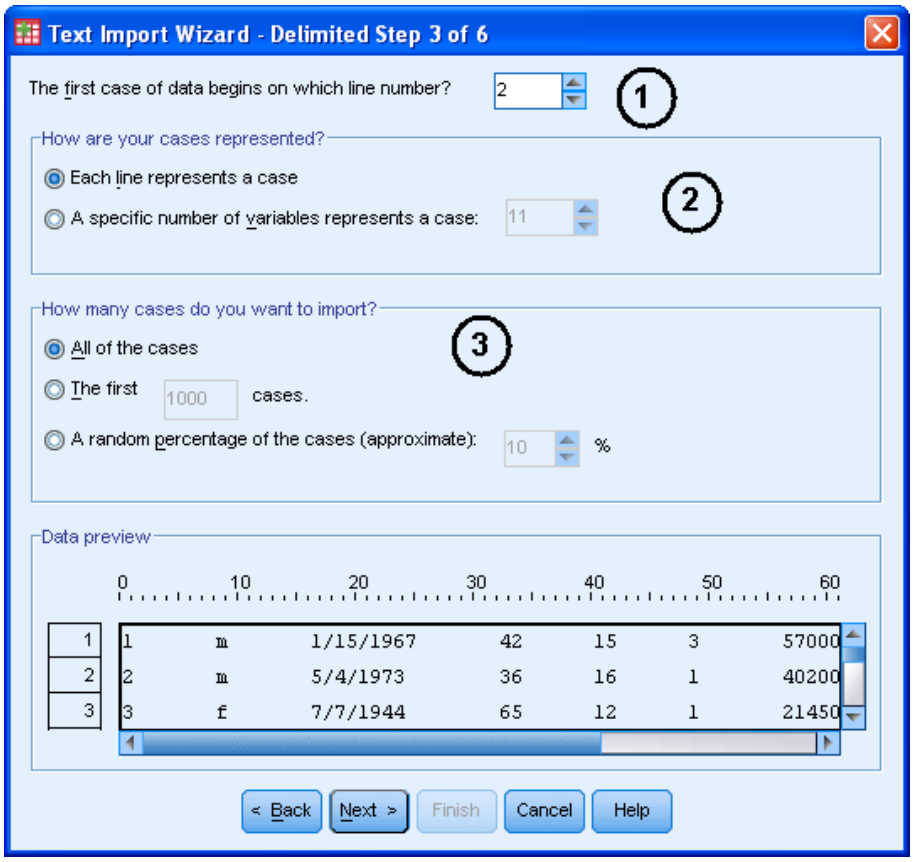


1) Select how the variables are arranged either *Delimited* file format or *Fixed width* format.
2) Indicate (Yes/No) whether the variable names are the first row of the text file. Examine the Data Preview section to determine this specification.

The following steps will differ depending on whether you select delimited or fixed width format. Here we go through the steps for reading a delimited text file.

In Step 3, specify which line (row) contains the first case of data. PASW Statistics automatically changes the default to line number 2 when you specify in the previous step that the first line contains variable names. By default, PASW Statistics reads all remaining rows as data with each row representing a case. You can specify a specific number of variables to represent each case. This is not often required. Finally, you have the choice to read only selected lines of data. By default, all other lines are read as data; but, you could choose to read only the first few cases or a random selection of cases.

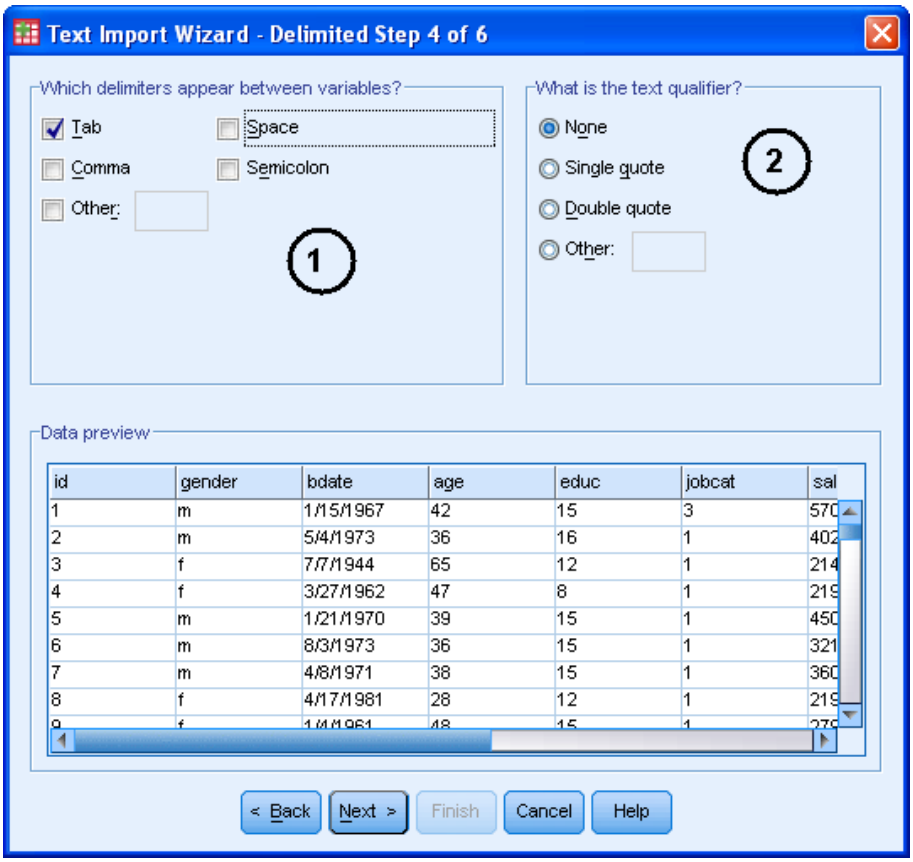
Figure 3.20 Text Import Wizard – Delimited Step 3 of 6



1) Specify the row containing the first case of data which automatically adjusted to row 2 if you specify variable names in the first row in the previous step.
2) Optionally, specify a specific number of variables to represent each case. The default, each line represents a case, is most common.
3) Optionally, you can select to import the first specified number of cases or a random specified percentage of cases. The default is to import all cases.

The Data preview in Step 4 provides a quick way to ensure that the data are being properly read by PASW Statistics. You can change the specification of delimiter characters and indicate if alphanumeric or string text has been entered with qualifiers such as quotes surrounding the text. Use the Data preview pane to examine the effect of any changes you make to these specifications.

Figure 3.21 Text Import Wizard – Delimited Step 4 of 6



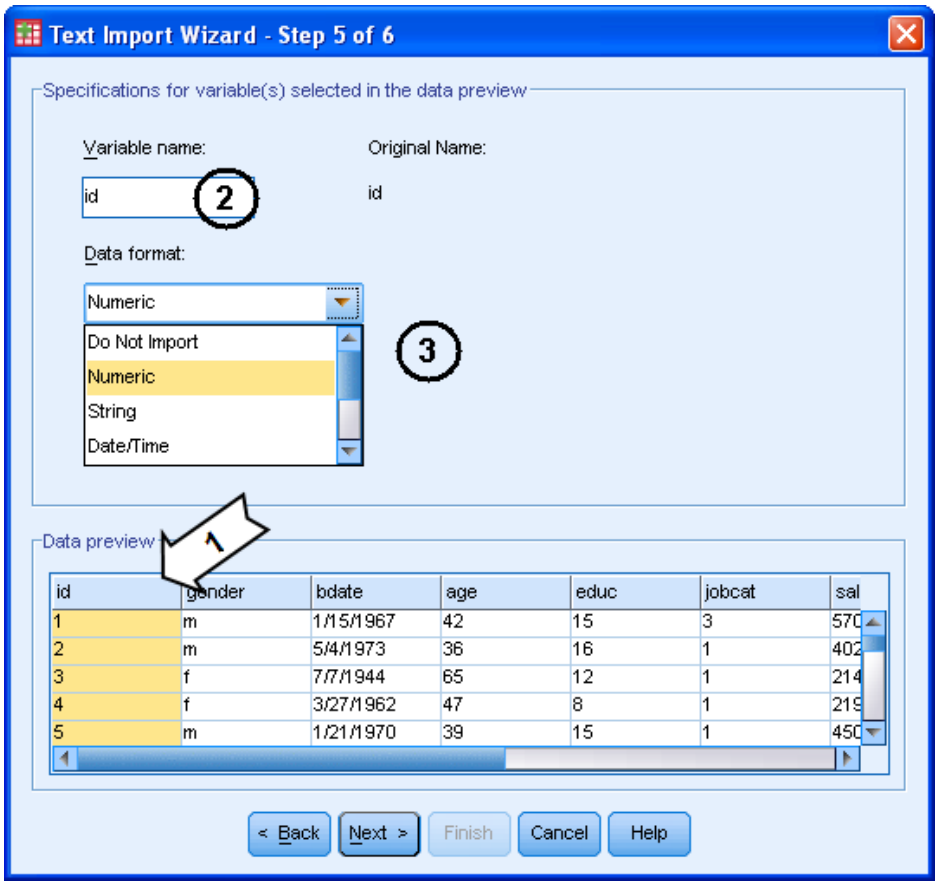
1) Select the delimiter character(s).
2) Specify the text qualifier character if used with the string or alphanumeric text.

Tip

The default delimiter characters may include the *Space* or another character that is not a valid delimiter in the text file. Examine the Data Preview carefully to determine if the default delimiter characters are correct and make any necessary changes.

In Step 5, you can change variable names and data formats for the variables as well as identify any variables that you do not want to import.

Figure 3.22 Text Import Wizard – Delimited Step 5 of 6



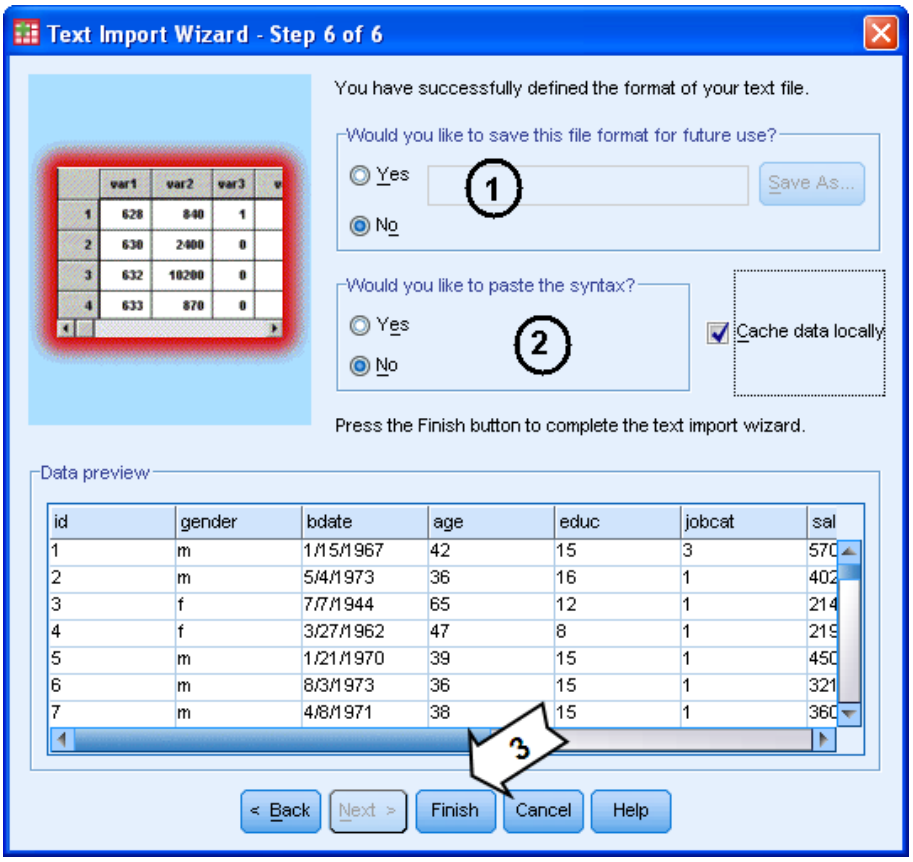
1) Highlight the variable to change in the Data Preview.
2) Type a new variable name into the Variable Name text box.
3) Change the Data format or select *Do not import* from the dropdown list to not import the selected variable.

Warning

Changing the data format can result in lost data values. For example, changing the data format for a string variable *gender* with values "f" and "m" into numeric will cause all data values to be missing.

In the sixth and last step, provide options to save a predefined format or to paste the syntax. These options provide ways of retaining the specifications such as variable names and data formats for later use with the same or other data files. The options differ only in the format in which this information is saved. The first option (save the format) will store the information in a format that can be retrieved, displayed and used in the Text Import Wizard, while the syntax option will produce a PASW Statistics command that will run independently of the Text Import Wizard. See the lesson *PASW Statistics Syntax*.

Figure 3.23 Text Import Wizard – Delimited Step 6 of 6



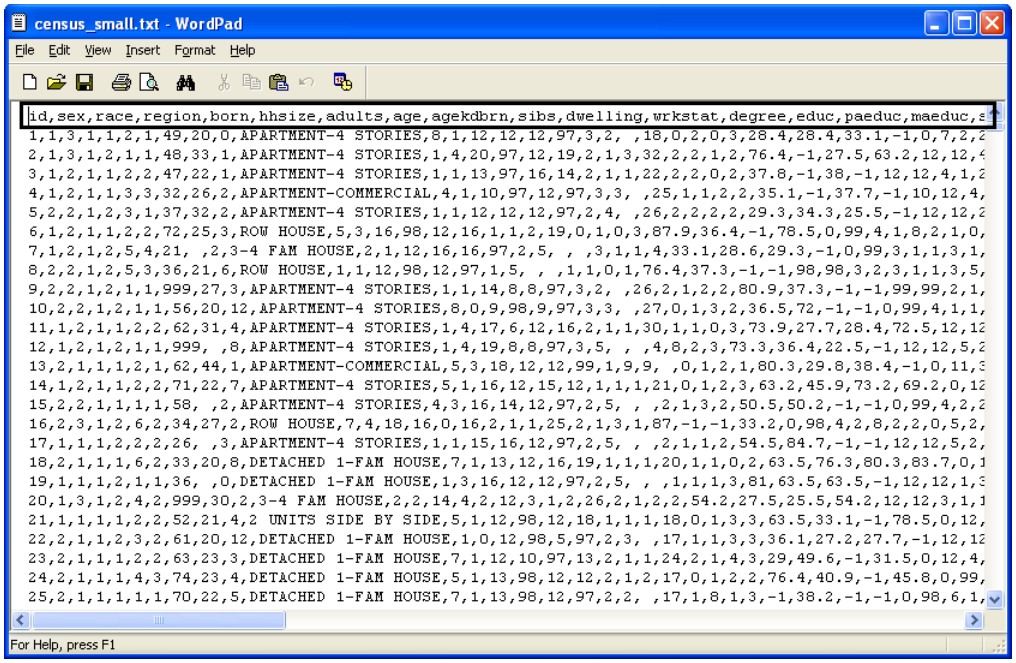
1) Save a file format for later use? Select *Yes* or *No*.
2) Paste the syntax command? Select *Yes* or *No*.
3) Select **Finish** to complete the specification.

If neither of these options is selected, the data are read immediately when the Finish button is selected, and the data are displayed in a new Data Editor window.

3.12 *Demonstration: Reading a Delimited Text File*

We will read the text file *census_small.txt*, a comma-separated text file, with variable names in the first row.

Figure 3.24 *Census_small.txt* Text Data File



Notice the variable names in the first row and the string values containing blank spaces in the *dwelling* variable.

We will use the Text Wizard to read this text file.

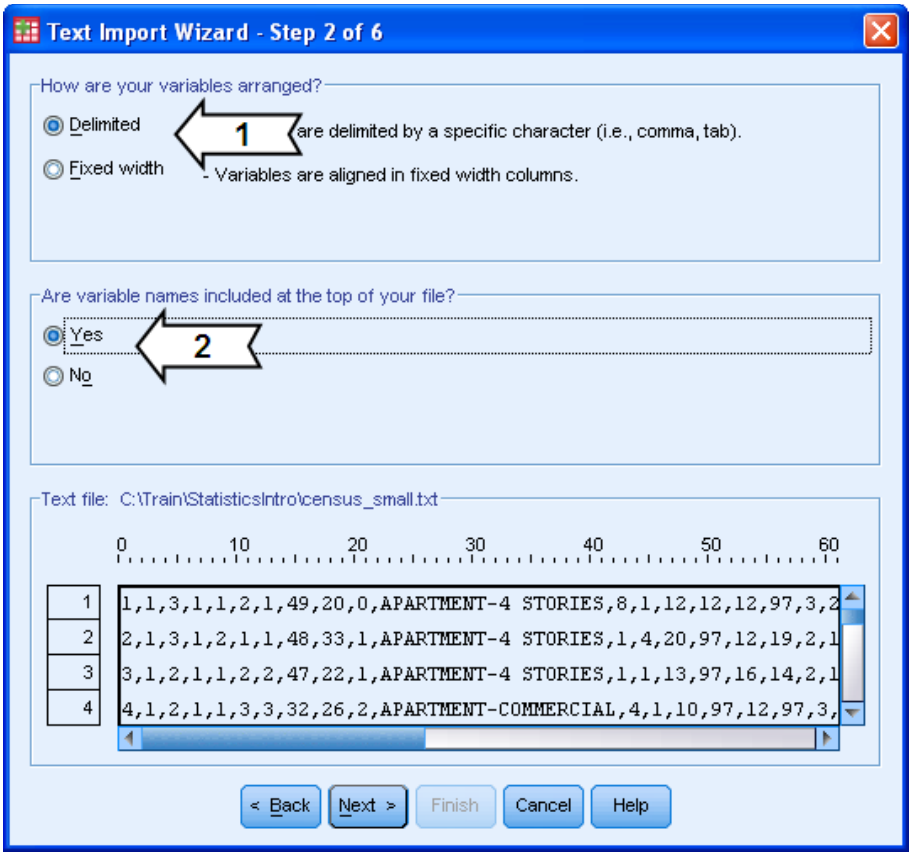
Detailed Steps for Reading a Delimited Text File

To read this comma-delimited file:

1) From the **File** menu, select **Read Text Data** and select the *census_sample.txt* file in the Open Data dialog box
2) In step 1 of the Text Wizard, accept the default of **No** for a predefined file format
3) In step 2, select the **Delimited** text format and select **Yes** that variable names are included at the top of the file.
4) In step 3, accept the defaults to read the first case from row 2
5) In step 4, deselect **Space** as a delimiter
6) In step 5, accept all defaults after examining the Data Preview to assure that the data are being read correctly now
7) In step 6, accept the defaults and do not save the file format or paste the syntax.
8) Select **Finish** to complete the specifications and read the data

The figure below shows the completed Step 2 dialog box.

Figure 3.25 Read census.txt file – Step 2

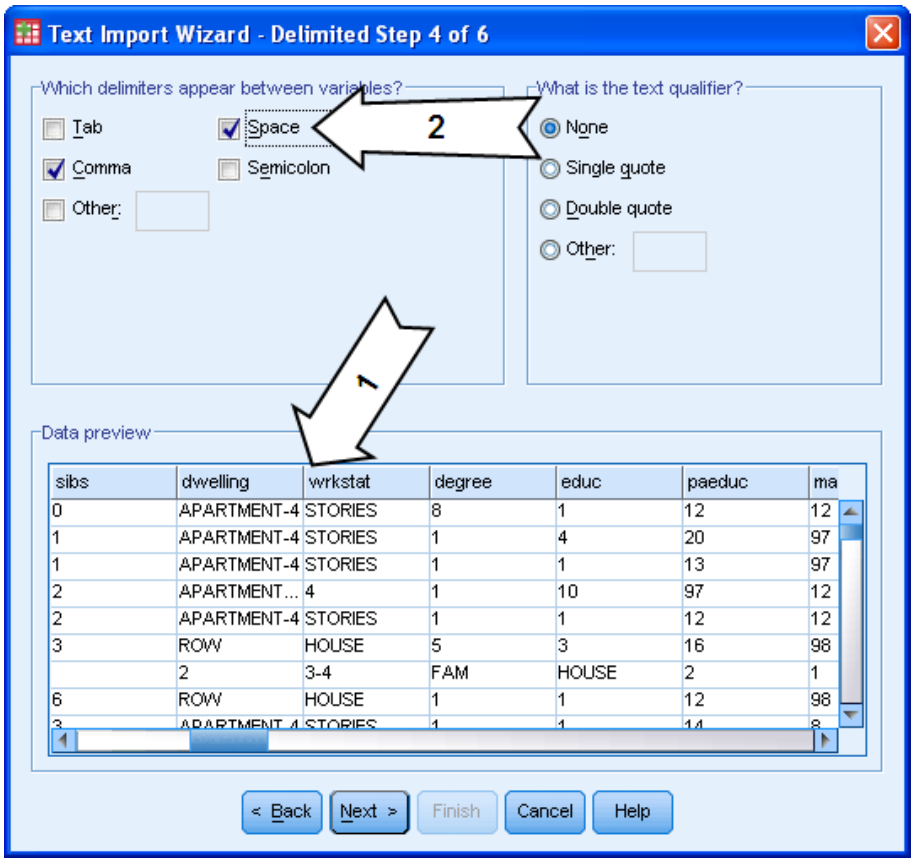


In step 2,

1) Select the **Delimited** text format.
2) Select **Yes** that variable names are included at the top of the file.

The Data Preview is automatically adjusted to now show row 2 in the file as the first data case.
In step 4 of the Text Import Wizard, the delimiters deserve special attention. Some of the values for string variable *dwelling* contain blank spaces, for example for value “APARTMENT-4 STORIES” in the first three cases. If space is used as delimiter, this would result in more than one variable for this field and misalignment of the variable names and data. This can be seen by examining the Data Preview.

Figure 3.26 Read census.txt file – Step 4 Data Preview with Space as Delimiter

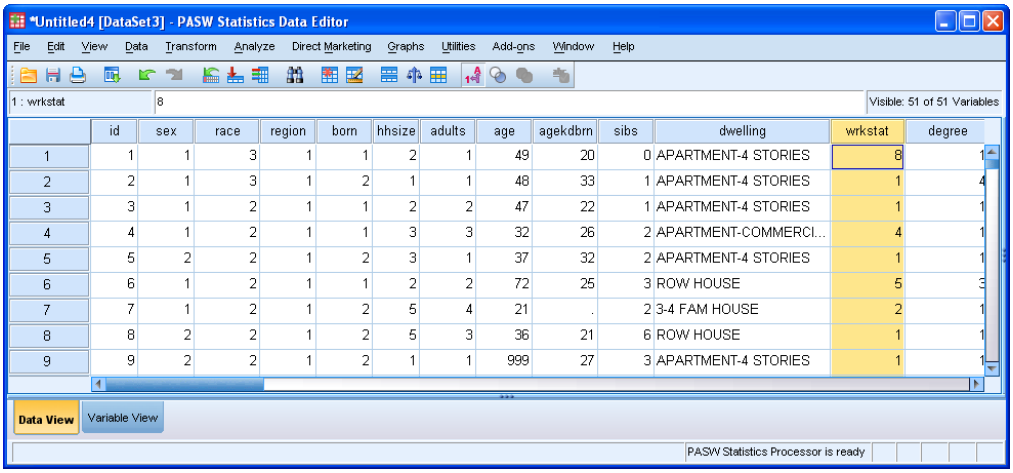


1) Examine the Data Preview.
2) Deselect **Space** as a delimiter.

Results: Reading a Text File

By default, a new Data Editor window is opened containing the data imported from the text file and we can verify that the data were read correctly. The names in the first row have been used as variable names, the remaining rows are read as data cases, and the variable *dwelling* values were read correctly.

Figure 3.27 Imported Text Data



	id	sex	race	region	born	hhsize	adults	age	agekdbrn	sibs	dwelling	wrkstat	degree
1	1	1	3	1	1	2	1	49	20	0	APARTMENT-4 STORIES	8	1
2	2	1	3	1	2	1	1	48	33	1	APARTMENT-4 STORIES	1	4
3	3	1	2	1	1	2	2	47	22	1	APARTMENT-4 STORIES	1	1
4	4	1	2	1	1	3	3	32	26	2	APARTMENT-COMMERCI...	4	1
5	5	2	2	1	2	3	1	37	32	2	APARTMENT-4 STORIES	1	1
6	6	1	2	1	1	2	2	72	25	3	ROW HOUSE	5	3
7	7	1	2	1	2	5	4	21	.	2	3-4 FAM HOUSE	2	1
8	8	2	2	1	2	5	3	36	21	6	ROW HOUSE	1	1
9	9	2	2	1	2	1	1	999	27	3	APARTMENT-4 STORIES	1	1

Apply Your Knowledge: Reading a Text File

1. Suppose a comment (with values such as *In general I am satisfied*) is imported from a comma separated text file. The delimiter should then be set to
    a. Comma and Space
    b. Comma only
    c. Space only

2. PASW Statistics can read text files in which of the following formats? (Answer all that apply)
    a. Fixed width format with fields in the same columns for all cases
    b. Comma delimited files with the file extension *.csv*
    c. Delimited text files with variable names in the last row of the file
    d. Delimited text files using both comma and tab as delimiters

3. All fields must be read from a text file. (True or False?)

3.13 *Lesson Summary*

In this lesson we described how to use the File menu to import data from different file formats. We demonstrated how to read Excel, Access, comma-delimited text files

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Import data from different types of file formats

And, they should also be able to:

- Describe choices on the File menu to read and save data files
- Read Excel files
- Read from an Access database
- Read delimited text files

3.14 *Learning Activity*

Supporting Materials

Excel, Access, and text file versions of the Employee data, *employee data.xls*, *employee data.mdb*, *employee data.dat*. This data contains salary and job information on employees of a bank.

The tab-delimited text file *customer satisfaction 2009.dat* which contains information on customer purchasing behavior and satisfaction is also used.

1. **Spreadsheet data:** A copy of the employee data is stored as a worksheet in an Excel file *employee data.xls.* Variable names are stored in the first row of each of this file. Read this file into PASW Statistics.

2. **Database data:** A copy of the employee data is stored as a table within an Access database named *employee data.mdb.* Read all the fields in this table into PASW Statistics. (Note that an ODBC Access database driver must be installed on your computer in order to complete this exercise.)

3. **Tab-delimited data:** A copy of the employee data stored as a tab-delimited text file is contained in *employee data.dat.* Variable names are stored in the first row. Read this file into PASW Statistics with the Text Import Wizard.

For those with extra time:

4. Read data from the tab-delimited text file *customer satisfaction 2009.dat*. This file has a comment field, *reason_cancel* that contains spaces and other special characters in the text. Review the data carefully to make sure that this field is read correctly.

5. Create a new data set by importing *employee data.mdb*. In the DataBase Wizard, Limit Retrieved Cases by selecting only those cases whose *gender = 'm'* (i.e. Males). (Note: Be sure to enclose the value 'm' in single quotes.) How many cases were retrieved? How did you find out this number?

Lesson 4: Variable Properties

4.1 *Objectives*

After completing this lesson students will be able to:

- Define, save and view variable properties

To support the achievement of this primary objective, students will be able to

- Describe all of the variable properties
- Define variable properties in the Variable View window
- Define variable properties using the Define Variable Properties dialog
- Save variable properties with data in a PASW Statistics data file
- View variable properties interactively using Variables Utility
- View variable properties in tables using Display Data Dictionary and Codebook procedure

4.2 *Introduction*

Variable properties such as labels and display formats can be associated with each variable in the data. These variable properties can be saved with the data values and are used when analyzing the data and displaying charts and tables. This lesson discusses these variable properties and how you can define, view, and save them in a PASW Statistics data file.

Business Context

Data values often are not alphanumeric. For example, the variable *gender* might contain data values 1 (representing “male”) and 2 (representing “female”) rather than the values “male” and “female”. Although this is a very efficient way of entering the values, more meaningful labels are desired on tables and charts. For example in a Frequencies table the labels “male” and “female” should be displayed and not the values 1 and 2.

Another data coding option used especially in survey research is to enter special values to represent types of missing information. For example, the researcher might code “No Answer” as value 99 and “Don’t Know” as “88” for a question asking a respondent’s attitude about some policy. In the analyses these values should be treated as non-valid or missing values, e.g. statistics such as the mean of a variable should exclude these user-defined missing values in computations.
These and other variable properties can be defined and saved with your data values.

Supporting Materials

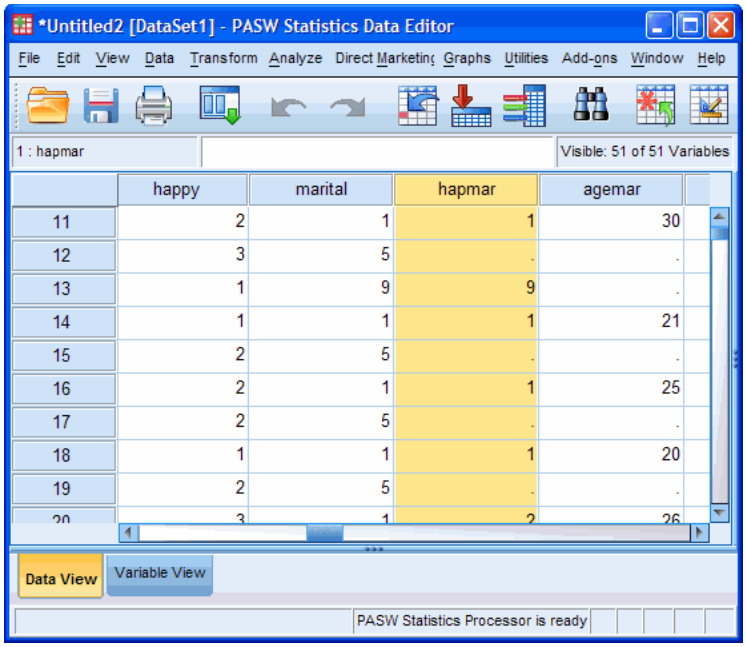
The *census_small.xls* and *census_small.sav* data files are used in this lesson. These data are a subset of demographic and attitudinal variables from a survey conducted in 2008 of a sample of the general population.

4.3 *Variable Properties*

Variable properties contain information about the data values; often referred to as metadata. In PASW statistics, each variable must have a unique variable name. Optionally additional properties can be defined for each variable. Some of the most commonly used properties are descriptive labels

for variables and data values, flags for values that indicate no answers or missing information, the type of variable, e.g. numeric, date, string, etc. and the measurement level of the variable. Variable properties can be illustrated by looking at data read from an Excel file. The figure below shows the Data Editor with data imported from Excel. If we run Frequencies on a variable, the result table is after the Data Editor figure.

Figure 4.1 Data Editor and Results Without Variable Properties Defined



	happy	marital	hapmar	agemar
11	2	1	1	30
12	3	5	.	.
13	1	9	9	.
14	1	1	1	21
15	2	5	.	.
16	2	1	1	25
17	2	5	.	.
18	1	1	1	20
19	2	5	.	.
20	3	1	2	26

hapmar

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	596	29.5	61.0	61.0
	2	343	17.0	35.1	96.1
	3	30	1.5	3.1	99.2
	8	1	.0	.1	99.3
	9	7	.3	.7	100.0
	Total	977	48.3	100.0	
Missing	System	1046	51.7		
Total		2023	100.0		

After importing data from Excel, no variable properties such as descriptive variable and value labels have been defined. As a result, the frequency table contains only the data values and cannot be interpreted without knowing the meaning of each variable name and data value.

- What does *hapmar* mean?
- What do the values 1, 2, 3, 8, 9 mean?
- Are values 8 and 9 special values?
- What is the "." value in the Data Editor?

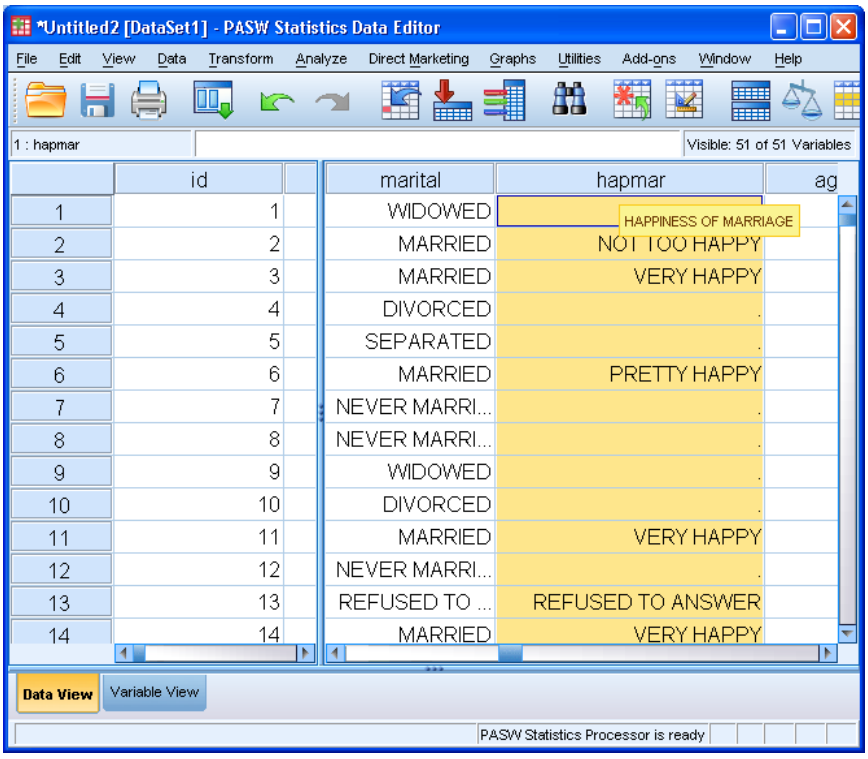
We want to attach labels and other definitions to the variable. Of course, we can only do so if we have the information available. In this case we do have this information.

- Variable *hapmar* means HAPPINESS OF MARRIAGE
- Values of *hapmar* are: 1: VERY HAPPY; 2: PRETTY HAPPY; 3: NOT TOO HAPPY; 8: DON'T KNOW, 9: NO ANSWER
- Values 8 and 9 represent values where an answer is missing. PASW Statistics allows you to flag these values as **user-missing,** so a case with this value will not be included in analyses by default. It is up to the researcher to decide what values will be defined as user-missing. For example, for some analysis "Don't know" might be considered as a missing value, in other cases, it may not. It is the decision of the researcher whether to treat this answer as valid or as missing.
- The variable *hapmar* shows a period (".") in the Data Editor when the respondent is not married. The period is called the **system-missing value** and represents no information for a variable. A blank or non-numeric value in a numeric variable in the data source (here: the Excel file) is converted to a system missing value. Or, if data for a new respondent are entered, but the data entry is not completed for that respondent, the remaining variables will have the system-missing value (."). In general, cases having the system-missing value on a variable will automatically be excluded from transformations and analyses for that variable.
- Variable *hapmar* has distinct ranked categories measuring the degree of happiness. The measurement level of ranked variables is defined as ordinal.

To summarize, we need to attach information about the data (metadata) in the form of variable properties to the variables in order to get labeled tables and charts and correct statistics.

The figures below show what we would like to accomplish.

Figure 4.2 Data Editor and Results With Labels Defined

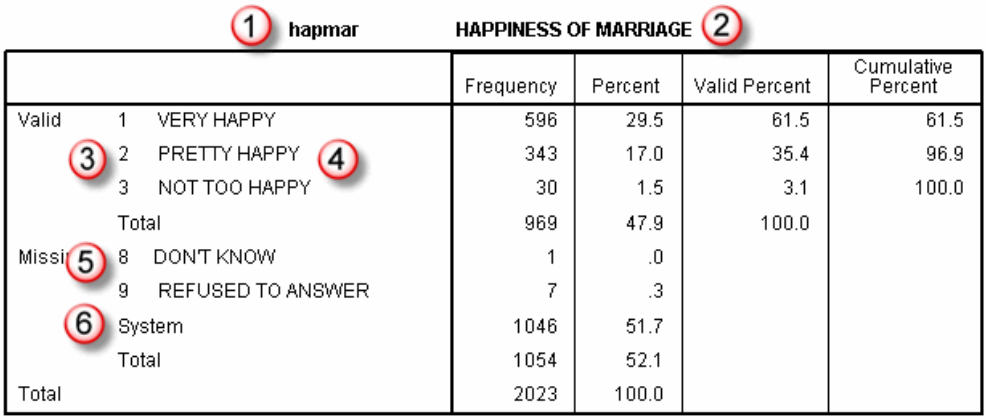


	id	marital	hapmar	ag
1	1	WIDOWED		
2	2	MARRIED	NOT TOO HAPPY	
3	3	MARRIED	VERY HAPPY	
4	4	DIVORCED	.	
5	5	SEPARATED	.	
6	6	MARRIED	PRETTY HAPPY	
7	7	NEVER MARRI...	.	
8	8	NEVER MARRI...	.	
9	9	WIDOWED	.	
10	10	DIVORCED	.	
11	11	MARRIED	VERY HAPPY	
12	12	NEVER MARRI...	.	
13	13	REFUSED TO ...	REFUSED TO ANSWER	
14	14	MARRIED	VERY HAPPY	

Although the data are still numeric values, we now have the value labels displayed in the Data Editor. Pointing at the variable name, the variable label pops up.

The Frequencies table below shows all of the applied variable properties for *hapmar* variable.

Figure 4.3 Frequency Table With Variable Properties Defined



1 hapmar HAPPINESS OF MARRIAGE 2

			Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	VERY HAPPY	596	29.5	61.5	61.5
	2	PRETTY HAPPY	343	17.0	35.4	96.9
	3	NOT TOO HAPPY	30	1.5	3.1	100.0
	Total		969	47.9	100.0	
Missi	8	DON'T KNOW	1	.0		
	9	REFUSED TO ANSWER	7	.3		
	System		1046	51.7		
	Total		1054	52.1		
Total			2023	100.0		

3 4 5 6

1) Variable name (*hapmar*)
2) Variable label (HAPPINESS OF MARRIAGE)
3) Values (1, 2, 3, 8, 9)
4) Value labels (VERY HAPPY, PRETTY HAPPY, etc)
5) User missing values (8,9)
6) System-missing value (“.” in the Data Editor)

The following 11 variable properties can be entered for each variable.

- **Name**: Each variable must be assigned a unique variable name no longer than 64 characters
- **Type:** The type or format of the variable (numeric, string, dollar, etc.)
- **Width:** The total number of columns (width) of the variable values
- **Decimals:** The number of decimal positions of the variable values
- **Label:** Variable label for the variable
- **Values:** Value labels for any value
- **Missing:** The values which should be flagged as user-missing and excluded by default from most analysis
- **Columns:** Changes the display width of the column in the Data View. (This can also be done directly in the Data View using the mouse.)
- **Align:** The alignment of the values in the Data View columns
- **Measure:** The level of measurement for the variable
- **Role:** Role of variable in analysis when used in new procedure dialogs

In this lesson we discuss two methods to attach these properties to the variables, using Variable View tab of the Data Editor and Define Variable Properties feature.

4.4 *Use Variable View*

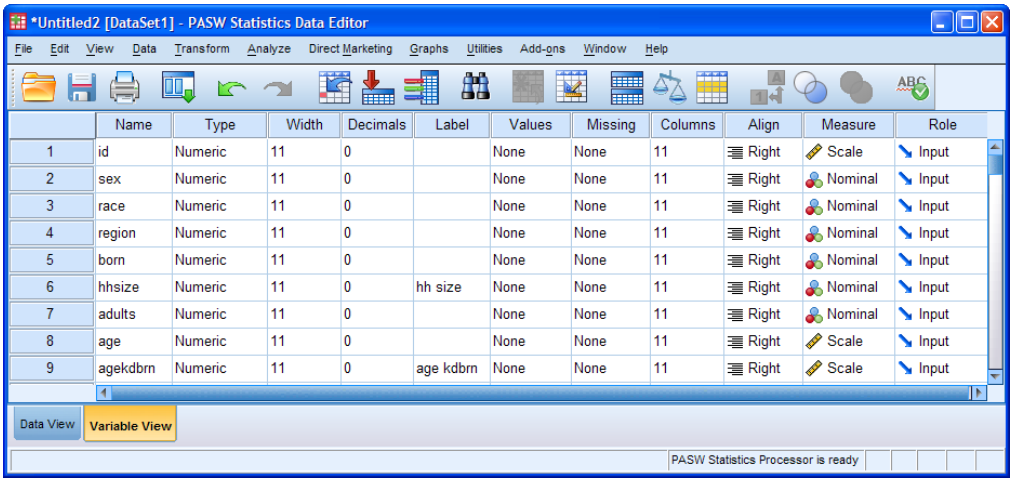
The Variable View provides a user-friendly interface to define and edit variable properties, variable-by-variable or using copy & paste operations to copy properties from one variable to a number of other variables. Also, use Variable View to create new variables and to define properties for the newly created variables.

Use the tab Variable View in the Data Editor to view, define and edit variable properties.

Tip

Double-click on a variable name in the Data View to display the Variable View window with focus at the selected variable.

Figure 4.4 Variable View: Variable Properties



Each row in the Variable View represents a variable and each column represents a variable property. The Variable View contains as many rows as the number of variables in the file. There are 11 columns for the 11 variable properties or attributes.

4.5 *Steps to Use Variable View*

Adding or editing variable properties using the Variable View is completed with these steps:

1) Select the cell for the variable and property that you want to change.
2) Type the text for Name and Label properties and optionally check spelling for text.
3) For all other attributes, select from the dropdown list or click ... in the selected cell. This will display a dialog box.
4) Complete your request in the dialog box.

4.6 *Procedure: Use Variable View*

The following properties are available for each variable:

Name

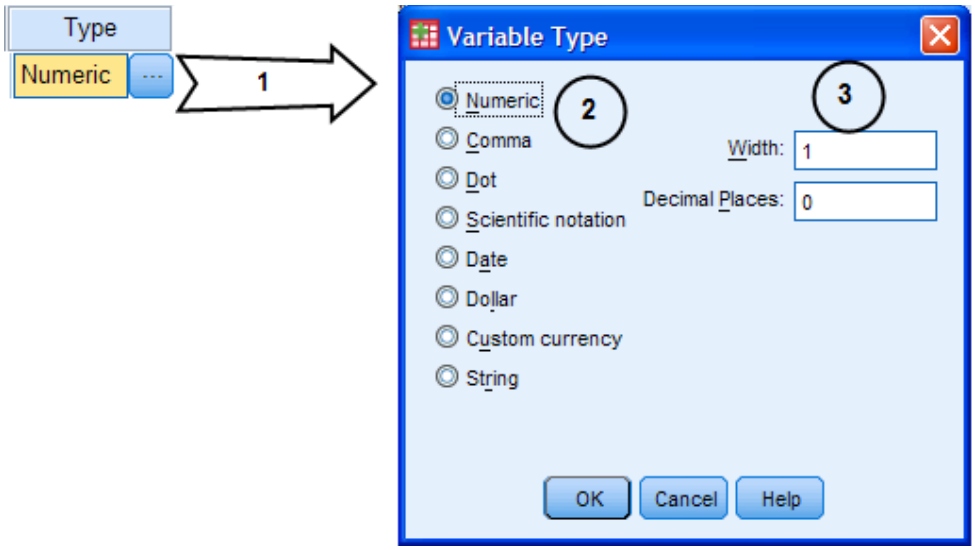
Type a variable name. Each variable must be assigned a unique name, constructed according to the following rules:

- Maximum of 64 characters in length
- Must begin with a letter; remaining characters can be any letter, any digit, a period, or most special characters.
- Cannot contain special characters (! ? ' *) or blanks
- Cannot end with an underscore or period
- Can specify any combination of lower and/or upper case characters

Type

Specify the type or display format of the variable. This affects the way the values are displayed in the Data Editor and in tables and charts.

Figure 4.5 Variable Type Dialog Box



1) Select the type cell for the variable and click on the tool button.
2) Select the variable type from the list.
3) For each variable type further properties can be defined.

For a numeric variable, define

Width The total number of columns (width) to display the variable values. The width should be large enough for the values to be entered as well as any special characters that are displayed with the requested format, For example, if the value 12 is entered, the width should be 2; if the value 12.34 is entered, width should be 5 (including the decimal separator).

Decimals The number of decimal positions used to display the variable values.

For string variables, specify the width or maximum number of characters in the variable. Comma and Dot formats display numeric values with comma or dots (periods). A date variable can be displayed in formats such as mm-dd-yyyy or dd/mm/yyyy).

Warning

Changing a variable's type can result in lost data values for that variable. For example, changing a string variable with values "01/05/2010" to a date variable will cause all data values to be missing. Use the Date and Time Wizard, discussed in the *Data Management and Manipulation Using PASW Statistics* course, to convert string variables to date variables.

Label

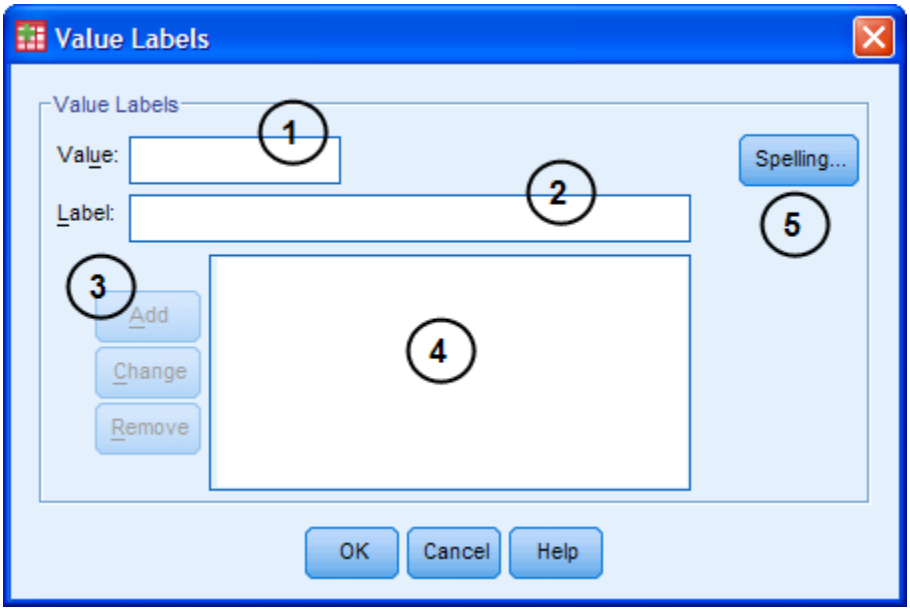
Optionally, type a variable label for each variable. You can assign a descriptive label up to 256 characters long. Variable labels can contain spaces and reserved characters that are not allowed in variable names. If desired a spelling check can be done on the variable labels.

Values

Optionally, type labels for any value of any variable. Each value label can be up to 120 characters long. For each value, the label is specified and added to the list of defined values. If desired, a spelling check can be done on the value labels.

To open the Value Labels dialog, select  tool.

Figure 4.6 Value Labels Dialog Box



1) Type the data value
2) Type the label text
3) Select **Add** to add a new label, **Change** to change a highlighted existing label, **Remove** to delete a highlighted label
4) Displays a list of all defined value labels for the variable. Highlight a label to change or remove it.
5) Optionally, select Spelling checker to check the spelling of the label text.

Further Information

The Spell Checker is also available for variable labels and alphanumeric data values. The *Working with the Data Editor* Lesson discusses the Spell Checker.

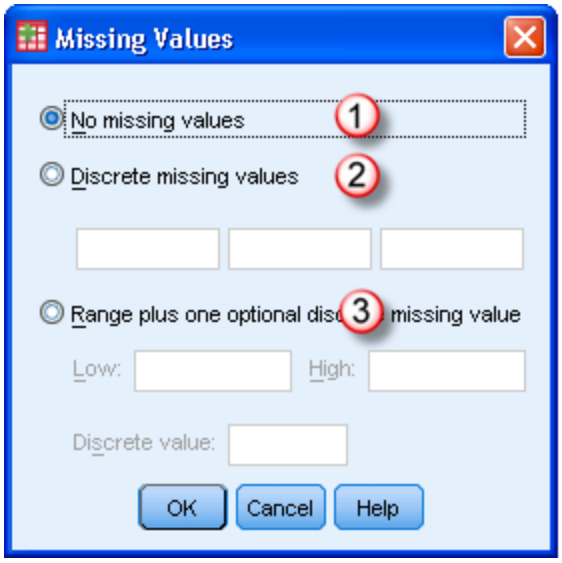
Value labels can be defined both for numeric and string variables. String values do not need to be quoted when entered in the Value Labels dialog box.

Missing

Specify data values to flag as user-missing values. In the Missing Values dialog box up to three distinct missing values can be specified, or a range of values plus one additional discrete value.

To open the Missing Values dialog, select  tool.

Figure 4.7 Missing Values Dialog Box



1) Select *No missing values* which is the default
2) Select *Discrete missing values* and type up to three discrete values with one value in each text box
3) Select *Range plus one optional discrete missing value*. Type the Low and High values and optionally one additional discrete value.

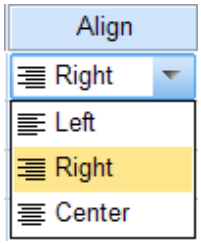
Missing values can be defined both for numeric and string variables. String values do not need to be quoted in the Missing Values dialog box. To specify a blank, use the space bar to indicate a blank value.

Columns

Specify the display width of the column in the Data View. The column width can be changed directly in the Data View of the Data Editor by clicking and dragging the width of the column.

Align

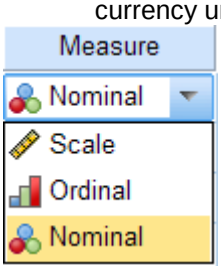
Specify the alignment of the values of the variable in the Data View columns. Choices in the dropdown list are **right**, **center**, or **left**. By default, numeric variable values are right aligned and string variable values are left aligned.



Measure

Specify the level of measurement for the variable. You can define one of three levels of measurement for each variable.

- **Nominal**. Categorical data where there is no inherent order to the categories. Examples are region of the country, marital status, and ethnic groups. All string variables are nominal.
- **Ordinal**. Categorical data where there is a rank order of categories. Examples are attitude rating scales such as "Strongly Agree", "Agree", "Disagree", and “Strongly Disagree".
- **Scale**. Data where the values indicate both order and distance. Examples are salary in currency units, temperature scales and age in years.



Many of the analysis and reporting procedures recognize these measurement level definitions. In these procedures, different options and statistics are available for variables depending on their level of measurement. For example, the Codebook reporting procedure recognizes the measurement level definitions and reports different statistics for nominal/ordinal variables than it does for scale variables.

Role

Dialog boxes for new procedures in Version 18 support predefined roles that can be used to select variables for analysis. When you open one of these dialogs, variables that meet the role requirements will be automatically displayed in the destination list(s). By default, the role of all variables is set to **Input** which means that all of the variables are available for analysis. Detailed discussion of role is beyond the scope of this lesson.

Tip

Variable property definitions can be copied from one variable and pasted to other variables by using standard Microsoft Windows procedures (keystrokes CTRL-C, CTRL-V; the Edit menu, or the right-click popup menu.)

Note

PASW Statistics syntax is not built when you add or modify variable attributes in the Variable View of the Data Editor. Use the Define Variable Properties feature to build and save syntax.

4.7 *Demonstration: Use Variable View*

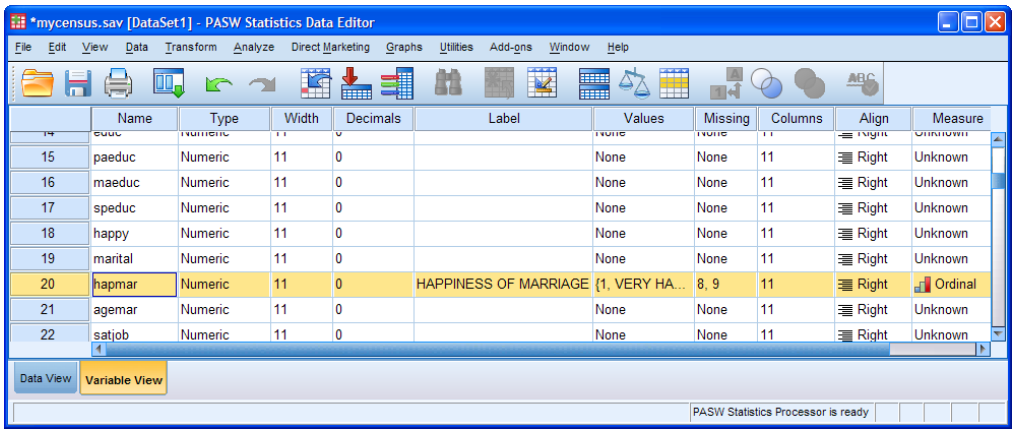
To demonstrate variable definitions, we start by reading the Excel file *census_small.xls*. After importing this file, we use the Variable View to add variable properties.

The figure below shows the definitions for the variable *hapmar*:

- **Name**: *hapmar*
- **Type:** Numeric

- **Width:** 1 (width 1 suffices to accommodate values 1, 2, 3, 8, 9)
- **Decimals:** 0 (no decimals will be displayed)
- **Label:** HAPPINESS OF MARRIAGE
- **Values:** Value labels 1: VERY HAPPY; 2: PRETTY HAPPY; 3: NOT TOO HAPPY; 8: DON'T KNOW; 9: NO ANSWER
- **Missing:** 8 , 9
- **Columns:** 8
- **Align:** Right
- **Measurement**: Ordinal

Figure 4.8 Variable View Displaying Variable Properties for *hapmar* Variable



Steps to add the Variable Properties

1) Type the variable label HAPPINESS IN MARRIAGE) in the Label column for *hapmar*
2) In the Values column, open the Value Labels dialog for *hapmar* using the ... tool , add labels for the values (1: VERY HAPPY, 2: PRETTY HAPPY; 3: NOT TOO HAPPY; 8 DON'T KNOW; 9 NO ANSWER)
3) In the Missing column, open the Missing dialog, select *Discrete Missing Values*, type value 8 and 9 into the first two text boxes.
4) In the Measure column, select the measurement level *Ordinal*

Variable properties for a few additional variables in the file are shown below.

Table 4.1 Variable Attributes for Additional Variables

Variable	Variable Label	Value Labels	Missing Values	Measurement Level
id	Respondent ID Number			Nominal
sex	Respondent's sex	1 Male 2 Female		Nominal
age	Age of respondent	999 no answer	999	Scale
marital	Marital Status	1 Married 2 Widowed 3 Divorced 4 Separated 5 Never Married 9 No Answer	9	Nominal

conmedic	Confidence in medicine	0 Not applicable 1 A great deal 2 Only some 3 Hardly any 8 Don't know 9 No answer	0, 8, 9	Ordinal
coneduc	Confidence in education	See conmedic	See conmedic	See conmedic
confinan	Confidence in banks & financial institutions	See conmedic	See conmedic	See conmedic

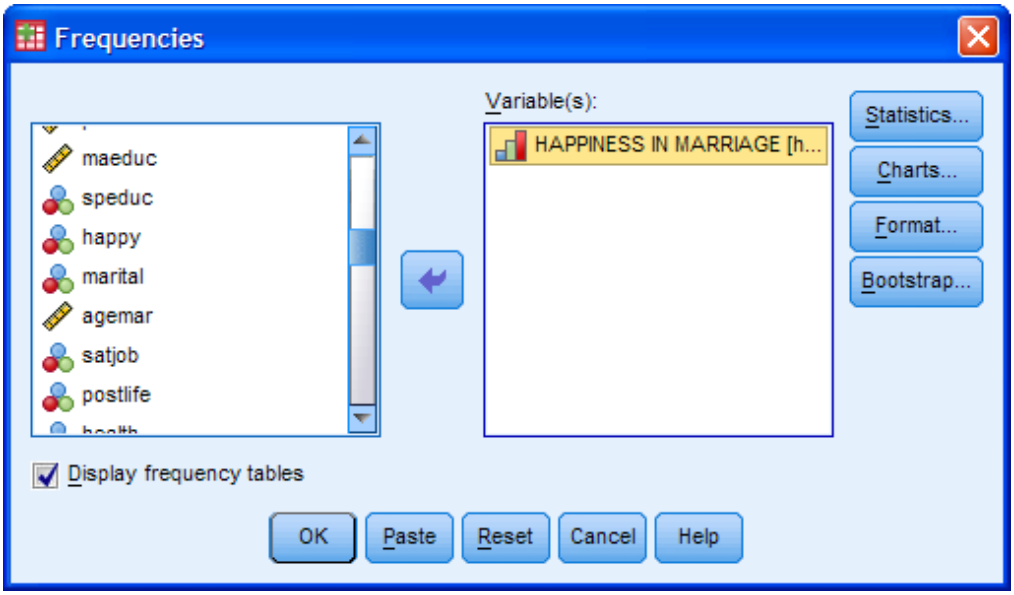
Note

The confidence variables have the same value labels, missing values and measurement level definitions, so copy & paste can be used to define these properties. However, this method does not perform checking operations as does the **Define Variable Properties** feature presented later in this lesson.

4.8 *Results: Use Variable View*

Variable definitions can be checked by running Frequencies or Codebook procedure on the variables. For example, to check the properties entered for *hapmar*, run the Frequencies procedure which is found on the Analyze…Descriptive Statistics menu.

Figure 4.9 Frequencies Dialog for *hapmar* Variable



Move the *hapmar* variable to the Variable(s) box and run. The Frequencies table is shown below.

Figure 4.10 Frequencies Table for *hapmar* Variable

HAPPINESS OF MARRIAGE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	VERY HAPPY	596	29.5	61.5	61.5
	PRETTY HAPPY	343	17.0	35.4	96.9
	NOT TOO HAPPY	30	1.5	3.1	100.0
	Total	969	47.9	100.0	
Missing	DON'T KNOW	1	.0		
	NO ANSWER	7	.3		
	System	1046	51.7		
	Total	1054	52.1		
Total		2023	100.0		

Apply Your Knowledge: Use Variable View

1. True or false: Two variables can have the same variable name.

2. True or false: In the Value Labels dialog box, PASW Statistics lists all values in the data that are unlabeled.

3. Mark all correct statements:
    a. The system-missing value can have a value label
    b. Up to 3 discrete user-missing values can be defined for a variable
    c. A string variable cannot have a user-missing value
    d. A string variable cannot have value labels

4. Mark the valid PASW Statistics variable names:
    a. gender
    b. gender of the respondent
    c. 1gender
    d. Var1:gender

4.9 *Define Variable Properties Feature*

Using the Variable View tab in the Data Editor is one way to define variable properties. However, the Variable View provides few checks without running a procedure such as Frequencies or Codebook. For categorical (nominal/ordinal) data, Define Variable Properties can help you define value labels and other variable properties. The Define Variable Properties feature has the following features:

- Scans the actual data values and lists all unique data values for each selected variable
- Identifies unlabeled values and provides an "auto-label" feature
- Provides a suggestion for measurement level of the selected variable
- Provides the ability to copy variable properties from another variable to the selected variable or from the selected variable to multiple additional variables
- Builds PASW Statistics syntax of the operations for later use

4.10 *Steps to Use Define Variable Properties*

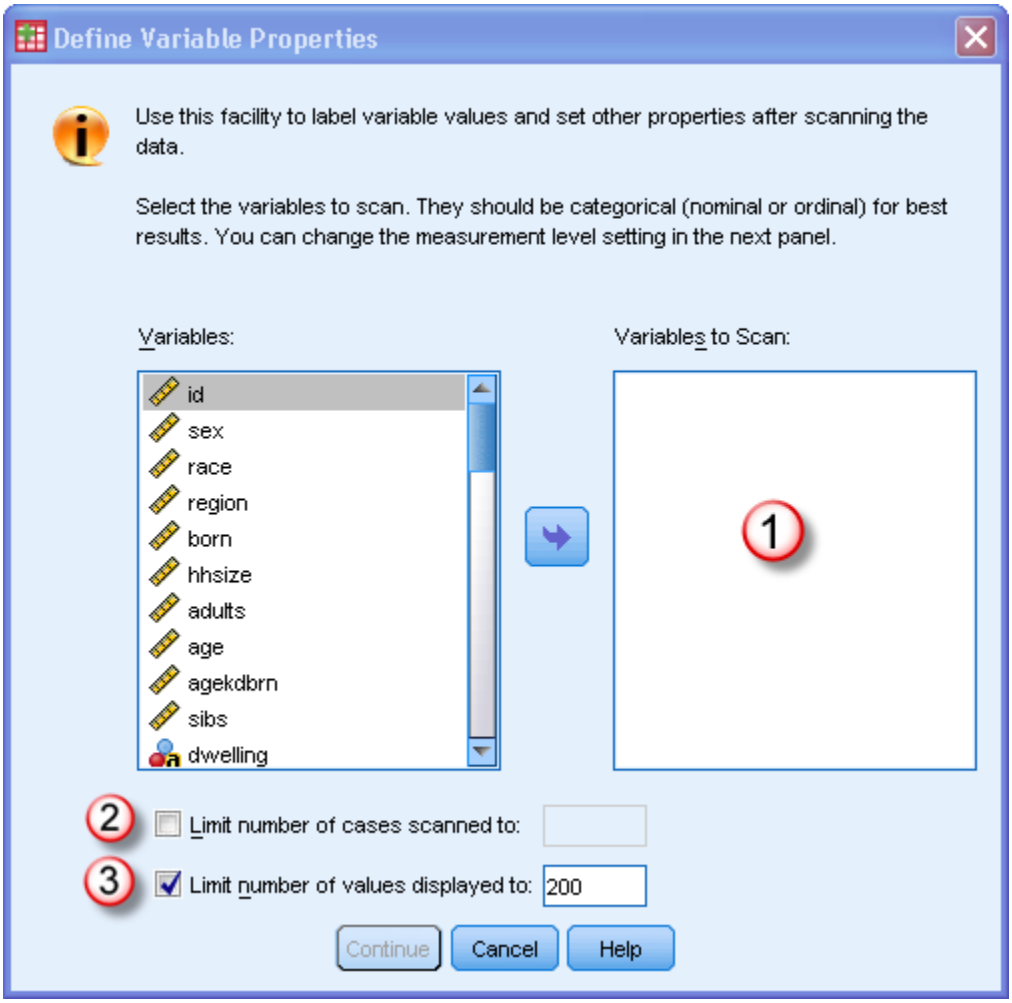
Define Variable Properties is accessed from the Data menu.

1) Select the variables that you want to scan
2) In the Define Variable Properties results, select a variable to view the scanned results and the existing variable properties
3) Enter variable properties
4) Optionally, copy variable properties to or from other scanned variables

4.11 *Procedure: Use Define Variable Properties*

Define Variable Properties is accessed from the Data menu. In the initial Define Variable Properties dialog box, select the variables for which you want to define or edit the properties.

Figure 4.11 Define Variable Properties Dialog: Step 1



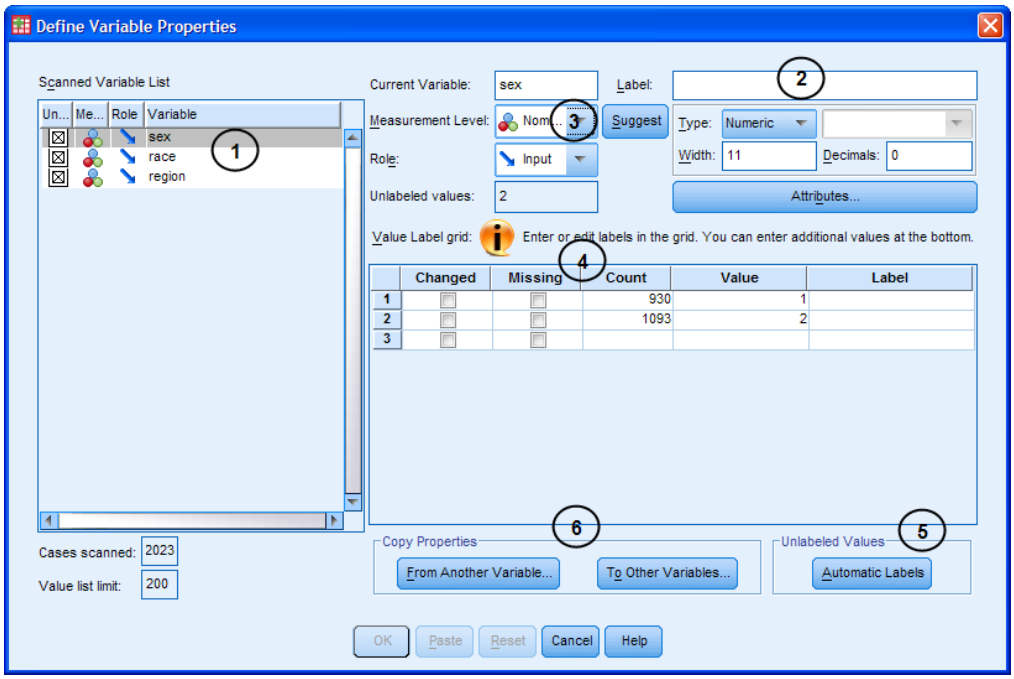
1) Select variables to scan and define variable properties
2) Check box to *Limit the number of cases scanned* and type the number of cases to scan. By default, all cases will be read.
3) Check box to *Limit the number of values displayed* for each variable and type the number of values. The default is to display a maximum of 200 values for each variable.
4) Select **Continue**

Tip

Limiting the number of cases to scan is usually not necessary unless you have over 10,000 cases and you are scanning a large number of variables.

In the next dialog box variable properties are edited.

Figure 4.12 Define Variable Properties Dialog: Step 2



1) Displays list of scanned variables. Select a variable from this list to display the variable's information.
2) Displays variable label for the selected variable or type a label if none exists.
3) Displays the measurement level. If you are unsure of the correct measurement level, ask to **Suggest** one based on number of data values.
4) Value Label grid displays the unique data values scanned and the count (number of cases) for each value.
5) Select to generate Automatic Labels for scanned values.
6) Options to **Copy Properties** from or to other variables in the scanned list.

The Value Label Grid displays all the unique data values for the selected variable, any defined value labels for these values, whether a value is defined as a user-missing value and the number of times (count) each value occurs in the scanned cases.

Further Information

Define Variable Properties provides a check on the values of the variable, Variable View doesn't. Define Variable Properties does not allow the user to change the variable name, or to create a new variable. Variable View allows both.

4.12 *Demonstration: Define Variable Properties*

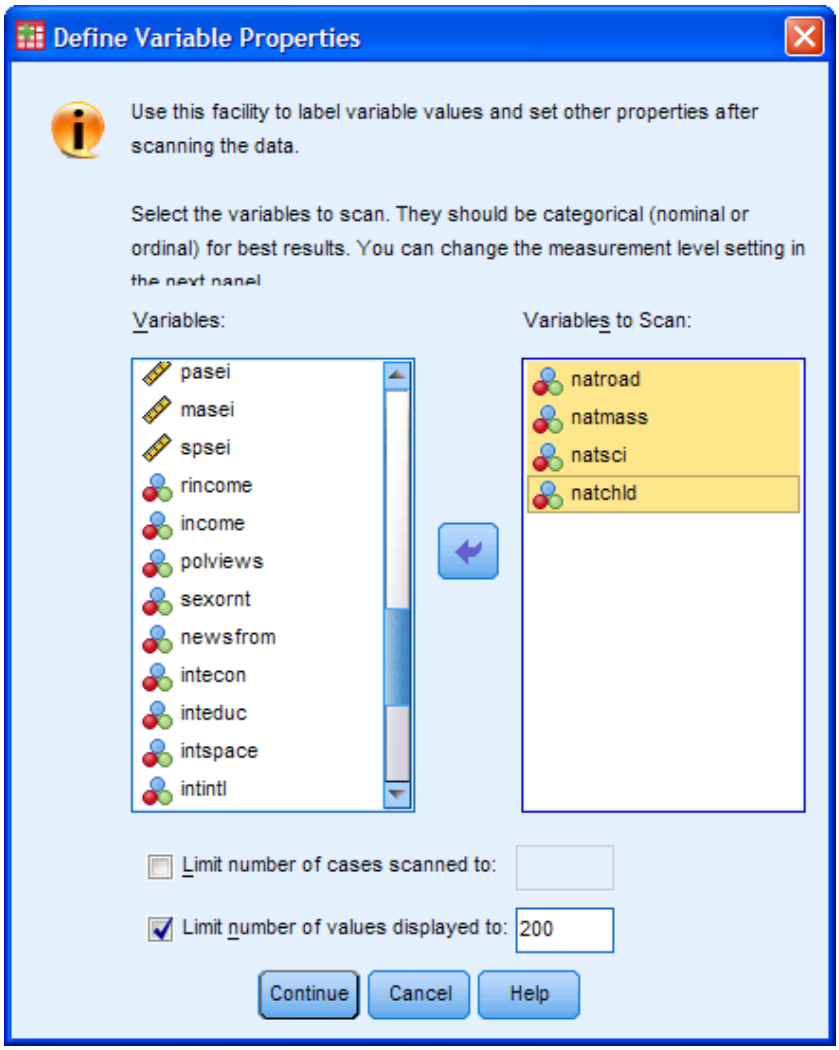
In this example we will demonstrate how to use Define Variable Properties to define the same variable properties for a set of variables. You often need to do this when you have a number of attitude variables that share the same possible values.

For example the variables *natroad, natmass, natsci* and *natchld* are variables that ask about level of government spending in a number of different areas such as road and bridge construction, mass transportation, scientific research, and childcare support. They share the same coding scheme: 1: TOO LITTLE; 2: ABOUT RIGHT; 3: TOO MUCH (and user missing values: 8: DON'T KNOW; 9: REFUSED TO ANSWER).
From the Data menu, select the feature Define Variable Properties.

Complete the following steps:

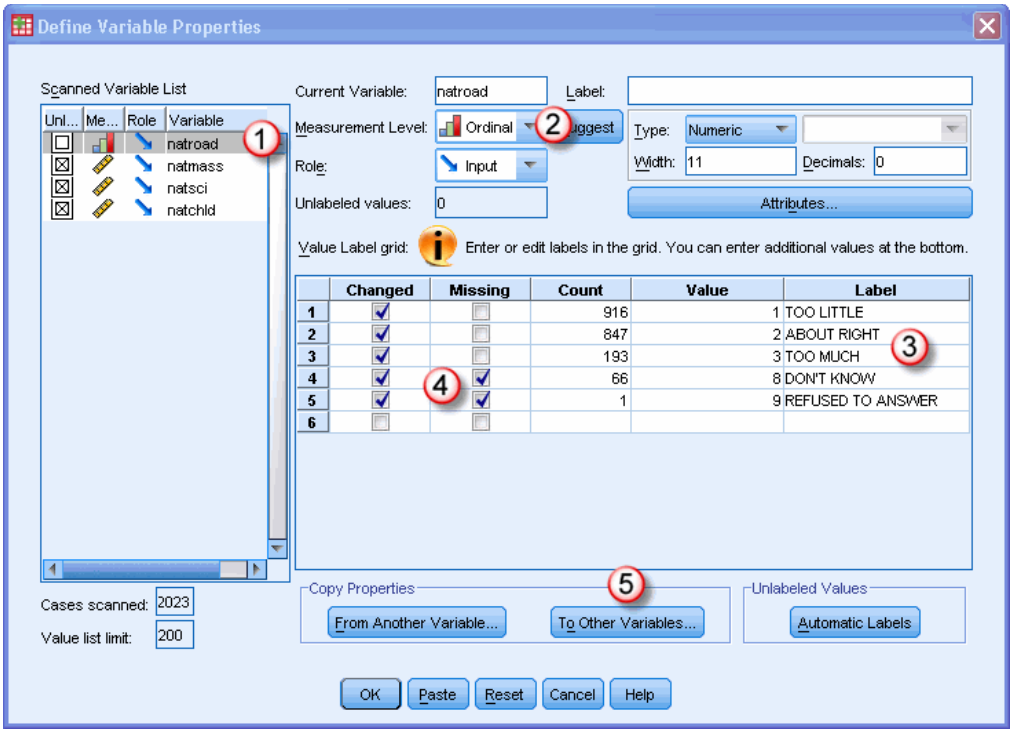
1) Select the variables *natroad* through *natchld* to be scanned
2) Define value labels, missing values and measurement level for one of the variables, *natroad*.
3) Copy the definitions to the other scanned variables.

Figure 4.13 Define Variable Properties Demonstration: Step 1



1) Move the variables *natroad* through *natchld* to the Variables to Scan list box

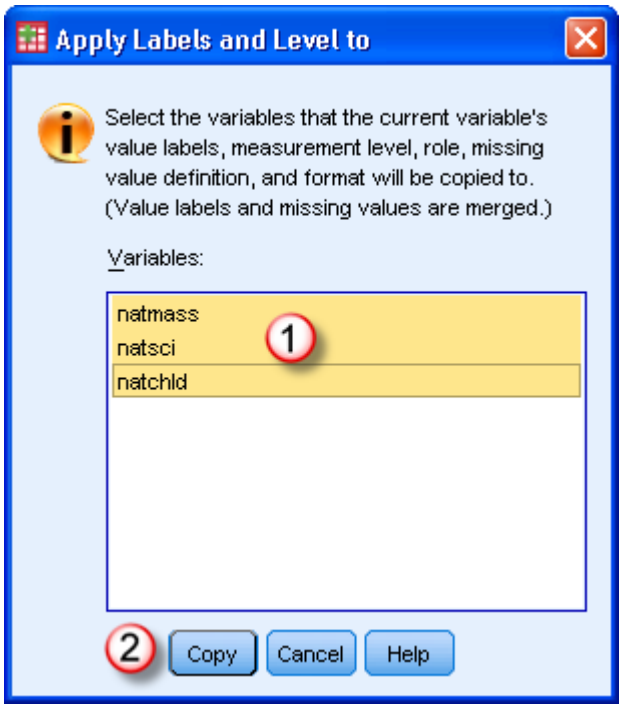
Figure 4.14 Define Variable Properties Demonstration: Step 2



1) Select *natroad*
2) Set measurement level to **Ordinal**
3) Type value labels as shown above
4) Define missing values by checking the box in the Missing column for values 8 and 9
5) Select the button **To Other Variables** in the Copy Properties area

In the next dialog box, select the variables to which the definitions are to be copied.

Figure 4.15 Define Variable Properties Demonstration: Step 3

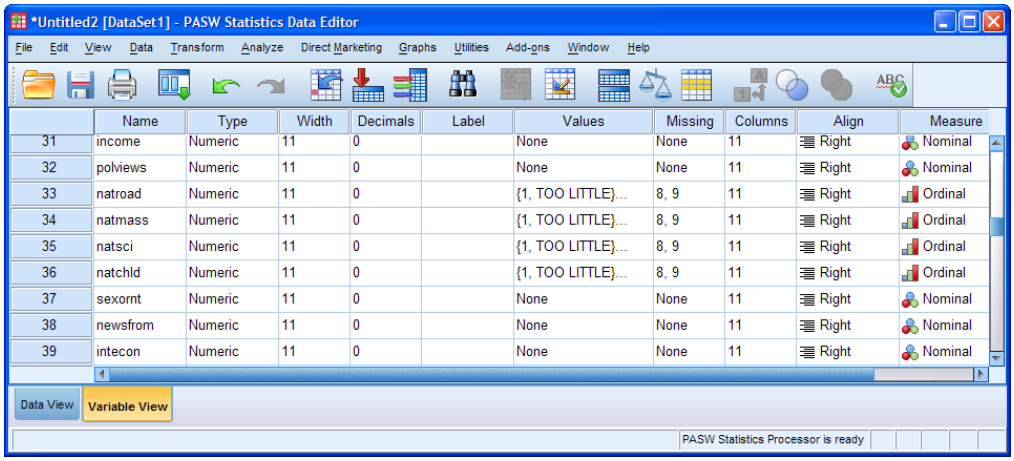


1) Select all variables
2) Select **Copy** to copy all variable properties from *natroad* to these variables

Results Using Data…Define Variable Properties

The Variable View shows that all of the scanned variables now have the same variable properties that we entered for the *natroad* variable.

Figure 4.16 Define Variable Properties Results: Variable View



Note

Variable labels are not copied in the Define Variable Properties feature. Use the Variable View to enter a unique variable label for each of these variables. Or, enter a variable label for each variable in the Define Variable Properties dialog.

Further Information

New variable properties are not automatically saved. After completing the variable definitions, save the changes in a new PASW Statistics data file. All additional analysis should be done on the saved PASW Statistics data file.

Apply Your Knowledge: Data…Define Variable Properties

For each of the following statements, answer True or False.

1. The variable name can be edited in Define Variable Properties.
2. Define Variable Properties provides a list of unique data values and the count for each value.
3. The Define Variable Properties feature automatically saves the new variable properties in an updated data file.

4.13 *Save the Data and Variable Properties*

The variable properties that are entered are not automatically saved. Rather the information and data values must be saved in a PASW Statistics data file for use in further analysis.

4.14 *Procedure: Save a PASW Statistics Data File*

To save a data file in PASW Statistics, use the **File** menu from the Data Editor window. Select **Save** to save the file with the same name or **Save As** to give the file a new name.

Tip

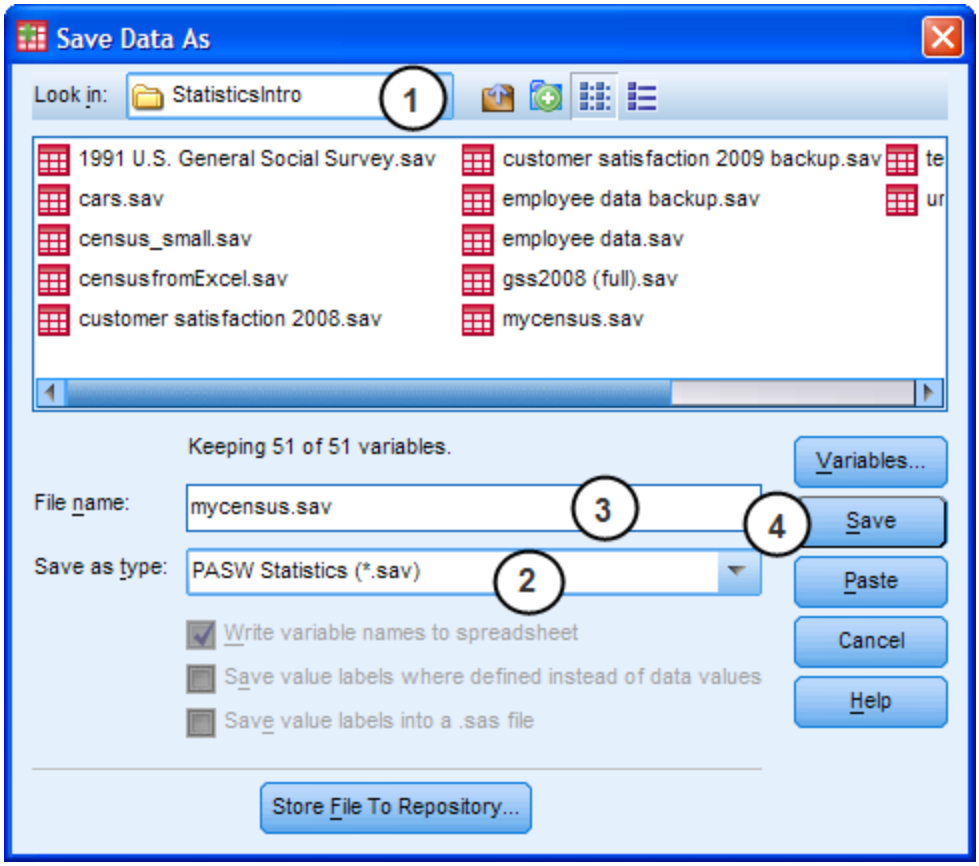
When you save a data file, we recommend using **Save As** and giving the saved file a new file name. This provides a backup if you have made significant changes such as adding many variable properties.

Detailed Steps to Save a PASW Statistics Data File

1) Browse to the folder in which to save the file using the **Look in:** dropdown list
2) Select Save as type: **PASW Statistics (*.sav).** This is usually the default setting.
3) Type the file name for the saved file.

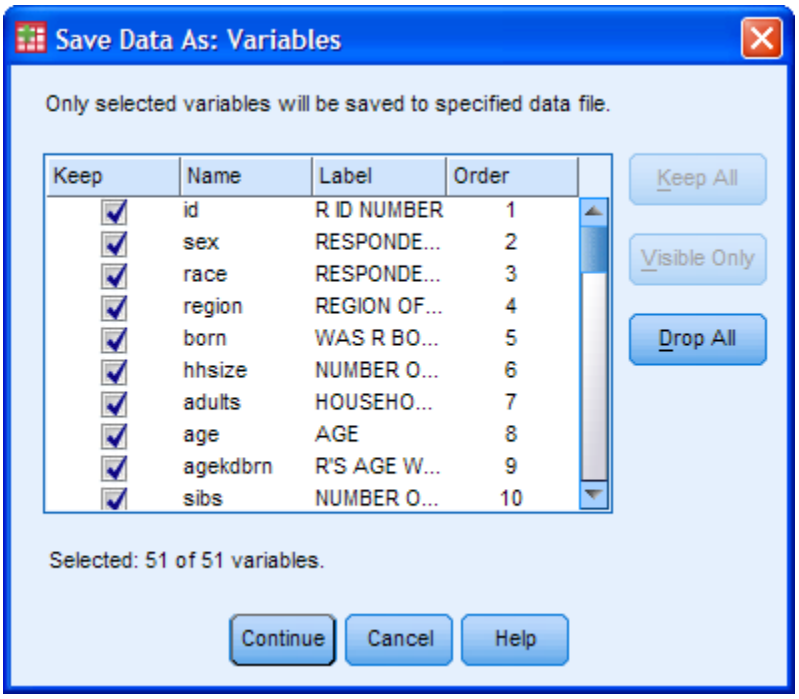
4) Select the **Save** button to save all of the variables and variable properties in the Data Editor.

Figure 4.17 Save Data As Dialog Box



Optionally, select the **Variables** button to save a subset of variables in the Data Editor window. This opens the Save As: Variables dialog shown below.

Figure 4.18 (Optionally) Save Data As: Variables Dialog Box



Check the box in the Keep column for those variables to be saved. The buttons on the right of the dialog, **Keep All**, **Visible Only**, and **Drop All** provide shortcuts for selecting and deselecting variables. The default is to keep all variables in the Data Editor.

Further Information

The **Save as type:** dropdown list provides a number of other formats such as Excel and comma-delimited to save the data. However, most variable properties will not be saved in these file formats. If saving to an Excel file, options are provided to write the variable names to the first row of the spreadsheet and to save the value labels instead of the data values.

4.15 *Demonstration: Save a PASW Statistics Data File*

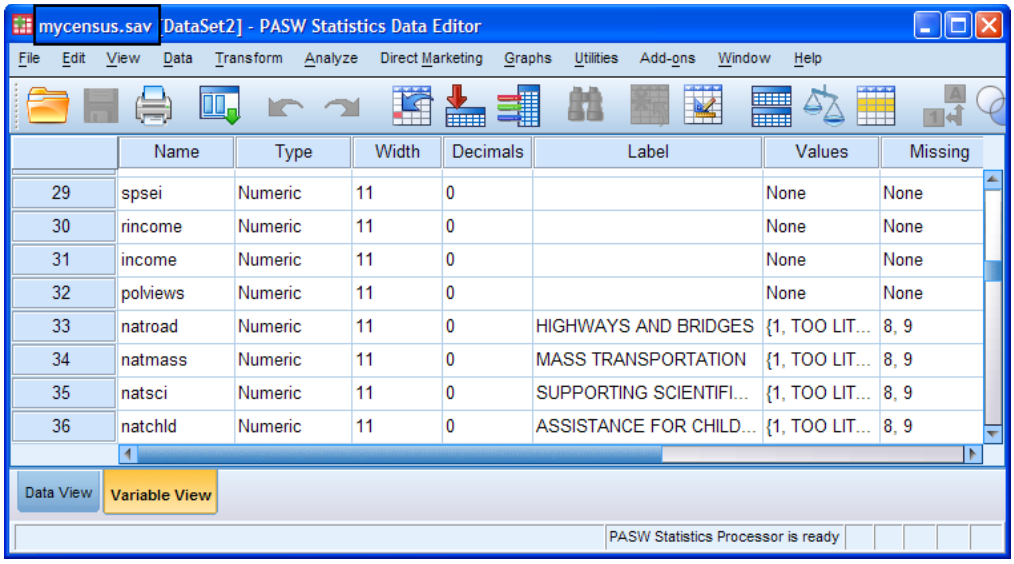
To demonstrate, we will save all the data from the Excel file that we read at the beginning of this lesson along with all of the variable properties that we have added in a PASW Statistics data file named *mycensus.sav*.

Detailed steps to save *mycensus.sav*:

1) If necessary, use the Look in: dropdown list to browse to the folder *c:\Train\StatisticsIntro* (or the folder being used for this instruction)
2) Select Save as type: **PASW Statistics (*.sav)** (if necessary)
3) Type the file name *mycensus*.
4) Select the **Save** button to save all of the variables and variable properties in the Data Editor.

The file has now been saved to use for further analysis. As well, the name of the Data Editor window has been changed to the saved file name.

Figure 4.19 Data Editor of *mycensus.sav* File



Apply Your Knowledge: Save Data and Variable Properties

For each of the following statements, answer True or False.

1. Variable properties are saved on all saved file formats.

2. Value labels are always written when saving an Excel file.

4.16 *View Variable Properties*

Most data files contain a large number of variables and it is not always easy to remember the properties of each variable. You may want to view the properties of variables interactively as you are completing a dialog box. Or, you may want to produce documentation, often referred to as a codebook that lists all of the information about the variables in the data. PASW Statistics provides several options for viewing the variable attributes. In this lesson, we will review three.

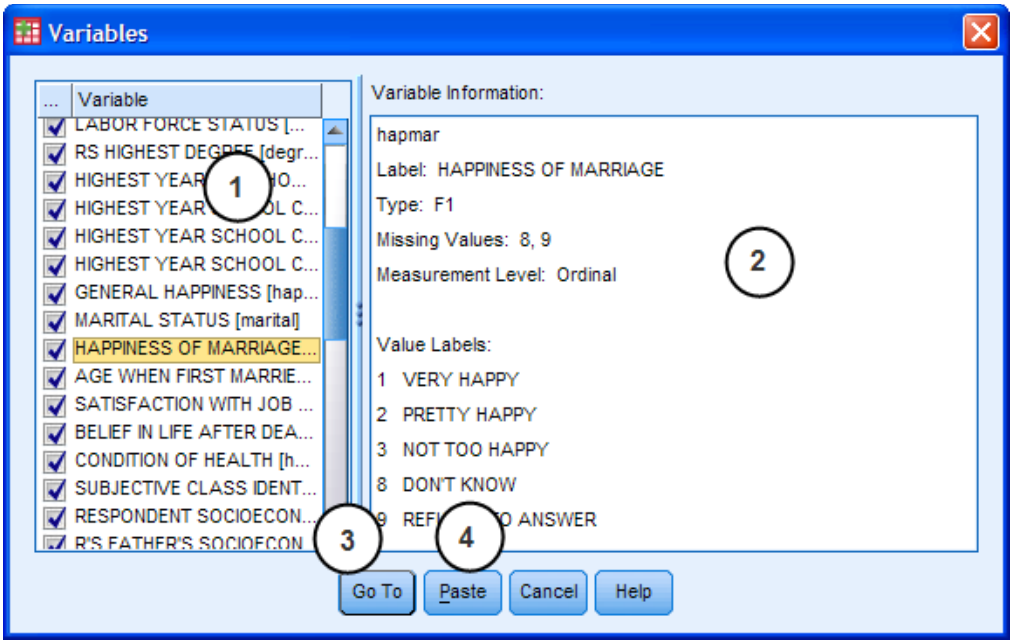
- **Variables Information Dialog**: View variable properties interactively in a dialog box
- **Display Data File Information**: Reports variable properties in two output tables
- **Codebook**: Reports variable properties and summary descriptive tables for each variable

4.17 *Procedure: Variables Information Dialog*

You can view the variable information for all your variables interactively in the Variables dialog box. This dialog box can remain open as a reference while making specifications in other dialog boxes.

To open the Variables dialog box, select **Utilities…Variables**. Alternatively, click the  button in the Toolbar.

Figure 4.20 Variables Information Dialog Box



1) Select a variable
2) Information for the selected variable is displayed
3) **Go to** button goes to the selected variable in the Data Editor
4) When working with PASW Statistics syntax, use the **Paste** button to copy the name of the selected variable into the syntax window

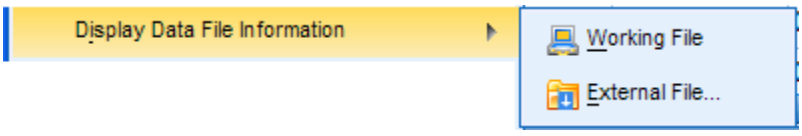
Tip

The Go To button is especially useful when you are working with a large number of variables.

4.18 *Procedure: Display Data File Information*

You can produce a document recording all the variable information as a useful permanent reference that can be shared with others.

To request data file information, select **File…Display Data File Information**.



Select *Working File* to produce data dictionary tables on the file in the active Data Editor window, or select *External File* for any external PASW Statistics data file. If *External File* is selected, the tables are produced in the Output Viewer window; but the data is not read or displayed in a Data Editor window

Results: Display Data File Information

Two tables display the variable attributes:

- **Variable Information table** displays the attributes for all the variables in the file.
- **Variable Values table** displays the value labels for any variable with at least one value label.

The next figure shows the first part of each of these tables for the *census_small.sav* file.

Figure 4.21 Data File Information Tables

Variable Information

Variable	Position	Label	Measurement Level	Role	Column Width	Alignment	Print Format	Write Format	Missing Values
id	1	R ID NUMBER	Nominal	Input	8	Right	F4	F4	
sex	2	RESPONDENTS SEX	Nominal	Input	6	Right	F1	F1	
race	3	RESPONDENTS RACE	Nominal	Input	5	Right	F1	F1	
region	4	REGION OF INTERVIEW	Nominal	Input	6	Right	F1	F1	
born	5	WAS R BORN IN THIS COUNTRY	Nominal	Input	5	Right	F1	F1	
hhsize	6	NUMBER OF PERSONS IN HOUSEHOLD	Scale	Input	5	Right	F2	F2	99
adults	7	HOUSEHOLD MEMBERS 18 YRS AND OLDER	Scale	Input	5	Right	F1	F1	99
age	8	AGE	Scale	Input	6	Right	F3	F3	999
agekdbrn	9	R'S AGE WHEN 1ST CHILD BORN	Scale	Input	8	Right	F2	F2	998, 999

Variable Values

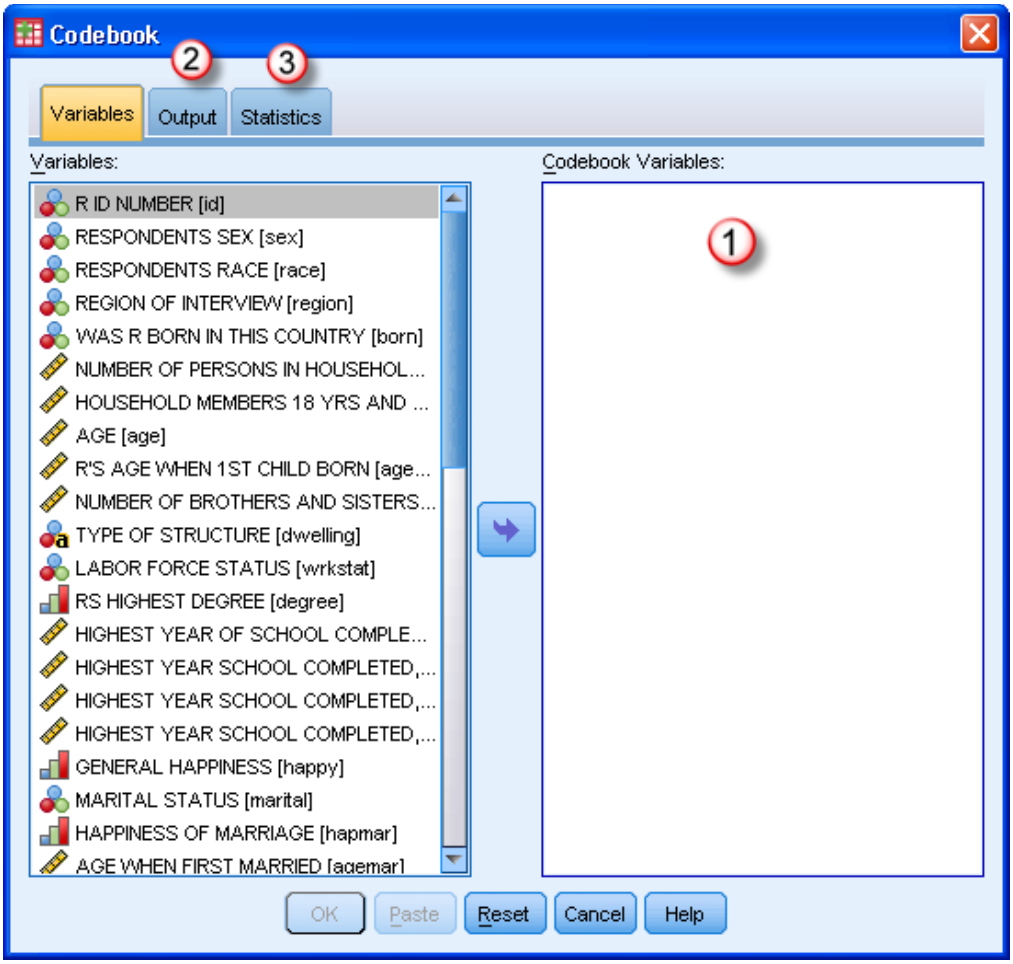
Value		Label
sex	1	MALE
	2	FEMALE
race	1	WHITE
	2	BLACK
	3	OTHER
region	1	NORTH
	2	SOUTH
	3	EAST
	4	WEST
	5	NORTH EAST
	6	NORTH WEST
	7	SOUTH EAST
	8	SOUTH WEST
born	1	YES
	2	NO
hhsize	99[a]	REFUSED TO ANSWER

4.19 *Procedure: Display Codebook Tables*

You can produce all of the data dictionary information as well as summary statistics such as counts and percentages combined in one table for each variable. This group of tables can be used to completely document a PASW Statistics data file. Survey researchers refer to this documentation as a codebook.

This is an easy way to document all information in a PASW Statistics data file which you can archive and share with others.
Codebook tables are requested by selecting **Analyze…Reports…Codebook** from the menus.

Figure 4.22 Codebook Dialog Box



1) Select the variables to display. You can run the codebook on selected variables or on all variables in the file.
2) Optionally on the **Output** tab, select variable attributes to display in each table and the order of the tables. By default, all variable attributes are displayed and the tables are in the order of the **Codebook Variables:** list of selected variables.
3) Optionally on the **Statistics** tab, select the statistics to display in the tables. By default, counts and percents are displayed for variables defined as nominal or ordinal measurement level and for labeled values of scale variables; mean, standard deviation, and quartile for scale variables.

Important

The Codebook procedure uses the measurement level defined for each variable to determine the summary statistics produced in the table.

Be sure that you have defined the measurement attribute appropriate for each variable before running Codebook tables.

Results: Display Codebook Tables

A table is displayed for each selected variable. Counts and percentages are displayed for variables of nominal or ordinal measurement level; means, standard deviations, and quartiles for scale variables.

The next figure shows a codebook table for the variable *sex* which is defined as a nominal variable and the table for the variable *age* which contains data values for years of age and is defined as a scale variable.

Summary statistics appropriate for the measurement level are displayed in each table and the set of tables documents the variable information associated with each variable as well as a summary of the data values for each variable.

Figure 4.23 Codebook Tables for Categorical and Scale Variables

sex

		Value	Count	Percent
Standard Attributes	Position	2		
	Label	RESPONDENTS SEX		
	Type	Numeric		
	Format	F1		
	Measurement	Nominal		
	Role	Input		
Valid Values	1	MALE	930	46.0%
	2	FEMALE	1093	54.0%

age

		Value	Count	Percent
Standard Attributes	Position	8		
	Label	AGE		
	Type	Numeric		
	Format	F3		
	Measurement	Scale		
	Role	Input		
N	Valid	1816		
	Missing	207		
Central Tendency and Dispersion	Mean	47.69		
	Standard Deviation	17.403		
	Percentile 25	34.00		
	Percentile 50	47.00		
	Percentile 75	60.00		
Labeled Values	999	REFUSED TO ANSWER	207	10.2%

4.20 *Demonstration: View Variable Properties*

In this exercise, we will view variable properties on selected variables in the *census_small.sav* PASW Statistics data file using each of the techniques we have reviewed.

If necessary, open the census_small.sav PASW Statistics data file to use for this demonstration.

View Variable Information Interactively

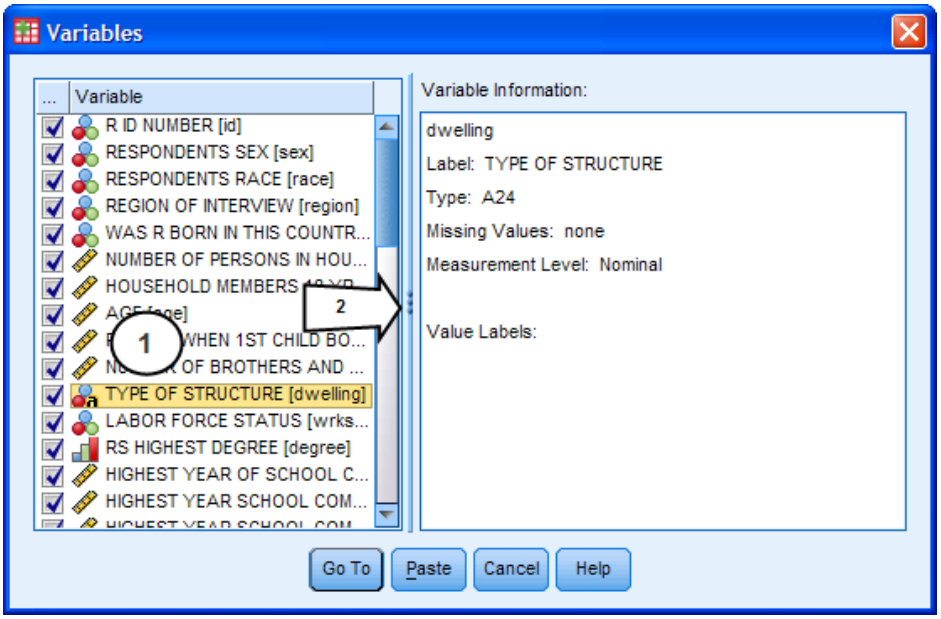
First, we view variable properties for variables *TYPE OF STRUCTURE [dwelling]* and *HAPPINESS OF MARRIAGE [hapmar]*.

Select **Utilities…Variables** (or the  button in the Toolbar).

Detailed Steps:

1) Select *TYPE OF STRUCTURE [dwelling]*
2) Optionally, drag the divider between the Variable and Variable Information boxes to widen the Variable display to display more of the variable label text.

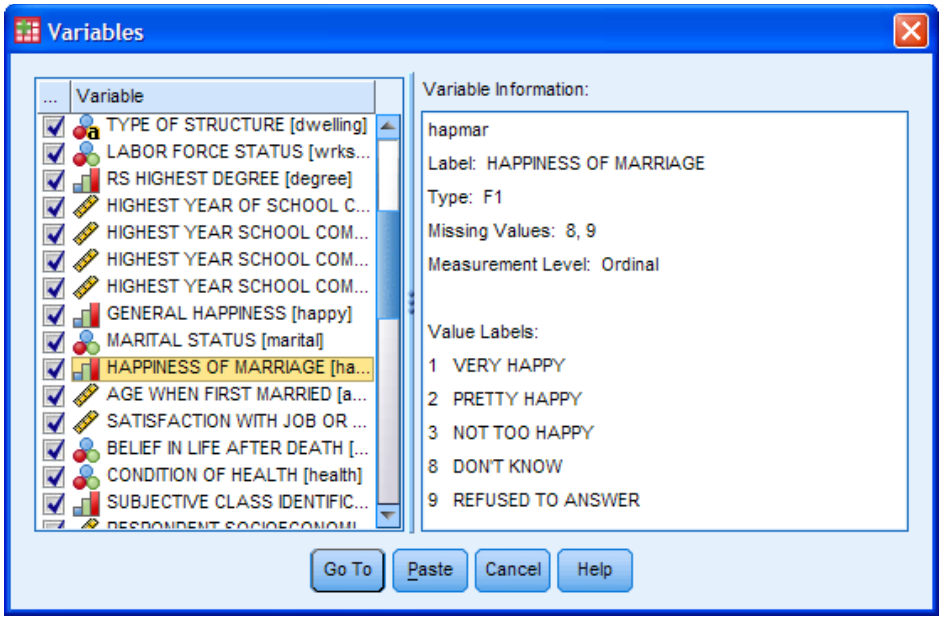
Figure 4.24 Variable Information Dialog for *dwelling*



TYPE OF STRUCTURE [dwelling] is Variable Type A24, an alphanumeric (string) variable, width 24. No value labels or missing values are defined for this variable.

1) Scroll down and select *HAPPINESS OF MARRIAGE [hapmar]*

Figure 4.25 Variable Information Dialog for *hapmar*



HAPPINESS OF MARRIAGE [hapmar] is variable type F1 which means the variable is a numeric variable that will display values of one digit . This is wide enough to accommodate values 1, 2, 3, 8, or 9. Missing values and value labels are defined for this variable.

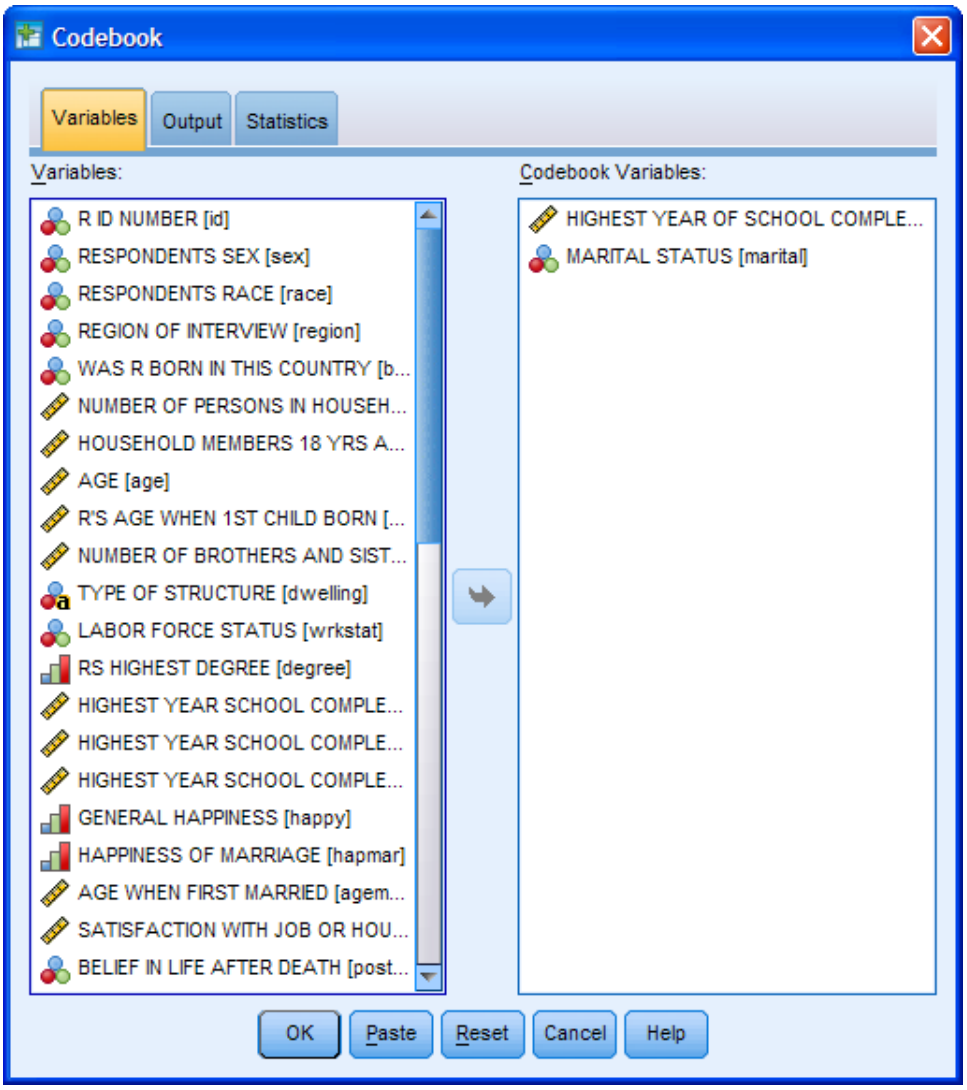
Produce Report Tables of Variable Information

If you want variable information produced in output tables, use either Display Data File Information or Codebook. Since the Codebook tables display summary statistics in addition to the variable properties, we will demonstrate the Codebook procedure.

Detailed Steps:

1) Select **Analyze…Reports…Codebook**
2) Select *HIGHEST YEAR OF SCHOOL COMPLETED [educ]* and *MARITAL STATUS [marital]*

Figure 4.26 Codebook Dialog to Display *educ* and *marital*



Results: Codebook Demonstration

The figure below shows the Codebook tables for *HIGHEST YEAR OF SCHOOL COMPLETED [educ]* and *MARITAL STATUS [marital].*

Figure 4.27 Codebook Tables for *educ* and *marital*

educ

		Value	Count	Percent
Standard Attributes	Position	14		
	Label	HIGHEST YEAR OF SCHOOL COMPLETED		
	Type	Numeric		
	Format	F2		
	Measurement	Scale		
	Role	Input		
N	Valid	2018		
	Missing	5		
Central Tendency and Dispersion	Mean	13.43		
	Standard Deviation	3.079		
	Percentile 25	12.00		
	Percentile 50	13.00		
	Percentile 75	16.00		
Labeled Values	98	DON'T KNOW	3	.1%
	99	REFUSED TO ANSWER	2	.1%

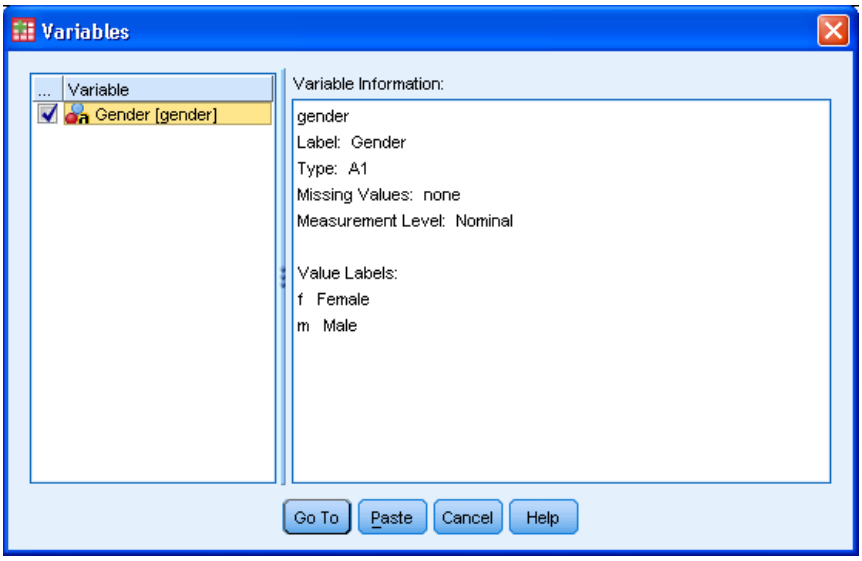
marital

		Value	Count	Percent
Standard Attributes	Position	19		
	Label	MARITAL STATUS		
	Type	Numeric		
	Format	F1		
	Measurement	Nominal		
	Role	Input		
Valid Values	1	MARRIED	972	48.0%
	2	WIDOWED	164	8.1%
	3	DIVORCED	281	13.9%
	4	SEPARATED	70	3.5%
	5	NEVER MARRIED	531	26.2%
Missing Values	9	REFUSED TO ANSWER	5	.2%

The variable *MARITAL STATUS [marital]* is a nominal variable, so counts and percentages are displayed; *HIGHEST YEAR OF SCHOOL COMPLETED [educ]* is a scale variable, so that table includes mean, standard deviation, and the other summary statistics as well as counts and percentage for the labeled value.

Apply Your Knowledge: Viewing Variable Properties

1. Mark all statements that apply to the variable *Gender* as shown in the Variables dialog below.
    a. The variable *gender* is a string variable
    b. The width of the variable *gender* allows the user to type in the value *unkown*
    c. There is no missing value defined for *gender*
    d. The measurement level could be changed to Scale



2. What variable property would need to be changed in order to produce a Codebook table with the mean statistic for the variable salary? Refer to table below.
    a. Remove the value label from the -1 value
    b. Change the measurement level to nominal
    c. Change the measurement level to scale
    d. Just rerun the Codebook procedure

educ

		Value	Count	Percent
Standard Attributes	Position	5		
	Label	Educational Level (years)		
	Type	Numeric		
	Format	F2		
	Measurement	Ordinal		
	Role	Input		
Valid Values	8	8	53	11.2%
	12	12	189	39.9%
	14	14	6	1.3%
	15	15	116	24.5%
	16	16	59	12.4%
	17	17	11	2.3%
	18	18	9	1.9%
	19	19	27	5.7%
	20	20	2	.4%
	21	21	1	.2%
Missing Values	-1	(MISSING)	1	.2%

4.21 *Lesson Summary*

In this lesson we described the variable properties, how to enter them using the Variable View of the Data Editor or the Define Variable Properties feature and how to save the data and variable properties in a PASW Statistics data file for further analysis. We also reviewed options of viewing the variable properties saved in a PASW Statistics data file.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Define, save and view variable properties

And, they should also be able to:

- Describe all of the variable properties
- Define variable properties in the Variable View window
- Define variable properties using the Define Variable Properties dialog
- Save variable properties with data in a PASW Statistics data file
- View variable properties interactively using Variables Utility
- View variable properties in tables using Display Data Dictionary and Codebook procedure

4.22 *Learning Activity*

Supporting Materials

The *employee data.xls* data which contains information about employees at a bank.

The *customer satisfaction 2009.xls* which contains information on customer purchasing behavior and satisfaction is also used.

1. Open *employee data.xls* (Excel file, variable names on first row).

2. Use the Variable View tab in the Data Editor to add variable labels, value labels (where appropriate), missing values (where appropriate) and measurement level. The definitions are listed below.

Variable name	Variable label	Value labels	Missing values	Measurement level
ID	Employee Code			nominal
salbeg	Beginning Salary		0	scale
gender	Employee Gender	0 Male; 1 Female; 9 Unknown	9	nominal
jobtime	Months on the job		0	scale
age	Employee Age (years)		0	scale
salnow	Current Salary		0	scale
edlevel	Educational Level (years)		0	scale

workexp	Work Experience (years)			scale
jobcat	Employment Category	0 Unknown; 1 Clerical; 2 Office Trainee; 3 Security Officer; 4 College Trainee; 5 Exempt Employee; 6 MBA Trainee; 7 Technical	0	nominal
minority	Minority Classification	0 White; 1 Nonwhite; 9 Unknown	9	nominal

3. Run Codebook on all of the variables. Review the codebook tables to make sure that the data values and labels are as you expect. Are there missing cases for those variables where you specified missing values? Note that the tables for the scale variables display summary statistics (mean, standard deviation, and quartiles) while tables for the nominal and ordinal variables display counts and percents.

4. Save the data file (now containing labels, missing values, etc) as *my employee data.sav*.

For those with extra time:

5. Open *customer satisfaction 2009.xls.*

6. Use **Data…Define Variable Properties** to scan and assign value labels to *perf1* to *satisfaction*: 1: --; 2: -; 3: +/-; 4: +; 5: ++, 9 'No Answer'. Assign user-missing to value 9.

7. Run Codebook on these variables to check your work.

8. Save your data as *my customer satisfaction 2009.sav*.

Lesson 5: Working with the Data Editor

5.1 *Objectives*

After completing this lesson students will be able to:

- Use the Data Editor to enter data values, work with Data Editor features and multiple Data Editor windows

To support the achievement of this primary objective, students will also be able to:

- Use features in the Data Editor
- Use Spell Checker
- Copy information from one dataset to another
- Use the Copy Data Properties feature

5.2 *Introduction*

In addition to importing data from a variety of different file formats, PASW Statistics allows the user to enter data directly into the Data Editor. This is useful for entering small amounts of data that have not been previously entered into any file. In this lesson we discuss how to start from scratch and enter data for an imaginary survey into PASW Statistics.

PASW Statistics allows the user to work with multiple datasets open at the same time. Information can be copied from one Data Editor to other open Data Editor windows although this method provides no checking operations. The Copy Data Properties feature allows the user to copy data properties from both open and external datasets, and provides guidance and checks for the operation.

Business Context

The topics in this lesson can help in answering questions such as:

- How can I enter the responses to this short paper survey directly into PASW Statistics?
- How can I copy all of the variable properties from last month's customer file to the new data for the current month?
- How can I view the data values on more than one data file?
- How can I check spelling of labels and data values?

Supporting Materials

The files *census_small.sav*, a PASW Statistics data file and *census_small.xls*, an Excel file. Different formats of data from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

5.3 *Enter Data in the Data Editor*

Data values can be entered directly into the Data Editor Data View tab after you have defined the variable properties in the Variable View tab. This is useful if you have data only on paper. For

example, suppose that you conducted a brief survey of married couples by distributing postcards to a small number of people. The example is shown below.

Table 5.1 Example Survey: Questionnaire

Valentine's Day Survey

ID: _____

Gender: ❒ Male ❒ Female

What is your age? _____ (years)

How happy are you in your marriage? ❒ Very happy ❒ Pretty happy ❒ Not too happy ❒ No answer

Additional Comments _____

In order to enter the data efficiently, a coding scheme must be designed that defines the variable type of the variables, the mapping of the responses to numeric values (where applicable), missing value definitions and measurement level.

Table 5.2 Example Survey: Coding Scheme

Variable Name	Variable Type	Variable Label	Value Labels	Missing Values	Measurement Level
id	Numeric Width: 8 Decimals: 0	Respondent ID Number			Nominal
gender	Numeric Width: 1 Decimals: 0	Gender of Respondent	1 Female 2 Male		Nominal
age	Numeric Width: 3 Decimals: 0	Age of Respondent	999 No Answer	999	Scale
hapmar	Numeric Width: 1 Decimals: 0	Happiness of Marriage	1 Very happy 2 Pretty happy 3 Not too happy 9 No Answer	9	Ordinal
comment	String Width: 50				Nominal

Best Practices

Although PASW Statistics allows string (alphanumeric) values to be entered, we recommend that you assign the responses numeric codes and enter the response text as value labels. This is much less prone to human error as you are entering the text only once for each variable. As well, the saved data file will be smaller in size.

Further Information

PASW Statistics is not designed for data entry of large amounts of data. Instead, consider using PASW® Data Collection Data Entry software which provides robust data entry functionality such as Skip & Fill and logical rule checks, data value checking and double entry verification..

Once a coding scheme is in place, the next step is to define variables and to enter data values.

5.4 *Steps to Enter Data Values in the Data Editor*

Entering data in PASW Statistics starts by giving each variable a name and defining their properties (e.g. value labels) in a new Data Editor window.

Spelling checking can be applied both in Variable View to check variable and value labels and in Data View for text in a string variable. This is especially useful if you have long comment variables and expect to use PASW® Text Analytics for Surveys to analyze the text and create categories.

Entering Data into the Data Editor is accomplished with these steps:

1) Open a new empty Data Editor window by choosing **File…New…Data** from the menu
2) In the Variable View, give each variable a name and define variable properties including variable and value labels, missing value flags, variable type, and measurement level
3) Optionally, spell check the variable and value labels
4) Enter data values in the Data View
5) Optionally, spell check any alphanumeric data values

5.5 *Procedure: Enter Data Values in the Data Editor*

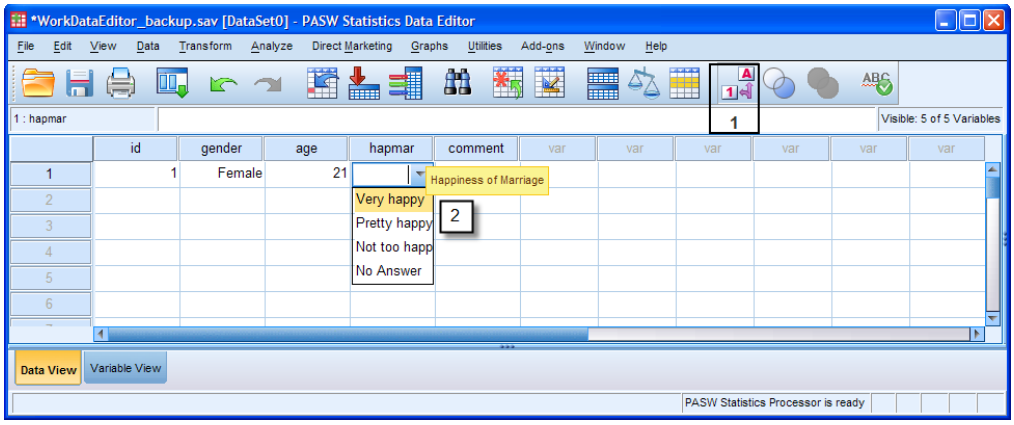
Open an empty Data Editor window from the menus **File…New…Data**. Go to the Variable View tab and enter a variable name for each variable to be entered as well as all other variable property information.

Further Information

Refer to the *Variable Properties* Lesson for detailed descriptions on how to enter variable properties.

Once the variables are defined, move to the Data View tab to begin entering data values. Each defined variable forms a column in the Data View. Enter one case on each row, entering a data value for each variable. If value labels are defined for a variable, you can enter the data value selecting from the value labels. Only values with value labels are listed, which helps to ensure that only valid values in the correct format are entered.

Figure 5.1 Enter Data Using Value Labels



Move to the Data View tab

1) Click the **Value Labels tool button** (Alternatively, select **View…Value Labels** from the menu)
2) In the data cell, select from the list of value labels

Note

Empty cells for numeric variables are assigned the system-missing value displayed as a period in the Data View. Empty cells for string variables are displayed as blank.

5.6 *Procedure: Spell Checker*

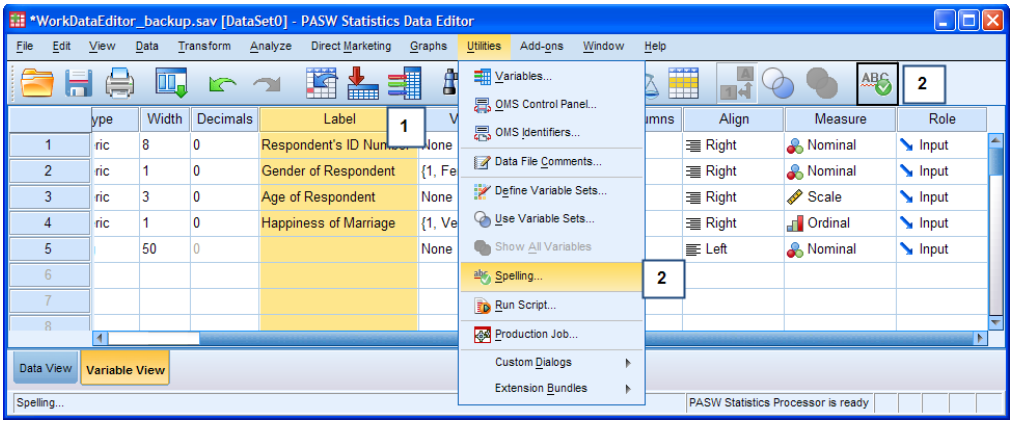
PASW Statistics provides a spell checker to use with

1) Variable labels
2) Value labels
3) String data values

Check Spelling: Variable Labels

Variable labels can be checked for spelling errors on the Variable View by selecting the variable labels that you want to check.

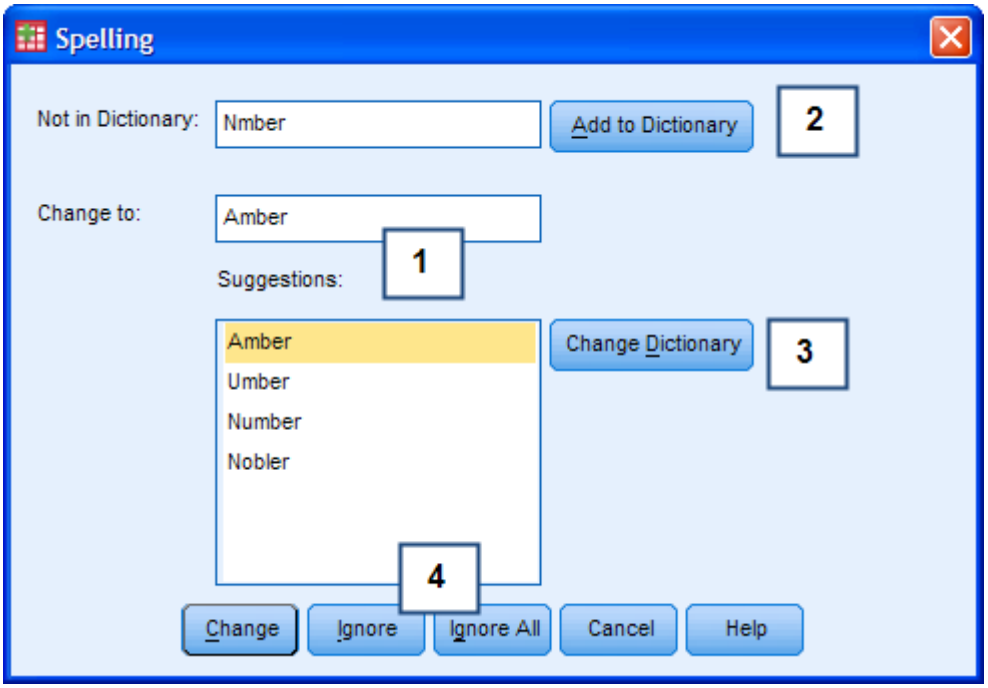
Figure 5.2 Spell Checker: Variable Labels



1) Select the variable labels to check. To select all variables, select the column header "Label"
2) Select the **Spell Checker tool button** (Alternatively, select **Utilities…Spelling**)

Either a message box will popup indicating that no spelling errors were found. Or, a dialog box will appear that tells you what is in error and gives you the opportunity to correct it.

Figure 5.3 Spelling Dialog Box



1) Type the text to change to or select one of the Suggestions:
2) Optionally, **Add to Dictionary** the word not in the dictionary
3) Optionally, **Change Dictionary** to a different language
4) Select an action button to **Change**, **Ignore** once, **Ignore All** occurrences, or **Cancel** the operation

Check Spelling: Value Labels

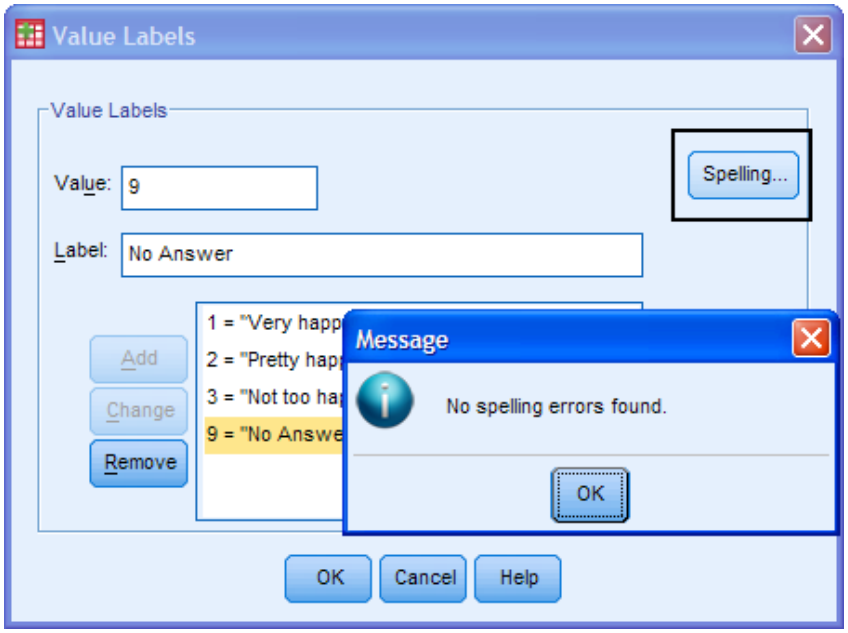
Value labels defined in the Variable View can be checked for spelling by selecting the cell under "Values" for the variable that you want to check.

1) Select the cell in the Values column for the variable's value labels to check. To select all variables' value labels, select the column header "Value"
2) Select the **Spell Checker tool button** (Alternatively, select **Utilities…Spelling**)

Either a message box will popup indicating that no spelling errors were found. Or, the Spelling dialog box (shown above) will appear that tells you what is in error and gives you the opportunity to correct it.

Value labels can also be checked in the Value Labels dialog as they are being entered.

Figure 5.4 Value Labels Dialog: Check Spelling



1) Click the **Spelling** button

Either a message box will popup indicating that no spelling errors were found (as shown). Or, the Spelling dialog box (shown earlier) will appear that tells you what is in error and gives you the opportunity to correct it.

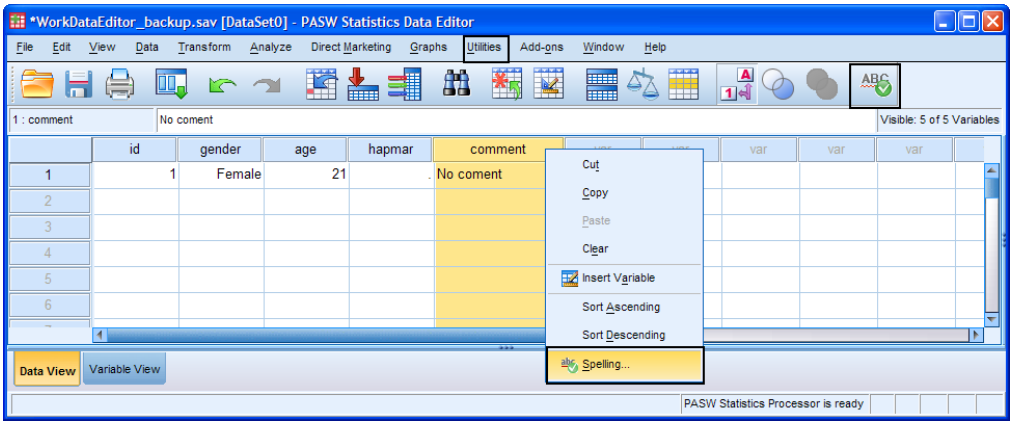
Spell Checking: String Data Values

Data values for string variables can be checked for spelling from the Data View. This assures consistent values before performing analysis.

Note

It is especially useful to check the spelling for long comment fields if you expect to use PASW Text Analytics for Surveys software to analyze the text and create categories.

Figure 5.5 Spell Checker: Data View



1) Select cell for the value to be checked or select the variable name in the column header to check all values of the string variable
2) Select the **Spell Checker tool button** (Alternatively, right-click and select Spelling from the popup list or select **Utitlites…Spelling** from the menu.

As for the labels, either a message box will popup indicating that no spelling errors were found. Or, the Spelling dialog box will appear that tells you what is in error and gives you the opportunity to correct it.

5.7 *Demonstration: Enter Data Values in the Data Editor*

To demonstrate entering data values directly into the Data Editor, we will define the variables and variable properties for the survey displayed at the beginning of this lesson. Then, we will enter the data values below for the first four respondents to the survey.

Table 5.3 Example Survey: Data for Four Respondents

id	gender	age	hapmar	comment
1	Male (2)	27	Very happy (1)	
2	Male (2)	33	Very happy (1)	My second marriage is better than the first
3	Female (1)	41	Pretty happy (2)	I have no comment
4	Female (1)	55	Not too happy (3)	We are in therapy

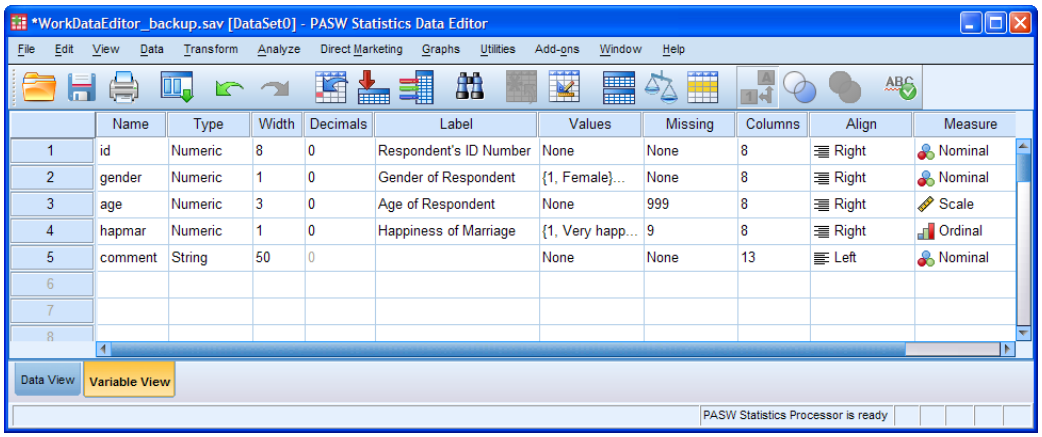
Detailed Steps for Entering Data Values in the Data Editor

1) Enter variable names, ***id***, ***gender, age, hapmar, comment*** to define the 5 variables on the survey
2) Enter the variable properties for each variable as listed in the table at the beginning of this lesson
3) Move to the Data View and select the Value Label  tool button
4) Check spelling of the variable and value labels
5) Enter the data values for the four respondents above
6) Check spelling of the data values for the string variable, **comment**

Results: Entering Data Values in the Data Editor

The Variable View should contain the variable properties as displayed in the figure below.

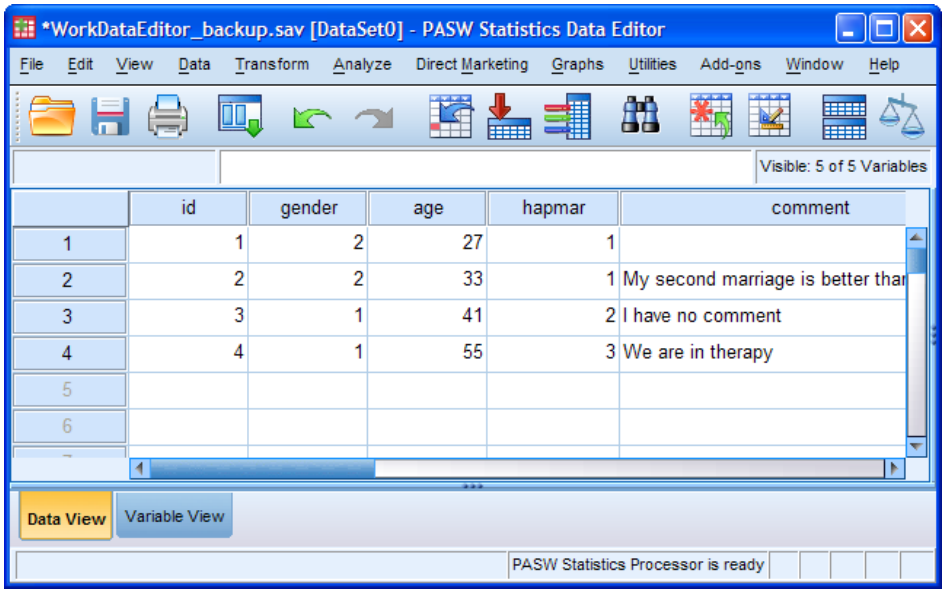
Figure 5.6 Variable View with Survey Variables Defined



7) Click off the Value Labels tool button to display the data values rather than the value labels.

And the data values are shown in the Data View below.

Figure 5.7 Data View with Data Values Entered



Next Steps

After data have been entered, typical data checks are:

- Run Frequencies or Descriptives, with statistics minimum and maximum to check for values outside the range
- Run Codebook to check that all variable properties have been entered correctly and get first descriptive statistics for the data

Apply Your Knowledge: Entering Data Values in the Data Editor

Answer True or False to each of the following statements.

1. If no value is entered for a numeric variable, the value is set to 0 by PASW Statistics.

2. If display of value labels is off in the Data View window, you can enter data for a variable by picking from the value labels of that variable.

3. Entering data in the Data Editor, PASW Statistics gives a warning if you enter a value not defined by a value label.

5.8 *Inserting and Deleting Variables and Cases*

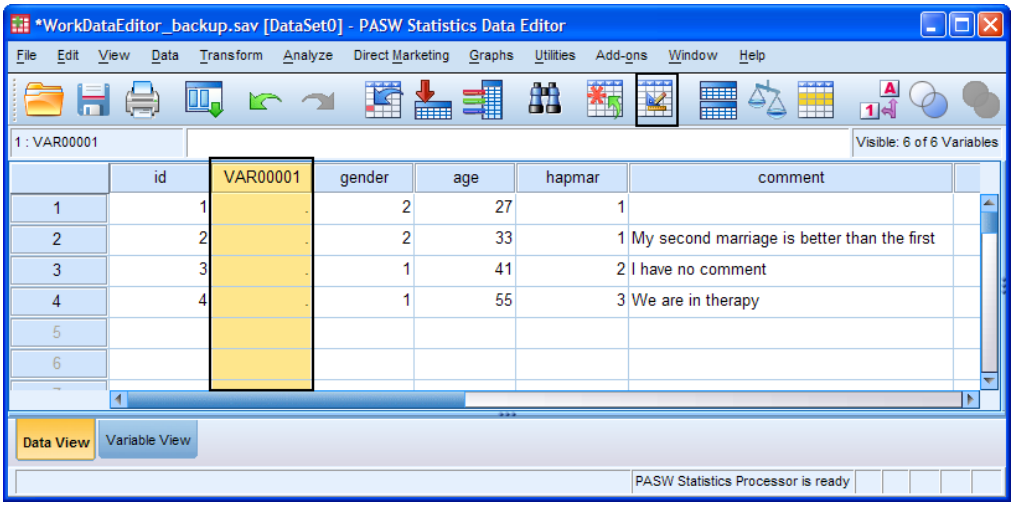
In order to work efficiently in the Data Editor, operations such as inserting, deleting and moving variables and cases are useful. These operations can be performed using the Edit menu or the tool button while in either the Data View or Variable View of the Data Editor

5.9 *Procedure: Inserting and Deleting Variables and Cases*

To insert a variable,

1) Select the column after variable is to be inserted
2) Select **Edit…Insert Variable** from the menu or click  tool button.

Figure 5.8 Data View with Variable Inserted: Default Variable Name



A new column is inserted with a default variable name of the form *VAR* followed by a unique number. Change the variable name and assign other attributes to this variable in the Variable View as we have demonstrated.

To delete a variable,

1) Select the variable
2) Select **Edit…Cut** (copy will be saved in Clipboard) or **Edit…Clear** (copy will not be saved in Clipboard). (Alternatively, right-click on the variable and select the option from the pop-up menu).

Best Practices	Cases also can be deleted in this way; however, it is not recommended. In the lesson *Selecting Cases* we will see better alternatives for case selection.

To insert cases into the middle of the data file,

1) Select the row following the location where you want to insert the case.
2) Select **Edit…Insert Case**s from the menu or click the  tool.

5.10 *View Values for Multiple Variables: Split Screen*

You may have two or more variables whose values you would like to compare but which appear at opposite ends of the Data Editor. The variables and cases in the dataset can be moved by dragging and dropping them to the new position. However, this changes their position in the data file when it is saved.

Rather than moving cases or variables, it may be easier to split the Data View screen so that each section can be scrolled independently. In this way, either the cases or variables of interest can be displayed side-by-side without physically moving them in the Data Editor. The screen can be split by either by case, variable or by both.

5.11 *Procedure: View Values for Multiple Variables: Split Screen*

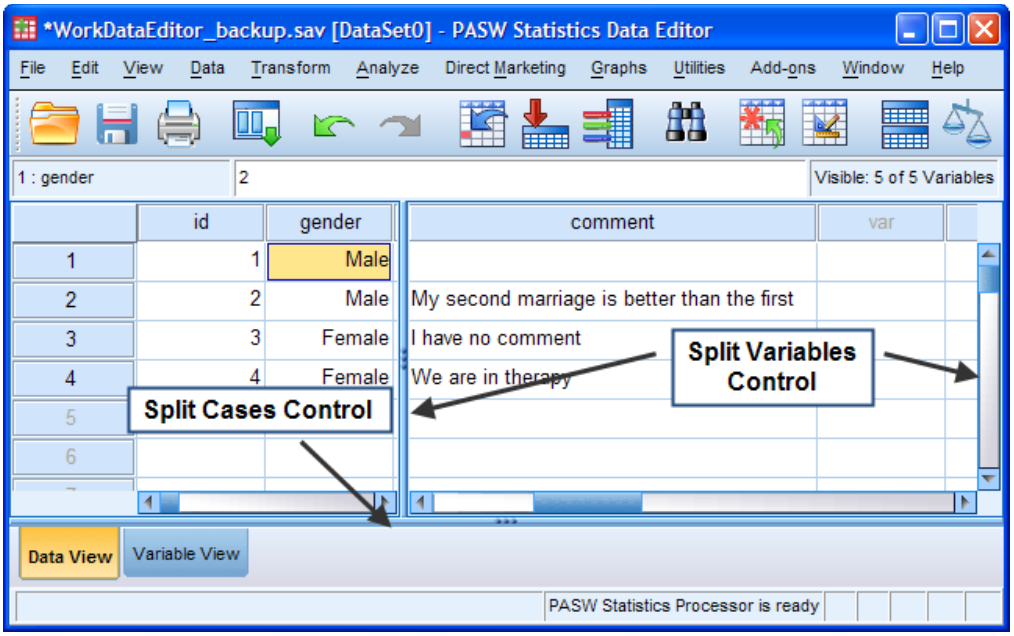
Options to split the screen are:

- To split the screen vertically (between variables) click and drag the grey bar on the right side of the window.
- To split the screen horizontally (between cases) click and drag the grey bar at the bottom of the window.
- Alternatively you can split the screen by selecting the case or variable where you want the split to occur and select **Window…Split** from the menus**.**

Options to remove the split screen:

- Click and drag the grey bar back to the original position
- Select **Window…Remove Split**.

Figure 5.9 Spit Screen Controls



Further Information

Shortcut keystrokes also can be used to navigate the Data Editor. For example, CTRL-END and CTRL-HOME bring you to the end/beginning of the data file. For additional information on all keyboard shortcut keys, see the Help topic *Keyboard Navigation* under *Accessibility* Help.

5.12 *Demonstration: Inserting, Deleting Variables and Split Screen*

To demonstrate these features of the Data Editor, we will: 1) insert a variable, *Region*, before the *Comment* variable, 2) move *age* immediately after the variable *id* and 3) split the screen vertically.

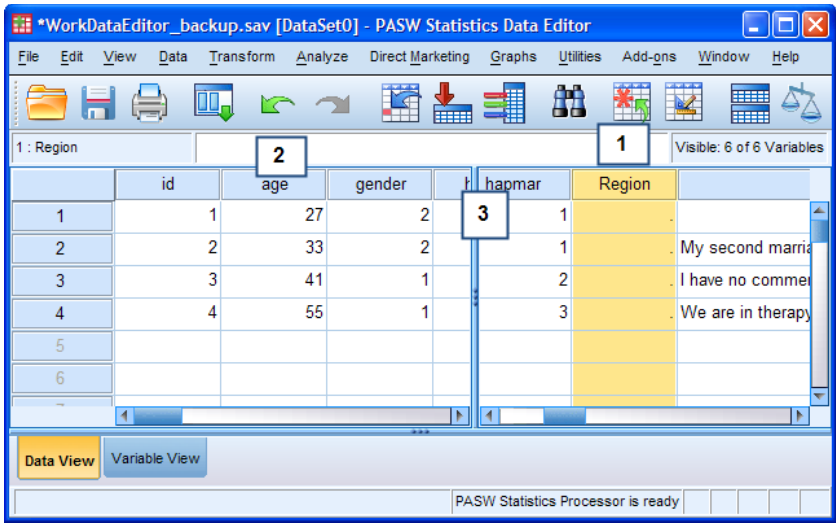
Detailed Steps

1) Select the column header for *comment*
2) Select **Edit…Insert Variable** from the menu or click  tool button
3) Select the column header for *age*.
4) Move to the Variable View tab
5) Replace the variable name for the inserted variable with the name *Region*
6) Drag and drop it immediately after the variable *id*. Note that a red line will appear after the *id* column to indicate the position the variable will be dropped.
7) Move to the Data View tab
8) Drag the vertical split screen bar to the left

Resulting Data Editor

The Data Editor displaying the changes is displayed in the figure below.

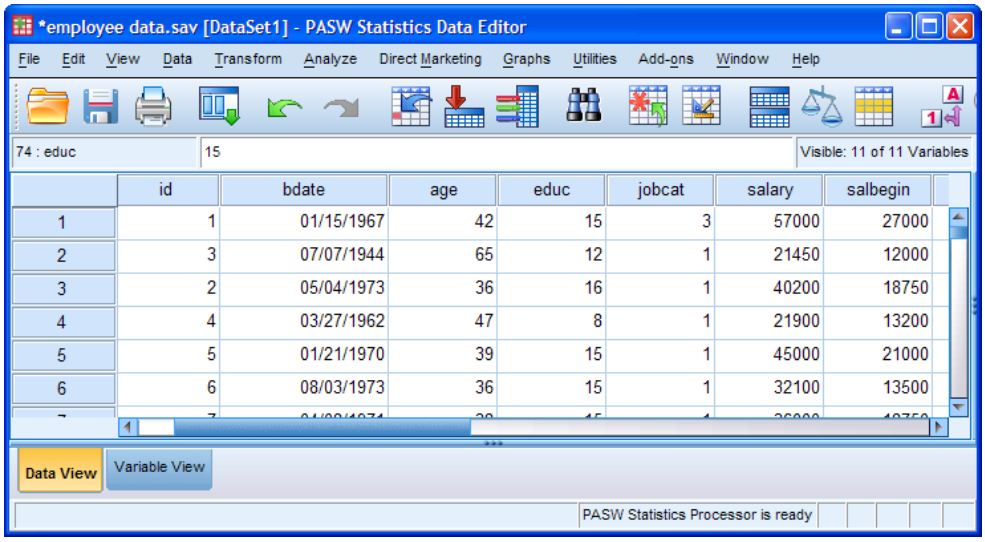
Figure 5.10 Modified Data Editor with Inserted Variable, Moved Variable and Split Screen



Apply Your Knowledge: Insert and Move Variables and Cases, Split Screen

1. True or False. A blank column must be inserted before you can move a variable in the Data Editor.

2. In the dataset below, what is the correct sequence of operations to insert a variable named *location* between the *educ* and *jobcat* variables?
   a. Select *educ*, then Edit…Insert Variable, and rename inserted variable to *location*
   b. Select *educ*, then Edit…Insert Case, and rename inserted variable to *location*
   c. Select *jobcat*, then Edit…Insert Variable
   d. Select *jobcat*, then Edit…Insert Variable, and rename inserted variable to *location*



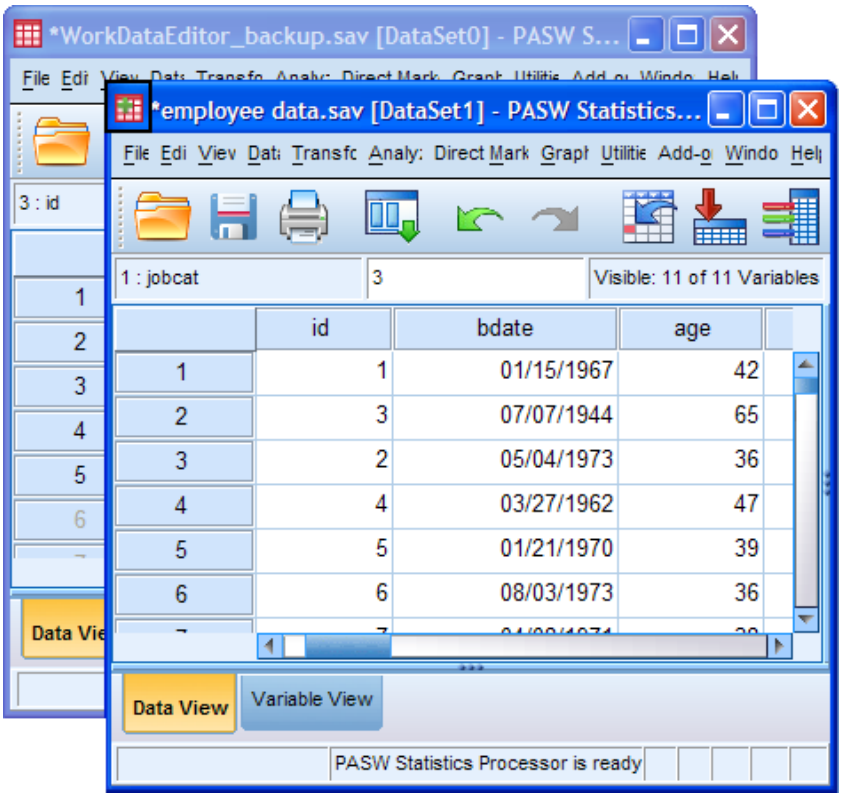
5.13 *Working with Multiple Datasets*

PASW Statistics allows multiple datasets to be open at the same time in multiple Data Editor windows. This feature can be used to view several datasets at one time or to copy variable definitions or data from one file to another. This can be done variable-by-variable or you can copy definitions of all variables in one dataset data to all corresponding variables in another dataset.

Suppose, for example, that you conduct a customer satisfaction survey monthly and that each month the data are collected in an Excel file. The first month you open the Excel file, enter the variable definitions in PASW Statistics and save the data in the PASW Statistics data format (.sav). The second month you open the Excel file for that month and then copy the variable definitions from the file already created in the first month.

If you have more than one dataset open, the green cross in the icon on the window title line designates the **active** dataset. The active dataset is the one which all operations from the menus will be run against.

Figure 5.11 Multiple Data Editor Windows: Active File is *employee_data.sav*



Only variables in the active dataset are available for analysis. You can change the active dataset simply by clicking anywhere in the Data Editor window for the data source that you want to use. Or, you can select the Data Editor window from the Window menu item.

Note

If you notice that PASW Statistics does not support multiple open datasets, check the options. In **Edit...Options**, select the **General tab** and deselect *Open only one dataset at a time.* If this option is checked, the currently open dataset will be closed each time you open a new data source. The default option is to allow multiple Data Editor windows.

5.14 *Procedure: Working with Multiple Datasets*

PASW Statistics must have one Data Editor window at a minimum, so the menu option **File…Close** will be grayed out if you have only one dataset open. If you have multiple datasets open, you can close all but one. If you close the last open Data Editor window, PASW Statistics automatically shuts down, prompting you to save changes first. If you exit from PASW Statistics with multiple Data Editor windows open, PASW Statistics will prompt you to save any changed dataset.

Copy and Paste Operations

You can copy both data and variable definition attributes from one dataset to another dataset in basically the same way that you can copy and paste information within a single Data Editor window.

If you copy and paste data from one dataset to another, the following rules apply:

- If you copy and paste selected data cells from one dataset to another, only the data values are pasted. No variable definition attributes as displayed in the Variable View tab are transferred.
- If you select the variable in the Data Editor Data View and copy and paste this variable to another dataset, both data values and the variable definitions are copied.
- If you select all or part of the variable definitions in the Data Editor Variable View, copy and paste to a variable in another dataset copies only the variable definitions, not the data values.

Variable attributes such as variable type must be the same in both files or the copy and paste operation will improperly define the data. For example, variable (numeric versus string) are not checked before copying and pasting data between files. Likewise when copying and pasting variable definition attributes, you must check that the attributes are appropriate for the variable in the receiving data source.

Warning

Copy and paste operations do NOT check the data structure or variable attributes for each of the files.

If the variable type does not match, data values can be lost. For example, pasting data from a string variable to a numeric variable will cause all of the data values to be converted to system-missing values.

Alternatives are available that perform many of these types of checks on the operation. You should consider using these especially if the datasets are less familiar to you. If the operation involves copying and pasting large amounts of information, these procedures will be more efficient as well.

Best Practices

Use alternative procedures:

- To copy data from one dataset to another dataset, the best option is Merge Files, described in detail in the *Data Management and Manipulation with PASW Statistics* course.
- To copy variable definitions for a number of variables, **Copy Data Properties** feature discussed in the next section is the best option.

5.15 *Demonstration: Multiple Datasets Copy and Paste*

In this demonstration, we assume that PASW Statistics has just been started, so we begin with one empty Data Editor window. We will demonstrate how to copy data properties from one dataset to another for a single variable using copy & paste. First we open the Excel version of the Census data and the PASW Statistics data file. Then copy and paste the variable attributes for *sex* from the open PASW Statistics data file to the dataset from the Excel file.

Detailed Steps for Copy and Paste Demonstration

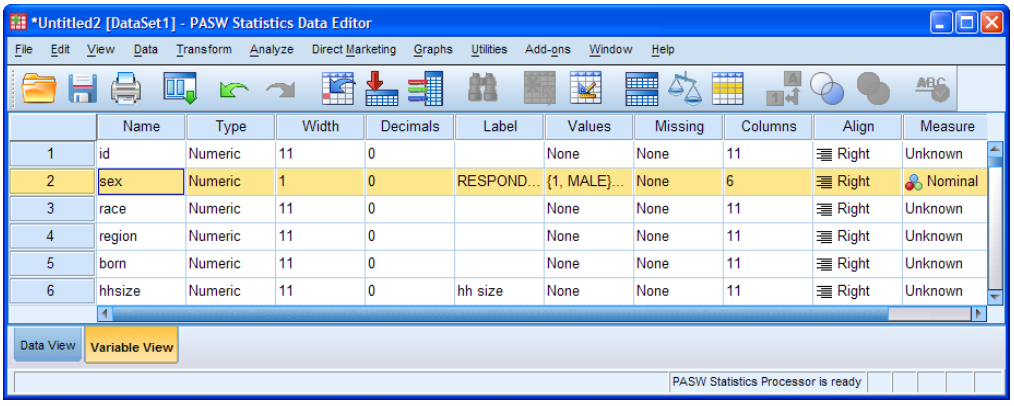
1) Open *census_small.xls*
2) Open *census_small.sav*

3) If necessary, make *gss2008_small.sav [DataSet2]* the active dataset
4) Move to the **Variable View** tab
5) Select the row number for *sex* and copy this variable's properties
6) Select **Untitled2 [DataSet1]**
7) In Variable View tab, select the row number for *sex* and paste to this variable

Results: Copy and Paste Demonstration

All of the variable properties for the variable *sex* have been defined in the dataset read from Excel.

Figure 5.12 Results of Copy and Paste Variable Properties for *sex*



5.16 *Copy Data Properties Feature*

The **Copy Data Properties** feature on the Data menu provides the facility to copy variable properties from one dataset to another. Dictionary information can be copied to the active data file from an open dataset or from an external PASW Statistics data file.

Business Context

Copy Data Properties feature is especially useful if you collect data periodically on the same variables but on different cases.

It can help with operations such as:

- Define variable properties for a patient satisfaction survey that is conducted each month using the same questionnaire.
- Define variable properties for a student record file containing the same variables as the data from last semester.

- Define variables from the data file that captures customer information from Store A to a new data file to collect data from the new Store B

5.17 *Procedure: Copy Data Properties*

Normally, you have to recreate the variable properties information every time you begin work on a new data file. To save time, you can use the **Copy Data Properties** feature to copy variable information from any PASW Statistics data file.

You can use this method to:

- Create new variables in the active data file based on variables in the source data file.
- Apply variable properties from the source data file to variables with the same names in the current active data file.
- Apply selected variable properties (e.g., defined value labels) from one variable in the source data file or the current active data file to multiple target variables in the current active data file.

Variable information from variables in one dataset is copied to matching variable names in another dataset. The variables do not need to be in the same order in both files, and variables that are not present in both files are unaffected.

5.18 *Requesting Copy Data Properties*

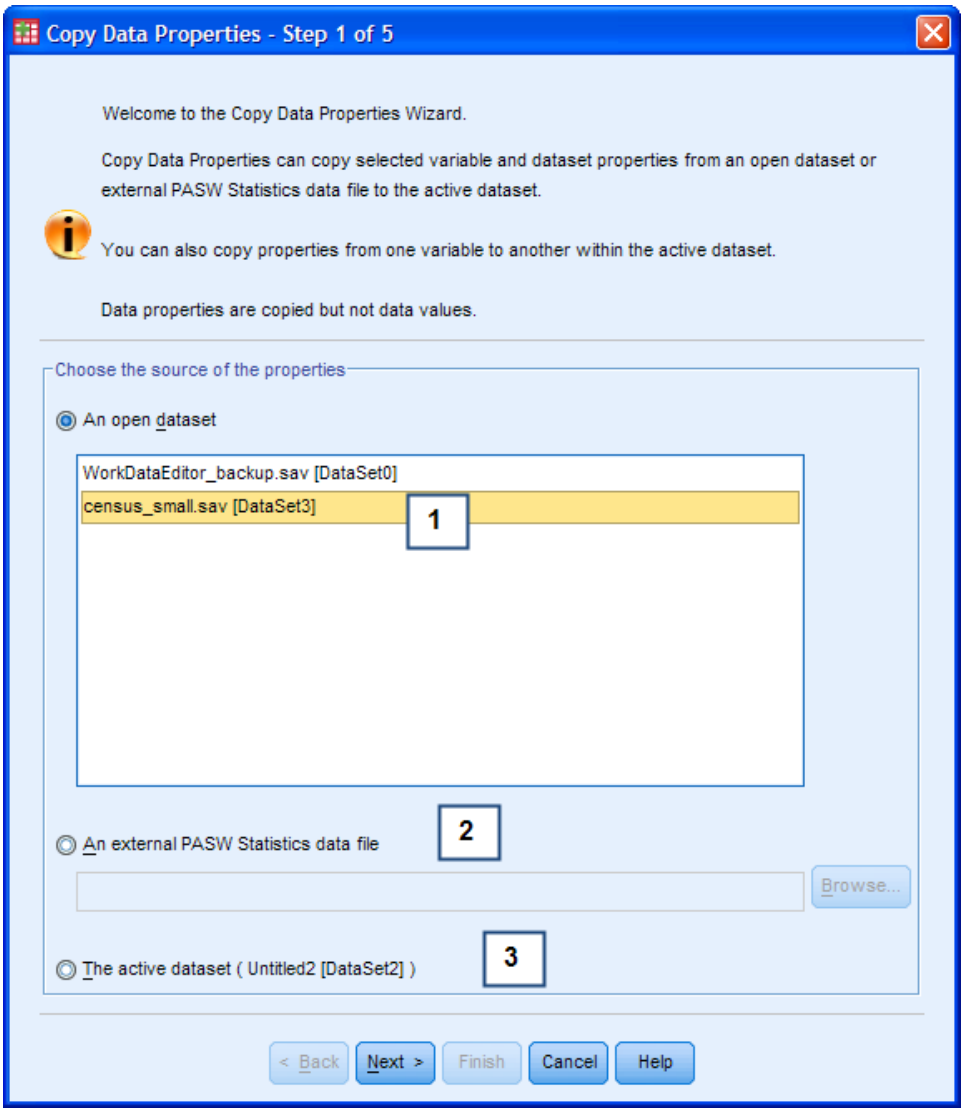
Copy Data Properties is accessed from the Data menu. The Copy Data Properties Wizard has five steps

1) Choose the source dataset for the variable properties
2) Choose the source and target variables. Optionally, create matching variables on the target dataset
3) Choose the variable properties to copy. The default is to copy all of them.
4) Choose dataset properties to copy. These are properties such as File Labels, Multiple Response sets that are associated with the dataset rather than individual variables
5) Execute the operation or write the PASW syntax to a Syntax Editor window

5.19 *Procedure: Copy Data Properties*

In the first step of the Copy Data Properties Wizard, specify the dataset from which the definitions are to be copied.

Figure 5.13 Copy Data Properties, Step 1



The data properties are copied to the active file. Choose the source dataset for the properties:

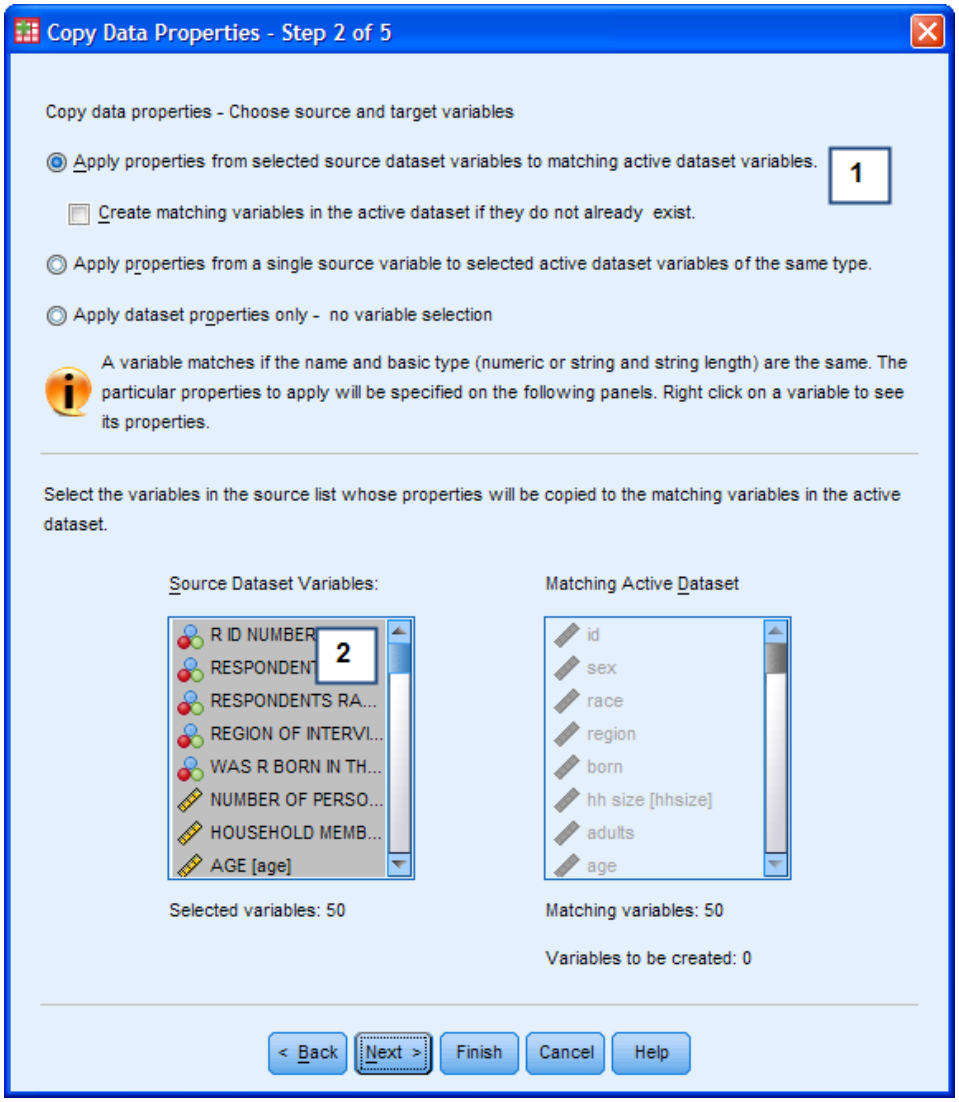
1) Select from the list of open datasets, or
2) Browse to select an external PASW Statistics data file, or
3) Select the active dataset as the source dataset. Use this option to create new variables from one variable.

In step 2, choose the source and target variables and select the option on how to apply the dataset information.

The option *Apply properties from selected source dataset variables to matching active data set variables* is the default. Selecting *Create matching variables in the active dataset if they do not already exist* will update the source list to display all variables in the source data file. If you select source variables that do not exist in the active dataset (based on variable name), new variables will be created in the active dataset with the variable names and properties from the source data file. If the active dataset contains no variables (a blank, new dataset), all variables in the source data file are displayed and new variables based on the selected source variables are automatically created in the active dataset.

The option, *Apply properties from a single source variable to selected active dataset variables of the same type* will apply the selected variable properties from one source variable to the selected variables in the active dataset. For example, this is useful if you want to copy value labels or missing value designations to a series of variables.
The option *Apply dataset properties only--no variable selection* will apply only file properties (for example, a dataset can have a file label and this file label will be copied) to the active dataset. No variable properties will be applied (not available if the active dataset is also the source data file).

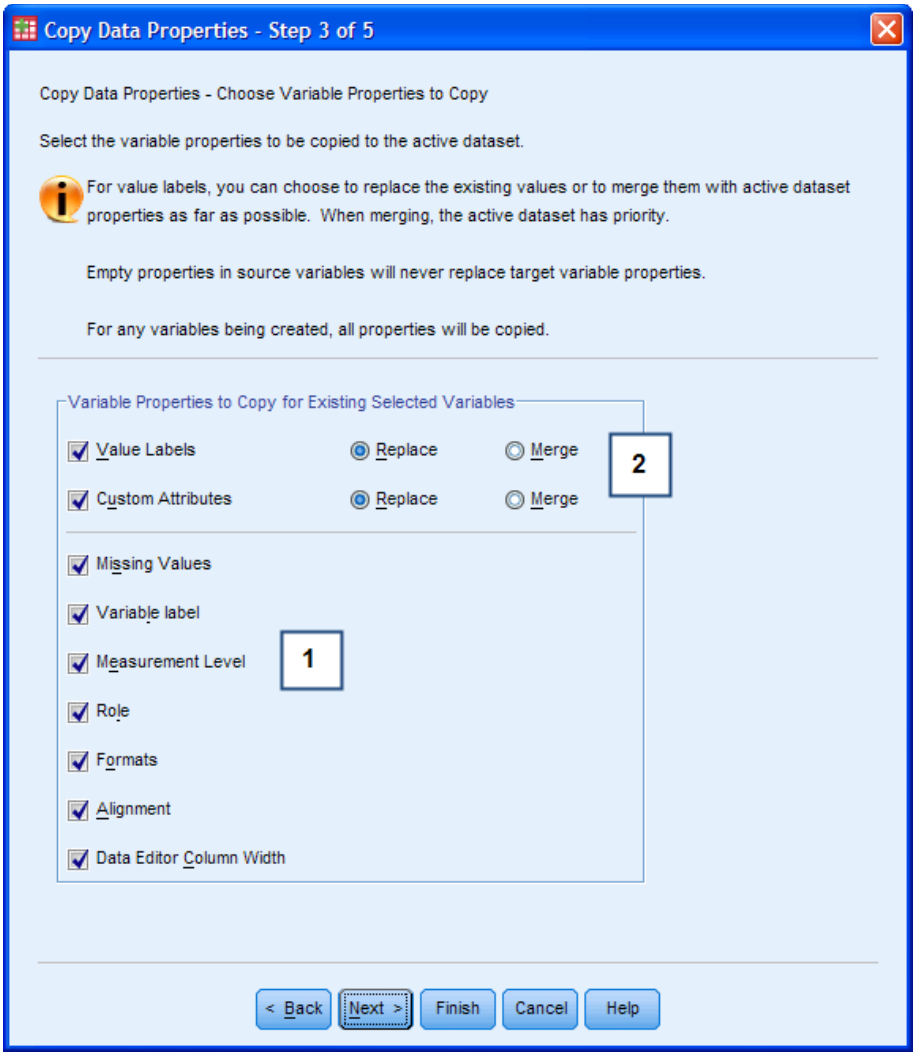
Figure 5.14 Copy Data Properties, Step 2



1) Select one of the three options listed above
2) Select the variables whose properties will be copied from the *Source Dataset Variables*: list

In the third step select the properties to copy. In general, applying dictionary information from one dataset to another will overwrite existing dictionary information. For value labels and custom attributes, you have the choice of replacing or merging the specifications.

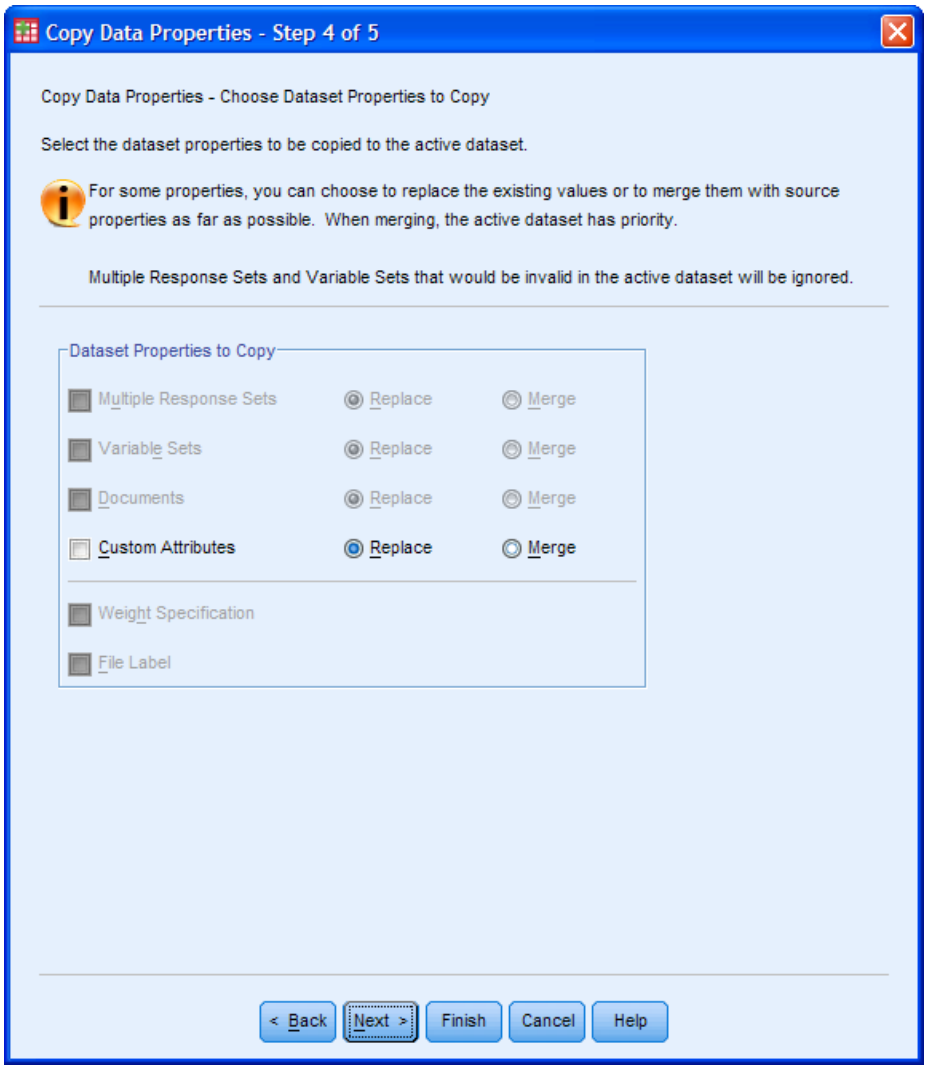
Figure 5.15 Copy Data Properties, Step 3



1) Select the variable properties to copy. By default, all are copied.
2) For variable labels and custom attributes, select whether to *Replace* or *Merge* with any existing ones.

In step 4, select the data properties to be copied.

Figure 5.16 Copy Data Properties, Step 4



Data properties include items such as Multiple Response Sets, Variable Sets, File Labels and other documentation. If the property does not exist in the source data file, it is grayed out in the dialog box.

Further Information

Several data properties including Multiple Response Sets, Variable Sets, and Custom Attributes are discussed in the *Data Management and Manipulation with PASW Statistics* course.

The fifth and final step in the wizard (not shown) allows you to paste the syntax for the Copy Data Properties specifications. If you need to replicate your work in the future, this is a very useful feature that is not available to you when you "cut and paste" information from one Data Editor window to another.

Further Information

For an introduction to using PASW Statistics syntax, refer to the *Syntax Basics* Lesson. More in depth coverage is provided in the *PASW Statistics Syntax I* and *PASW Statistics Syntax II* courses.

5.20 *Demonstration: Copy Data Properties*

We continue our demonstration by using Copy Data Properties to copy the variable properties from all variables in the *census_sample.sav* file to the dataset read from Excel

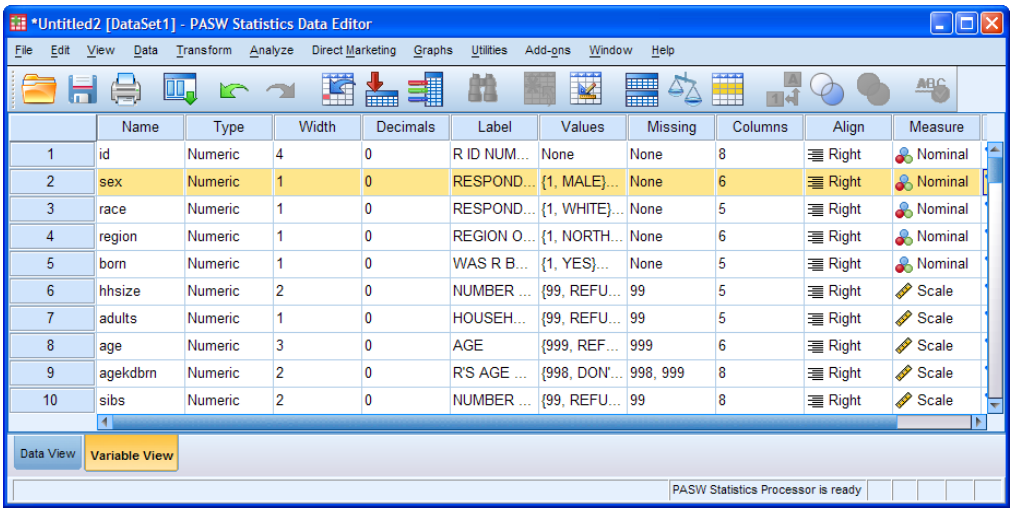
Detailed Steps for Copy Data Properties

1) Make *Untitled2 [DataSet1]* the active dataset
2) Select **Data…Copy Data Properties**
3) In step 1 of the Wizard, select *An open dataset*: and select *census_small.sav [DataSet2]* from the list of open datasets
4) In step 2, choose option *Apply properties from selected source dataset variables to matching active dataset variables*
5) In step 2, select all variables from the source dataset, except *sex*
6) In step , check that all properties will be copied
7) Select **Finish** to execute the specifications

Results Copy Data Properties

The dataset created from reading the Excel file now has variable definitions. This can be checked by running facilities such as Display Data File Information, Codebook, or Frequencies, opening the Variables Information dialog box or reviewing the Variable View information. The figure below shows the Variable View with all of the variable properties define.

Figure 5.17 Copy Data Properties Results



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	4	0	R ID NUM...	None	None	8	Right	Nominal
2	sex	Numeric	1	0	RESPOND...	{1, MALE}...	None	6	Right	Nominal
3	race	Numeric	1	0	RESPOND...	{1, WHITE}...	None	5	Right	Nominal
4	region	Numeric	1	0	REGION O...	{1, NORTH...	None	6	Right	Nominal
5	born	Numeric	1	0	WAS R B...	{1, YES}...	None	5	Right	Nominal
6	hhsize	Numeric	2	0	NUMBER ...	{99, REFU...	99	5	Right	Scale
7	adults	Numeric	1	0	HOUSEH...	{99, REFU...	99	5	Right	Scale
8	age	Numeric	3	0	AGE	{999, REF...	999	6	Right	Scale
9	agekdbrn	Numeric	2	0	R'S AGE ...	{998, DON'...	998, 999	8	Right	Scale
10	sibs	Numeric	2	0	NUMBER ...	{99, REFU...	99	8	Right	Scale

Apply Your Knowledge: Copy Data Properties

1. True or False? If the variable name is the same on both files, the variable type must also be the same in order for Copy Data Properties to copy the variable properties.
2. True or False? A variable that does not appear on both files is dropped during the Copy Data Properties operation.

5.21 *Lesson Summary*

Lesson Objectives Review

Students who have completed this lesson should be able to:

- Use the Data Editor to enter data values, work with Data Editor features and multiple Data Editor windows

And, they should also be able to:

- Use features in the Data Editor
- Use Spell Checker
- Copy information from one dataset to another
- Use the Copy Data Properties feature

5.22 *Learning Activity*

In this exercise you will use the PASW Statistics data file *employee data.sav*.

Supporting Materials

The two files *employee data.sav* and *customer satisfaction 2009.sav* are used in this exercise in addition to a file that you define.

We want to create a data file for employees of company X, with variables:

Variable Name	Variable Type	Variable Label	Values and Labels	Missing Values	Measurement Level
id	Numeric	Employee ID			Nominal
gender	Numeric	Sex of Employee	0 Male 1 Female		Nominal
jobcat	Numeric	Job category	1 Clerical 2 Custodial 3 Manager 4 CEO		Nominal
salary	Numeric	Salary	-1 missing	-1	Scale

1. Create a dataset with these variable names, but do not define variable label, value labels, missing values and measurement level.

2. Enter data for some employees as below:

id	gender	jobcat	salary
1	1	4	200000
2	1	1	50000
3	0	3	100000

3. Use the **Copy Data Properties** feature to copy the variable definitions from *employee data.sav* to the dataset you just created.
   Explain why the value labels for *gender* were not copied.

For those with extra time:

4. Open *customer satisfaction 2009.xls* and use **Data…Copy Data Properties** to copy the data definitions from *customer satisfaction 2008.sav* to the 2009 data. Save the data as *my customer satisfaction 2009.sav*.

Lesson 6: Summarizing Individual Variables

6.1 *Objectives*

After completing this lesson students will be able to:

- Summarize individual variables using tables and graphs

To support the achievement of this primary objective, students will also be able to:

- Define levels of measurement
- Use the Frequencies procedure to produce tables and charts appropriate for nominal variables
- Use the Frequencies procedure to produce tables and charts appropriate for ordinal variables
- Use the Frequencies and Descriptives procedures to produce tables and charts for scale variables

6.2 *Introduction*

As a first step in analyzing your data, you must review the overall distribution of the individual variables and check for any unusual or unexpected values. You often want to examine the values that occur in a variable and the number of cases for each. For some variables, you can study the distribution of the variable by examining simple summary measures, including minimum and maximum values for the range. Frequently used summary measures describe the central tendency of the distribution, such as the arithmetic mean, and dispersion, the spread around the central point.

The **Level of Measurement** of a variable determines the appropriate tables, statistics and charts used to describe the data. Intuitively, you can see that while a table of counts and percentages for each category of marital status is a useful description of the data, a mean or average for years of age would be more useful for that variable. The concept of measurement level formalizes these differences.

Business Context

There are a number of reasons for performing single variable analyses.

- Check the data for invalid or unusual values for each variable. Unusual values would be displayed in a frequency table.
- Establish base rates for the population sampled. For example, what percentage of our customers is satisfied with services this year?
- Studying a frequency table containing many categories might suggest ways of collapsing groups for a more succinct and statistically appropriate table.

Supporting Materials

The *Census_Small.sav* PASW Statistics data file is used in this lesson. These data are a subset of demographic and attitudinal variables from a survey conducted in 2008 of a sample of the general population.

6.3 *Definition: Levels of Measurement*

Levels of measurement refer to the coding scheme or assignment of codes or numbers for each variable. Different statistical measures are appropriate for different types of variables, depending on the level of measurement. There are four levels of measurement which combine into two main types: **categorical** and **scale**. Each level of measurement is defined by certain properties.

Categorical

Categorical variables have a limited number of distinct values or categories (for example, gender or marital status). Categorical variables can be string (alphanumeric) or numeric. Numeric variables use numeric codes to represent categories (for example, for gender: 1 = male and 2 = female). There are two basic types of categorical data:

- **Nominal:** Categorical data where there is no inherent order to the categories. Numbers are arbitrarily assigned to distinguish between different properties. For example, a job category of "sales" isn't higher or lower than a job category of "marketing" or "research." Other examples of nominal variables are marital status, region of residence, and ethnic group identification.
- **Ordinal:** Categorical data where there is a meaningful order or rank to the categories, but there isn't a measurable distance between categories. For example, there is an order to the values high, medium, and low, but the "distance" between the values can't be calculated. Other examples of ordinal variables are attitudinal questions with categories such as Strongly Disagree (value 1), Disagree (value 2), Agree (value 3), and Strongly Agree (value 4); or variables such as age or income coded into categories representing ranges of values.

Scale

Scale variables have data values that indicate both the order of values and the distance between values. For example, a salary of 72,000 is higher than a salary of 52,000, and the distance between the two values is 20,000. These types are also referred to as quantitative or continuous data. There are two types of scale data:

- **Interval:** The distance between adjacent points is the same throughout the scale, unlike ordinal scales. For example, with a variable such as age in years, the difference between 20 and 21 (1 unit, i.e., 1 year) is equal to the difference between 45 and 46. In other words, they have equal intervals between points on the scale.
- **Ratio:** Ratio data have all the properties of interval variables with the addition of a true zero point representing the complete absence of the property being measured. Thus, in addition to differences between values, ratios of values are interpretable. For example, temperature measured in degrees Fahrenheit is measured on an interval scale, because the zero point does not represent a "true" zero. Therefore, this variable is interval level. However, the variable, number of visits to the theatre per year, is a ratio variable because 0 indicates no visits. Ratios can be calculated (e.g., 4 visits to the theatre represent twice as many visits as 2).

The difference between these two types of scale variables is rarely important in statistical analysis. In fact, PASW Statistics combines both interval and ratio variables as **scale** level in the measurement level attribute.

An icon indicating the measurement level is displayed preceding the variable name or label in the variable lists of all dialog boxes. The next table summarizes the level of measurement, shows the most common icons used for the measurement levels, lists descriptive statistics and charts appropriate for each, and the procedures in PASW Statistics that can be used to produce the summaries.

Table 6.1 Summary of Measurement Level, Icons, Descriptive Statistics and Graphs

	NOMINAL	ORDINAL	SCALE
Definition	Unordered Categories	Ordered Categories	Metric/Numeric Values
Examples	Eye color, gender, marital status	Satisfaction ratings, age bands	Income, height, weight
Icon for numeric variable			
Icon for string variable		Not Applicable	Not Applicable
Descriptive Statistics	Percentage, Mode	Percentage, Mode, Median	Median, Mean, Min/Max/Range, Standard Deviation
Chart	Pie or Bar	Bar or Pie	Histogram
PASW Statistics Procedures	Frequencies Chart Builder	Frequencies Chart Builder	Frequencies, Descriptives Chart Builder

Note

Special data types, such as Date and Time variables, have distinct icons not shown in this table.

Further Information

See the lesson *Variable Properties* for how to define the measurement level attribute.

Apply Your Knowledge

1. What is the level of measurement of a variable about product preference for a newly introduced computer model, with the response choices of “Dislike,” “Both like and dislike,” and “Like”?
    a. Nominal
    b. Ordinal

c. Interval

2. What is the level of measurement of a variable measuring preference for season of the year, with categories of 1=Spring, 2=Summer, 3=Fall, and 4=Winter?
   a. Nominal
   b. Ordinal
   c. Interval

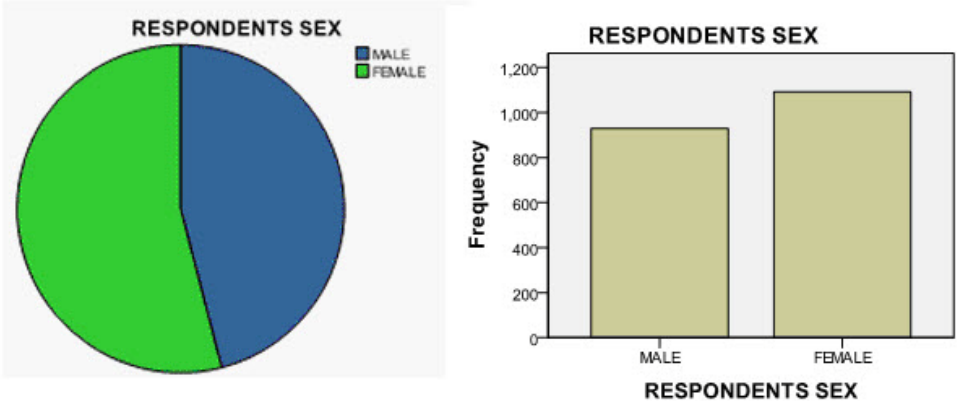
6.4 *Summarizing Nominal Variables*

For nominal data, the typical summary measure is the number or percentage of cases in each category. We can also use the mode, a measure of central tendency, which indicates the category with the highest frequency.

A nominal variable can be displayed graphically in either a pie chart or bar chart.

- A pie chart displays the contribution of parts to a whole. Each slice of a pie chart corresponds to a group that is defined by a single grouping variable.
- A bar chart displays the count for each distinct value or category as a separate bar, allowing you to compare categories visually.

Figure 6.1 Examples of a Pie Chart and a Bar Chart For Gender



The **Frequencies** procedure on the *Analyze ...Descriptive Statistics* menu is the basic procedure to display counts and percentages for each of the categories of a single variable. We will illustrate the features of **Frequencies** as we create output for nominal variables. These same features will be used for ordinal and interval variables.

6.5 *Requesting Frequencies*

Requesting **Frequencies** for nominal variables is completed with these steps:

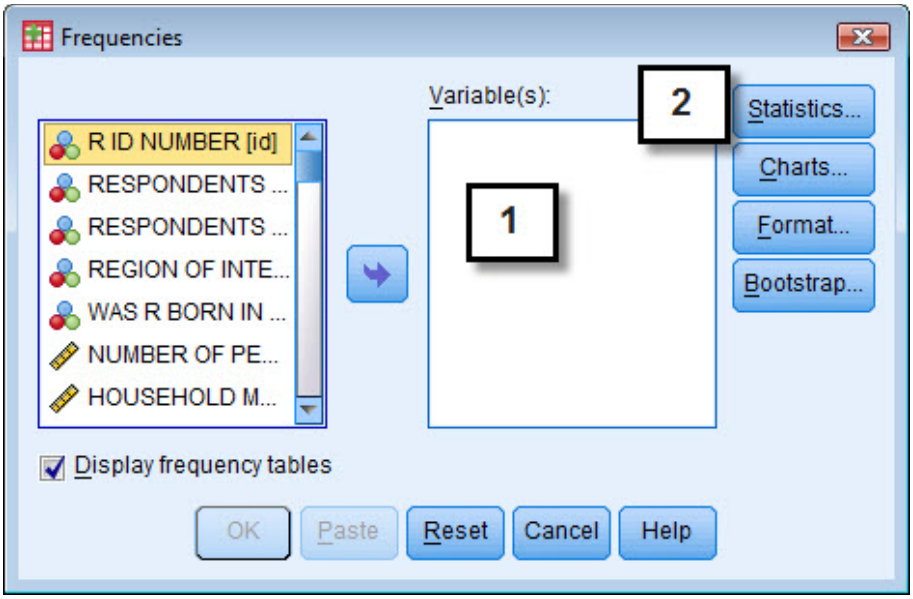
1) Select one or more variables to analyze
2) Request appropriate graphs
3) Request appropriate statistics

6.6 *Procedure: Frequencies for Nominal Variables*

Frequencies is accessed from the *Analyze…Descriptive Statistics…Frequencies* menu choice. In the opening dialog box

1) Place one or more variables in the Variable(s) box
2) Select the *Statistics* button to specify statistics

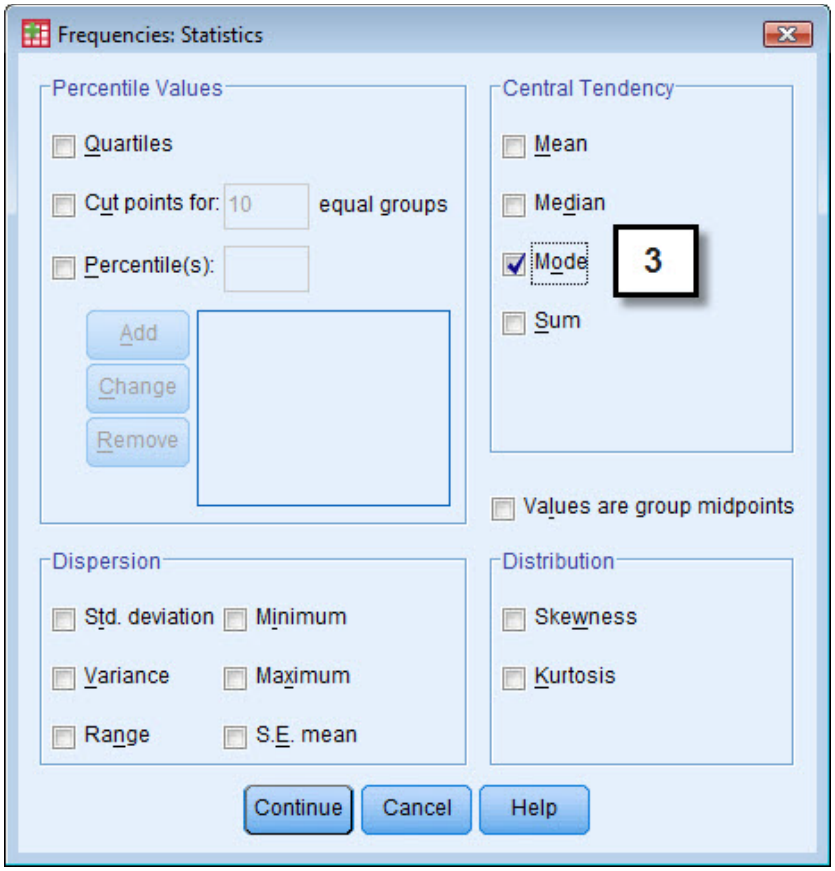
Figure 6.2 Frequencies Main Dialog



The Statistics subdialog provides many statistics, but it is critical that you only request those appropriate for the level of measurement of the variables you placed in the Variable(s) box. For nominal variables, only the mode is suitable.

3) Select the *Mode* for a nominal variable

Figure 6.3 Frequencies Statistics Subdialog

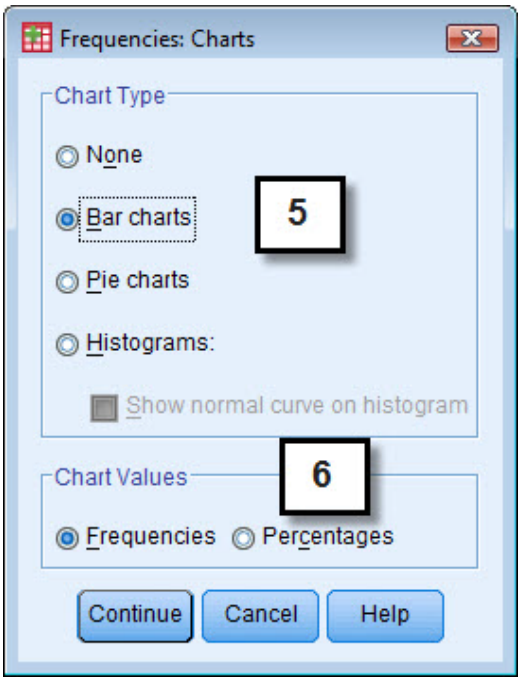


4) Select the *Charts* button to request graphs

The Charts subdialog box has options for pie charts and bar charts. Either type of chart is acceptable for a nominal variable. Charts can be built using either counts or percentages. Normally, percentages are a better choice.

5) Select a chart type
6) Select a Chart Value

Figure 6.4 Frequencies Charts Subdialog



The Formatting subdialog (not shown here) has several options for customizing the frequency table. For example, the table can be arranged according to the actual values in the data or according to the count (frequency of occurrence) of those values, and the table can be arranged in either ascending or descending order by either of these choices.

Tip

For a nominal variable (where the order of the categories is arbitrary) sorting the table and graph descending on counts makes it easy to see what the main categories are.

6.7 *Demonstration Summarizing a Nominal Variable*

We will use the data file *census_small.sav* in these examples. For a demonstration with a nominal variable, we will use *MARITAL STATUS [marital]* and request the mode and a bar chart with percentages.

Detailed Steps for Summarizing a Nominal Variable

1) Place the variable ***marital*** in the Variable(s) box
2) Select the **Statistics** button
3) Select **Mode**
4) Select the **Charts** button
5) Select **Bar charts** in the Chart Type area
6) Select **Percentages** in the Chart Values area

Results for Summarizing a Nominal Variable

The first table produced in the Viewer is the table labeled Statistics. This table shows the number of cases having a valid value on *marital* (2018), the number of cases having a (user- or system-) missing value (5) and the Mode (1). The mode—the category with the highest frequency—is listed by its numeric value, not the value label. The value of 1 is assigned to married people, so they are the largest group.

Figure 6.5 Statistics Table for Marital Status

Statistics

MARITAL STATUS

N	Valid	2018
	Missing	5
Mode		1

The second table shows the frequencies and percentages for each category. The columns are defined as:

- **Frequency:** The count of number of cases in each category
- **Percent:** The percentage based on the total number of cases (here 2023)
- **Valid Percent:** The percentage based on non-missing cases (here 2018)
- **Cumulative Percent:** The percentage of non-missing cases with values less than or equal to that row's value.

The table confirms that most people are married. Since there are almost no missing data for marital status, the percentages in the Percent column and in the Valid Percent column are almost the same.

Figure 6.6 Frequency Table for Marital Status

MARITAL STATUS

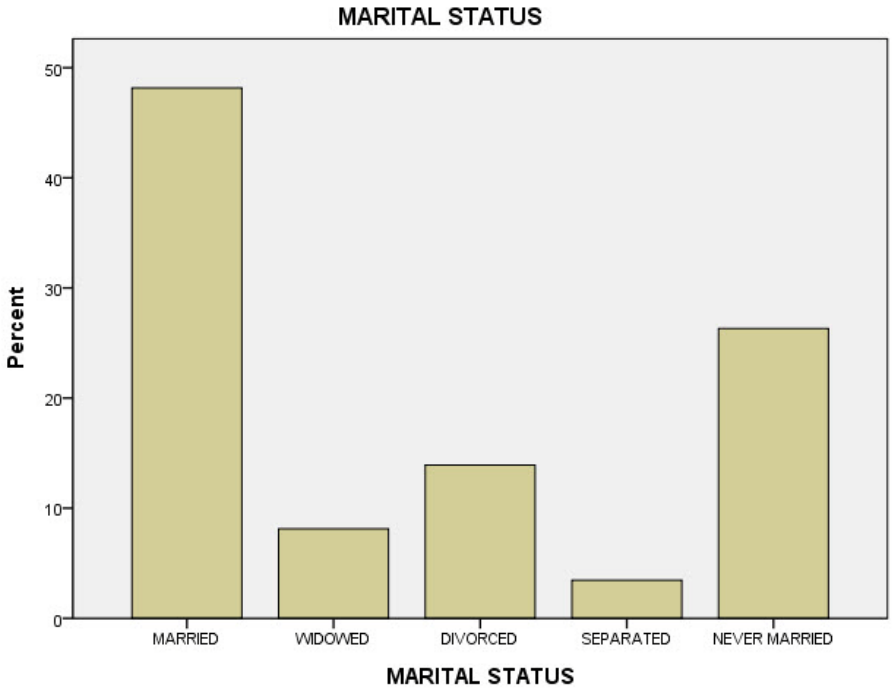
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MARRIED	972	48.0	48.2	48.2
	WIDOWED	164	8.1	8.1	56.3
	DIVORCED	281	13.9	13.9	70.2
	SEPARATED	70	3.5	3.5	73.7
	NEVER MARRIED	531	26.2	26.3	100.0
	Total	2018	99.8	100.0	
Missing	REFUSED TO ANSWER	5	.2		
Total		2023	100.0		

Note

For nominal variables, the cumulative percent may or may be an appropriate statistic. For marital status, one could say, for example, that 56.3% of the respondents are married or widowed. This may or may not be substantively useful.

The last item in the Viewer is the bar chart. In agreement with the mode and Frequencies table, the category MARRIED has the highest frequency/percent.

Figure 6.7 Bar Chart for Marital Status



Apply Your Knowledge

1. True or false? The Cumulative Percent column in a Frequencies table includes percentages for missing values?

2. In the Frequencies table shown below, which category is the mode?
   a. YES
   b. NO

BELIEF IN LIFE AFTER DEATH

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	YES	1455	71.9	81.4	81.4
	NO	332	16.4	18.6	100.0
	Total	1787	88.3	100.0	
Missing	DON'T KNOW	231	11.4		
	REFUSED TO ANSWER	5	.2		
	Total	236	11.7		
Total		2023	100.0		

6.8 *Summarizing Ordinal Variables*

The number and percentage of cases in each category is the most useful summary for most ordinal data. In addition, the median (the value above and below which half the cases fall) may also be a useful summary measure if there is a large number of categories. Graphically, you can display the frequencies or percentages in a bar chart or pie chart, as with nominal variables.

As for nominal variables, the Frequencies procedure is the basic procedure to display counts and percentages for each of the categories of a single ordinal variable.

6.9 *Demonstration: Summarizing an Ordinal Variable*

As an example of an analysis for an ordinal variable, we will run a Frequencies for the variable *SUBJECTIVE CLASS IDENTIFICATION [class].* This variable is coded into four categories: 1 = LOWER CLASS; 2 = WORKING CLASS; 3 = MIDDLE CLASS; 4 = UPPER CLASS.

Detailed Steps for Summarizing an Ordinal Variable

1) Place the variable ***class*** in the Variable(s) box
2) Select the **Statistics** button
3) Select **Median**
4) Select the **Charts** button
5) Select **Bar charts** in the Chart Type area
6) Select **Percentages** in the Chart Values area

Note

The mode, the data value of the category containing the most cases, would also be appropriate for an ordinal variable. Since we can also determine this from the frequency table, we won't request that statistic.

Results for Summarizing a Nominal Variable

The value of the median is displayed in the Statistics table along with the number of valid and missing cases. The median value for the 2007 valid cases is 2, which represents the WORKING CLASS category.

Figure 6.8 Statistics Table for Subjective Class Identification

Statistics

SUBJECTIVE CLASS IDENTIFICATION

N	Valid	2007
	Missing	16
Median		2.00

The frequencies table for subjective class identification is next.

Looking at the Cumulative Percent column in the Frequency table, we observe that the 50th percentile (or mid-point) falls in the WORKING CLASS category. We also observe that the majority of the cases are in the middle two categories, WORKING CLASS and MIDDLE CLASS. The mode is WORKING CLASS. And, we can see that the 16 missing cases were divided into two categories: 11 respondents who answered DON'T KNOW and 5 who REFUSED TO ANSWER.

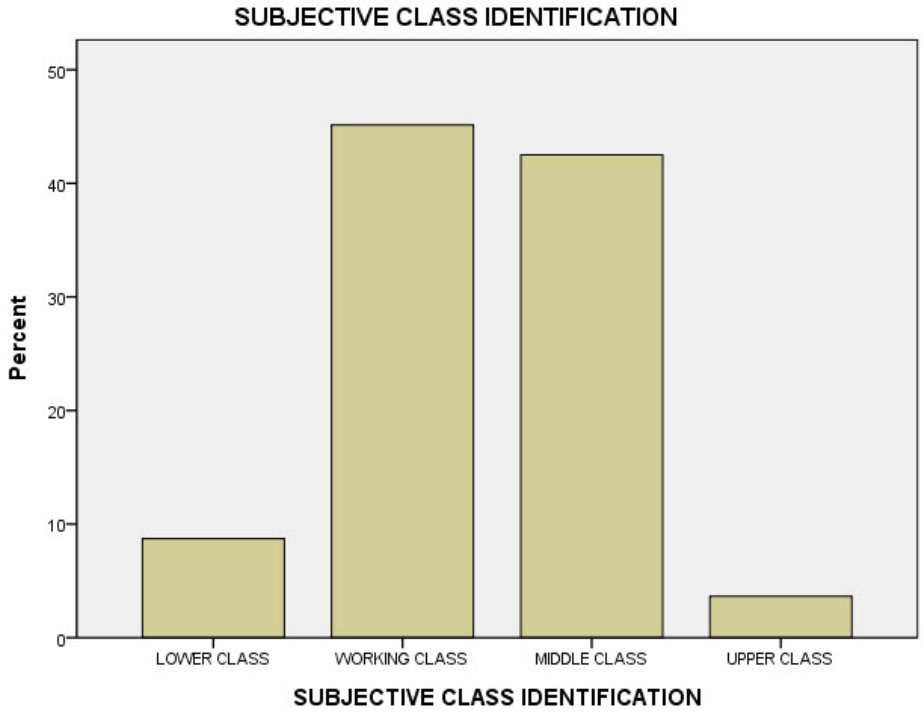
Figure 6.9 Frequencies Table for Subjective Class Identification

SUBJECTIVE CLASS IDENTIFICATION

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	LOWER CLASS	175	8.7	8.7	8.7
	WORKING CLASS	906	44.8	45.1	53.9
	MIDDLE CLASS	853	42.2	42.5	96.4
	UPPER CLASS	73	3.6	3.6	100.0
	Total	2007	99.2	100.0	
Missing	DON'T KNOW	11	.5		
	REFUSED TO ANSWER	5	.2		
	Total	16	.8		
Total		2023	100.0		

The bar chart summarizes the distribution that we observe in the frequency table. Note that it is difficult to determine the median value from this chart. Thus, both displays are important to our understanding of the data. And for those who are visually-oriented, the bar chart provides an easy method to see the distribution without the details in the Frequency table.

Figure 6.10 Bar Chart of Subjective Class Identification



Best Practice

Sorting the categories on descending/ascending counts (which was useful for nominal variables) will disturb the natural order of categories and so is typically not recommended for an ordinal variable.

6.10 *Summarizing Scale Variables*

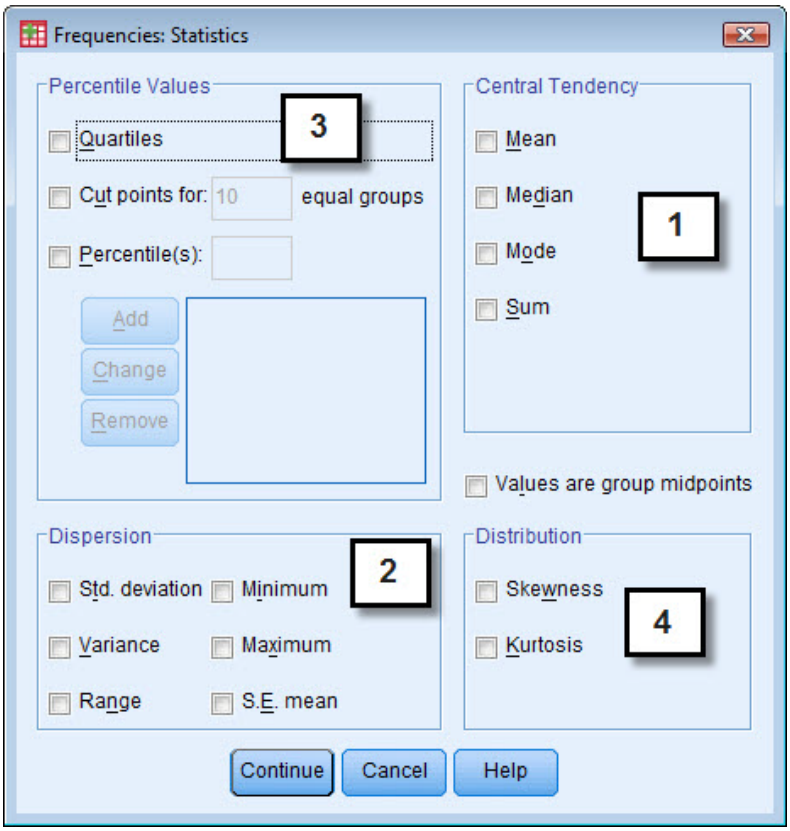
Several summary statistics are appropriate for scale variables. The statistics can be divided into those summarizing the central tendency and those measuring the amount of variation (or spread) in the data:

- **Measures of central tendency.** The most common measure is the mean (arithmetic average).
- **Measures of dispersion.** Statistics that measure the amount of variation or spread in the data include the standard deviation, minimum, and maximum.

The statistics in the Statistics subdialog are organized into groups.

1) **Central Tendency.** The mean, median, mode and sum are included in this group.
2) **Dispersion.** Measures of dispersion, in addition to the standard deviation and minimum and maximum, also include range, variance, and standard error of the mean (S.E. mean).
3) **Percentile Values.** You can request quartiles, or specify cut points for percentiles, or even an exact percentile value.
4) **Distribution.** There are two measures of the shape of the distribution, skewness and kurtosis.

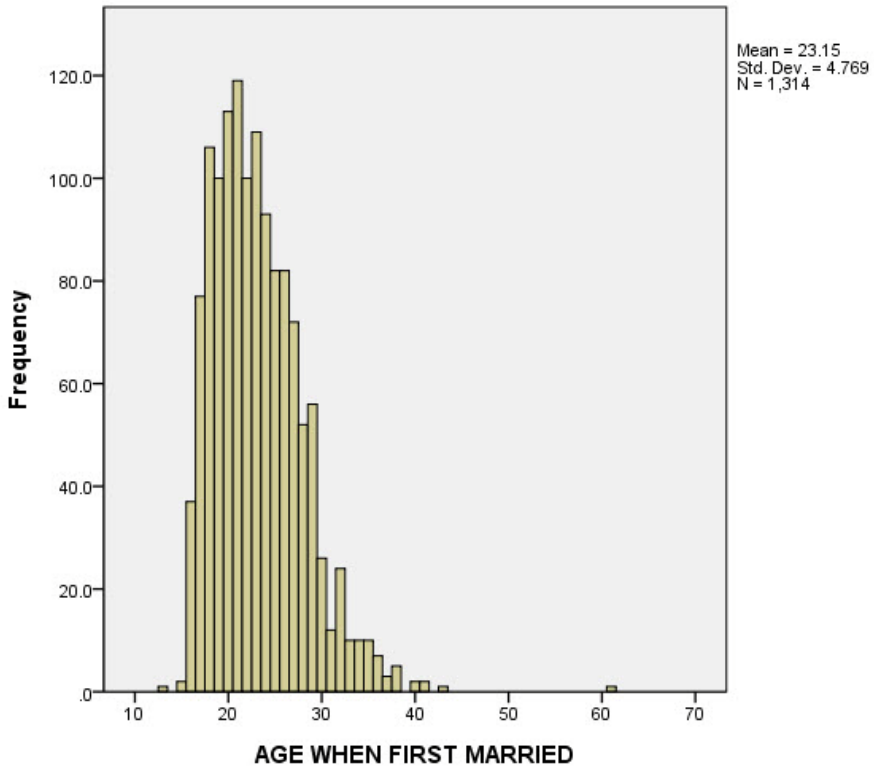
Figure 6.11 The Groups in the Statistics Subdialog



A histogram is used to graphically display the distribution of a scale variable. A histogram also has bars, like a bar chart, but they are plotted along an equal interval scale, and there is no gap between the bars. The height of each bar is the count of values of a quantitative (scale) variable falling within the interval. A histogram shows the shape, center, and spread of the distribution.

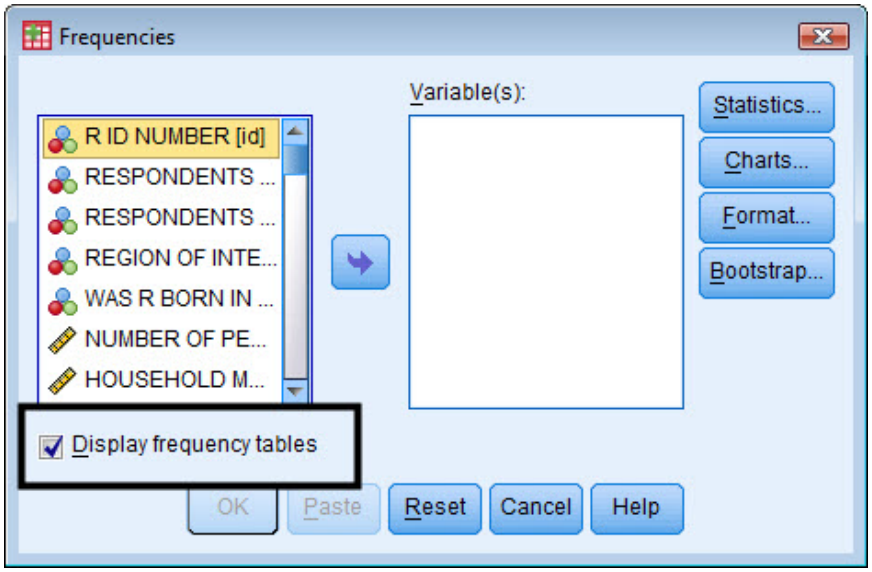
A normal curve can be superimposed on the histogram and helps you to judge whether the variable is normally distributed, which can be helpful for various statistical analyses.

Figure 6.12 Histogram of a Scale Variable



The Frequencies procedure is often used to summarize a scale variable, as a histogram can be requested in the Charts subdialog. If the Frequencies table itself is of no interest, uncheck the option *Display frequency tables* in the Frequencies dialog box.

Figure 6.13 Option to Display Frequency Table



6.11 *Demonstration: Summarizing a Scale Variable*

As an example of an analysis for a scale variable, we will run Frequencies for the variable *age*, requesting appropriate statistics and a histogram.

Detailed Steps for Summarizing a Scale Variable

1) Place the variable ***age*** in the Variable(s) box
2) Deselect **Display frequency tables** check box
3) Select the **Statistics** button
4) Select **Median**, **Mean**, **Std. deviation**, **Minimum**, and **Maximum** statistics
5) Select the **Charts** button
6) Select **Histogram** in the Chart Type area

Frequencies tables are usually not useful for scale variables since there may be almost as many distinct values as there are cases in the data file.

Results for Summarizing a Scale Variable

The Statistics table contains the requested statistics. The minimum value is 18 and the maximum value 97 for *age*. Note, that there is a 0.69 year difference between the mean (47.69) and the median (47) indicating that the variable is not terribly skewed in one direction or another, and is roughly normally distributed within the observed range.

Figure 6.14 Statistics for Age

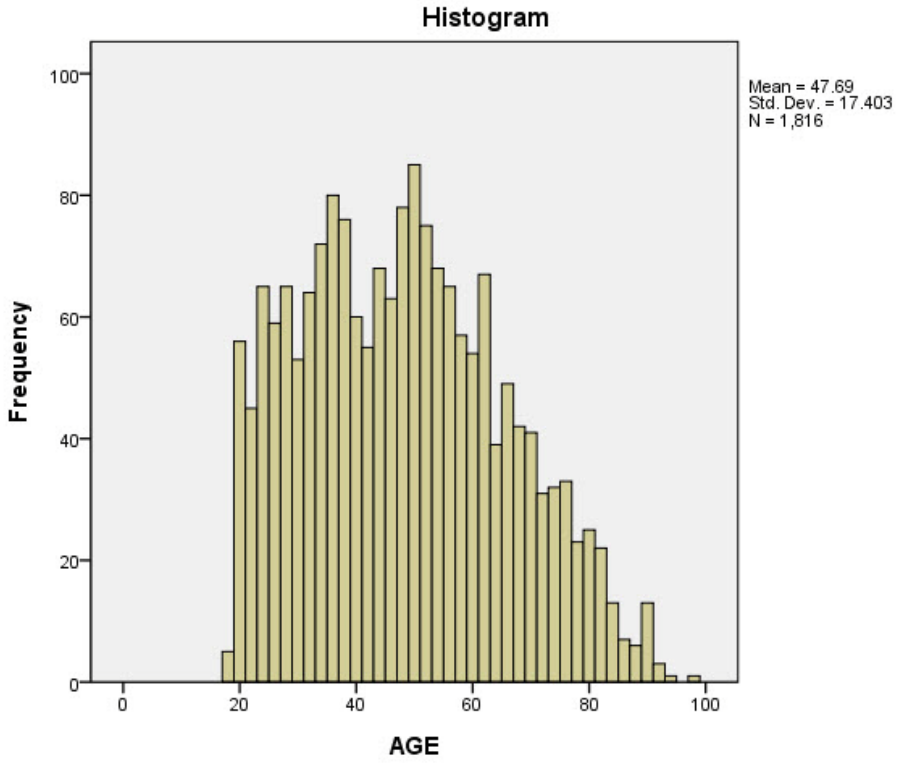
Statistics

AGE

N	Valid	1816
	Missing	207
Mean		47.69
Median		47.00
Std. Deviation		17.403
Minimum		18
Maximum		97

We can visually check the distribution with the histogram. We see that lower range of values is truncated at 18, and the number of people is highest in the middle age values (the "baby boomers") with the number of cases tapering off at the higher ages as we would expect. Thus, the age of respondent for this sample of adults is roughly normally distributed.

Figure 6.15 Histogram of Age



6.12 *Summarizing Scale Variables using Descriptives*

The **Descriptive**s procedure is a good alternative when the objective is to summarize scale variables. Descriptives is usually used to provide a table of statistical summaries (means, standard deviations, variance, minimum, maximum, etc.) for several scale variables. The Descriptives procedure also provides a succinct summary of the number of cases with valid values for each variable included in the table as well as the number of cases with valid values for all variables included in the table. These summaries are quite useful in evaluating the extent of missing values in your data and in identifying variables with missing values for a large proportion of the data. Descriptives does not provide a table listing each value of a variable.

The figure below shows the Descriptives table for a few variables.

Figure 6.16 Descriptive Statistics Output

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
RESPONDENTS SEX	2023	1	2	1.54	.498
AGE	1816	18	97	47.69	17.403
R'S AGE WHEN 1ST CHILD BORN	1489	13	61	24.28	5.033
Valid N (listwise)	1333				

The minimum and maximum values provide an efficient way to check for values outside the expected range. In general, this is a useful check for categorical variables as well. For example, although the

mean and standard deviation are not relevant for respondent's sex, the minimum and maximum for this variable show that there are no values outside the expected range.

The last row in the table labeled Valid N (listwise) gives the number of cases that have a valid value on all of variables appearing in the table. In this example, 1333 cases have valid values for all the three variables listed. Although this number is not particularly useful for this set of variables, it would be useful for a set of variables that you intended to use for a specific multivariate analysis. As you precede with your analysis plans, it is helpful to know how many cases have complete information and which variables are likely to be the main sources of potential problems.

6.13 *Requesting Descriptives for Scale Variables*

Requesting **Descriptives** is completed with these steps:

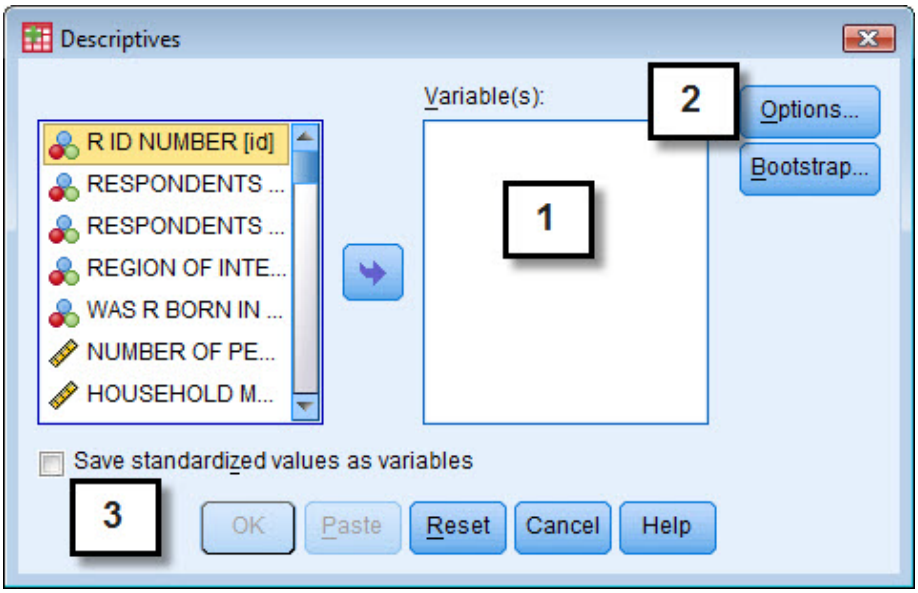
1) Select one or more variables to analyze
2) Request appropriate statistics

6.14 *Procedure: Descriptives for Scale Variables*

Descriptives is accessed from the *Analyze…Descriptive Statistics…Descriptives* menu choice. In the opening dialog box

1) Place one or more variables in the Variable(s) box
2) Select the *Statistics* button to specify statistics
3) Optionally, select *Save standardized values as variables* to create z-scores for the variables

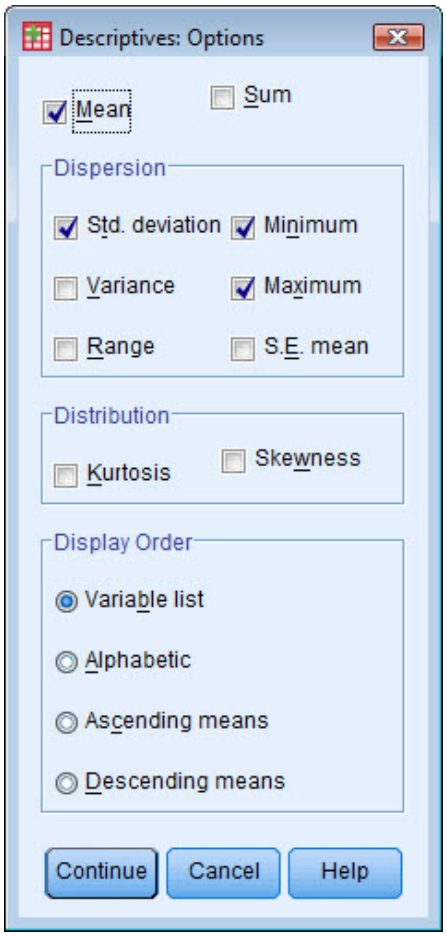
Figure 6.17 Descriptives Main Dialog



Only numeric variables appear in the Descriptives dialog box. The *Save standardized values as variables* feature creates new variables that are standardized forms of the original variables. These new variables, referred to as z-scores, have values standardized to a mean of 0 and standard deviation of 1.

The Options dialog allows you to select additional summary statistics to display. By default, minimum, maximum, mean, and standard deviation will be displayed. You can also select the display order of the variables in the table (for example, ascending or descending mean value).

Figure 6.18 Descriptives Option Subdialog



Tip

Display Order choices are useful if a number of similar variables are summarized. For example, suppose we have 10 satisfaction variables and want to report the means. The means can be sorted in descending order so the variable showing the highest satisfaction will be first, and the variable showing the lowest satisfaction will be last, in the table.

6.15 *Demonstration: Descriptives for Scale Variables*

In this demonstration we will request descriptive statistics on a few scale variables. For easier viewing, we will limit the size of the table and use only a subset of the scale variables in the data file. However, in practice you might run the procedure on all the numeric variables, as well as on subsets of variables that are of particular interest for planned analyses.

Detailed Steps for Descriptives for Scale Variables

1) Place the variables ***educ***, ***paeduc***, ***maeduc***, ***speduc***, ***sei***, ***pasei***, ***masei***, and ***spsei*** in the Variable(s) box
2) Select **Options** button
3) Select **Descending means**

Results for Descriptives for Scale Variables

The column labeled N shows the number of valid observations for each variable in the table. We see there is some variation in the number of valid observations: HIGHEST YEAR SCHOOL COMPLETED, which is educational attainment in years for the respondent, has more than 2000 valid observations, while R'S SPOUSE'S SOCIOECONOMIC INDEX has only 902 valid observations because many respondents are unmarried. Notice that only 400 cases have valid values for the entire set of questions.

Sorting the variables by their mean makes it easy to see that a respondent's education is higher than that of his or her spouse, or his or her parents. Interestingly, that relationship doesn't hold for socioeconomic index, where the highest mean is for the spouse.

Figure 6.19 Descriptive Statistics for Education and Sei Variables

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
R'S SPOUSE'S SOCIOECONOMIC INDEX	902	17	97	50.86	19.239
RESPONDENT SOCIOECONOMIC INDEX	1911	17	97	48.76	19.506
R'S FATHER'S SOCIOECONOMIC INDEX	1571	19	97	46.99	18.772
R'S MOTHER'S SOCIOECONOMIC INDEX	1283	17	97	43.12	19.175
HIGHEST YEAR SCHOOL COMPLETED, SPOUSE	963	0	20	13.57	3.019
HIGHEST YEAR OF SCHOOL COMPLETED	2018	0	20	13.43	3.079
HIGHEST YEAR SCHOOL COMPLETED, FATHER	1487	0	20	11.48	4.212
HIGHEST YEAR SCHOOL COMPLETED, MOTHER	1780	0	20	11.43	3.735
Valid N (listwise)	400				

Apply Your Knowledge

1. Suppose we have the following Descriptives table for the variable *age*. What is the reason for the low mean value?

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AGE	2015	-999	97	-55.68	312.781
Valid N (listwise)	2015				

2. Which measures of central tendency are appropriate for scale variables? Name all that apply.
   a. Mode
   b. Median
   c. Mean

6.16 *Lesson Summary*

In this lesson we defined measurement level and how to summarize individual variables with appropriate tables and charts based on the measurement level, using the **Frequencies** and **Descriptives** procedures.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Summarize individual variables using tables and graphs

And, they should also be able to:

- Define levels of measurement
- Use the Frequencies procedure to produce tables and charts appropriate for nominal variables
- Use the Frequencies procedure to produce tables and charts appropriate for ordinal variables
- Use the Frequencies and Descriptives procedures to produce tables and charts for scale variables

6.17 *Learning Activity*

In this exercise you will use the PASW Statistics data file *employee data.sav*. You will be asked to produce descriptive summaries on both the demographic and salary variables.

Supporting Materials

The data file *employee data.sav* contains information on employees of a major bank. Included is data on beginning and current salary position, time working, and demographic information.

1. Open the PASW Statistics data file *employee data.sav*.

2. Which variables in the file are nominal in level of measurement? Don't rely on the Measure variable attribute in the Data Editor to answer this question. Run Frequencies on these nominal variables (Hint: there should be 3). Request the mode plus a bar chart. Then rerun the procedure, requesting a pie chart.

3. Which job category had the most employees, and the fewest? Why isn't the mode reported for *gender*? Do you prefer a bar chart or pie chart to graphically summarize the distributions? Does you answer vary by variable? Why?

4. Run Frequencies for *salbegin* and *salary*, requesting summary statistics and a histogram with a superimposed normal curve. Suppress the display of the frequency table. How would you describe the distribution of each variable?

5. Produce a frequency table for *educ* showing the table in descending order of frequency. What are the most frequent, and least frequent, categories? Can you use the frequency table to determine the median value for *educ*?

6. Use the Descriptives procedure to produce summary statistics for *jobtime* and *prevexp*. Was anyone hired at the bank with no previous experience? How many cases have valid values for both variables?

Lesson 7: Modifying Data Values: Recode

7.1 *Objectives*

After completing this lesson students will be able to:

- Use various methods to group values of variables

To support the achievement of this primary objective, students will also be able to:

- Use the features of Variable Binning to group a scale variable
- Use the features of Recode into a Different Variable for categorical variables
- Use the features of Automatic Recode to create a numeric variable from a string variable

7.2 *Introduction*

The original values of a variable are often not the most suitable for your analysis or reporting needs. For reporting or analysis you may prefer to group categories together, or convert a continuous variable into one with a small number of categories. This can make tables and graphs easier to read for your intended audience. In this lesson, we discuss:

- **Creating a grouped version of a scale variable:** Instead of using income in dollars, you may want to use it in a discrete set of categories, defined by range. This can be accomplished with **Visual Binning**.
- **Combine several response categories into a single category:** You may want to create a "top-box" measure for a satisfaction question with the highly satisfied responses in one category, and all others in a second category. This can be accomplished with **Recode into Different Variables**.
- **Automatically create a numeric variable from a string variable:** For purposes of statistical analysis, you may need a numeric version of a string variable, such as region of the country coded originally as "West," "East," etc. This can be accomplished with **Automatic Recode**.

Business Context

The topics in this lesson can help in answering questions such as:

- How can I group together segments of customers for reporting and analysis?
- How can I create a grouped version of customer revenue with a small number of categories?

Supporting Materials

The file *census_small.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

7.3 *Visual Binning*

To bin data means taking two or more contiguous values and grouping them into the same category. This type of grouping is performed with the **Visual Binning** feature, where you can easily create a new categorical variable from an ordinal or scale variable by visually viewing the distribution of the

variable and choosing from several options for binning the values into discrete groups. You can also automatically create value labels for the new variable using this feature.

7.4 *Requesting Visual Binning*

Visual binning can only be done with variables that are ordinal or scale in measurement. In fact, only variables that are ordinal or scale will appear in the Visual Binning variable list. If necessary, you can change the level of measurement beforehand in the Data Editor Variable View tab so that a variable does appear.

Requesting **Visual Binning** is accomplished with these steps:

1) Choose at least one ordinal or scale variable to bin
2) Choose a method to make cutpoints, including equal width intervals, percentiles, or ones based on the mean and standard deviation
3) Decide on how to treat endpoints of bins
4) Name the new (binned) variable
5) Optionally, create labels for the new bins

7.5 *Visual Binning Output*

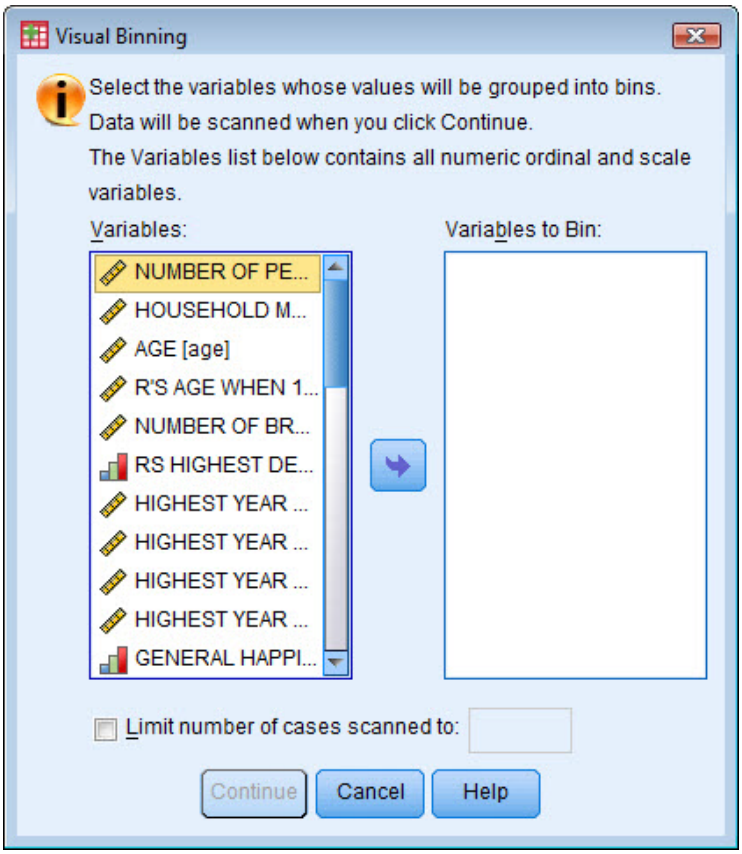
Visual Binning is a transformation, not a procedure. This means that it doesn't create any output as such, just new binned variables. To see the result, we need to use a table or graph to display the distribution of the new binned variable.

7.6 *Procedure: Visual Binning*

The **Visual Binning** transformation is accessed from the *Transform…Visual Binning* menu choice. In the opening dialog box

1) Place a variable to be binned in the Variables to Bin: box

Figure 7.1 Visual Binning Variable Selection Dialog



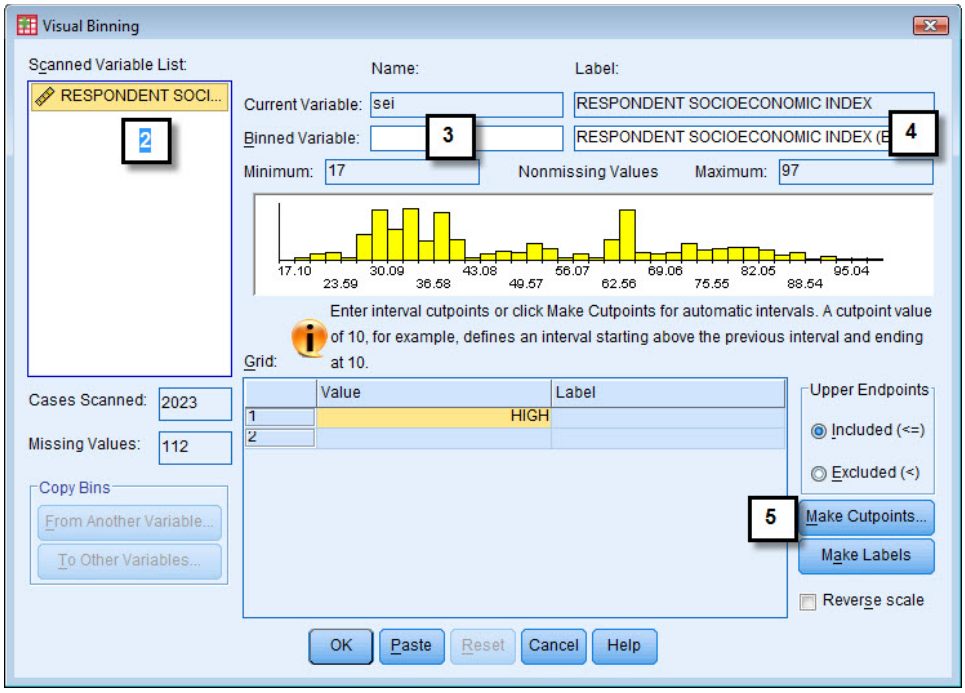
When you select Continue, the variable's values are scanned so it can be displayed in the subsequent dialog box.

2) Click a variable to bin in the Scanned Variable List

When a variable is selected in the Scanned Variable List, its distribution is displayed with a histogram. The Grid below the histogram is where the bins will be listed, along with their labels. Only cutpoints between bins are listed in the Value cells.

3) Add a name for the Binned Variable
4) A default variable label is supplied, but you can modify it in that text box

Figure 7.2 Visual Binning Main Dialog



5) Select the *Make Cutpoints* button to choose a method of binning

Options available for binning include:

- **Equal width intervals**. Generates binned categories of equal width (for example, 1–10, 11–20, and 21–30) based on any two of the following three criteria:
  - First Cutpoint Location. The value that defines the upper end of the lowest binned category (for example, a value of 10 indicates a range that includes all values up to 10).
  - Number of Cutpoints. The number of binned categories is the number of cutpoints plus one. For example, 9 cutpoints generate 10 binned categories.
  - Width. The width of each interval. For example, a value of 10 for the variable age would bin it into 10-year intervals.
- **Equal Percentiles Based on Scanned Cases**. Generates binned categories with an equal number of cases in each band, based on criteria of either the Number of Cutpoints or the Width or Percentage of cases in each category. For example, specifying three cutpoints will generate four categories, each containing approximately 25% of the cases. Specifying a width (%) of 25 would generate the same four categories
- **Cutpoints at Mean and Selected Standard Deviations Based on Scanned Cases.** Generates categories based on the values of the mean and standard deviation of the distribution of the variable. If you do not select one of the standard deviation intervals, two categories will be created, with the mean as the cutpoint dividing the categories. You can select any combination of standard deviation intervals based on one, two, and/or three standard deviations. For example, selecting all three would result in eight categories--six categories in one standard deviation intervals and two categories for cases more than three standard deviations above and below the mean.

Note

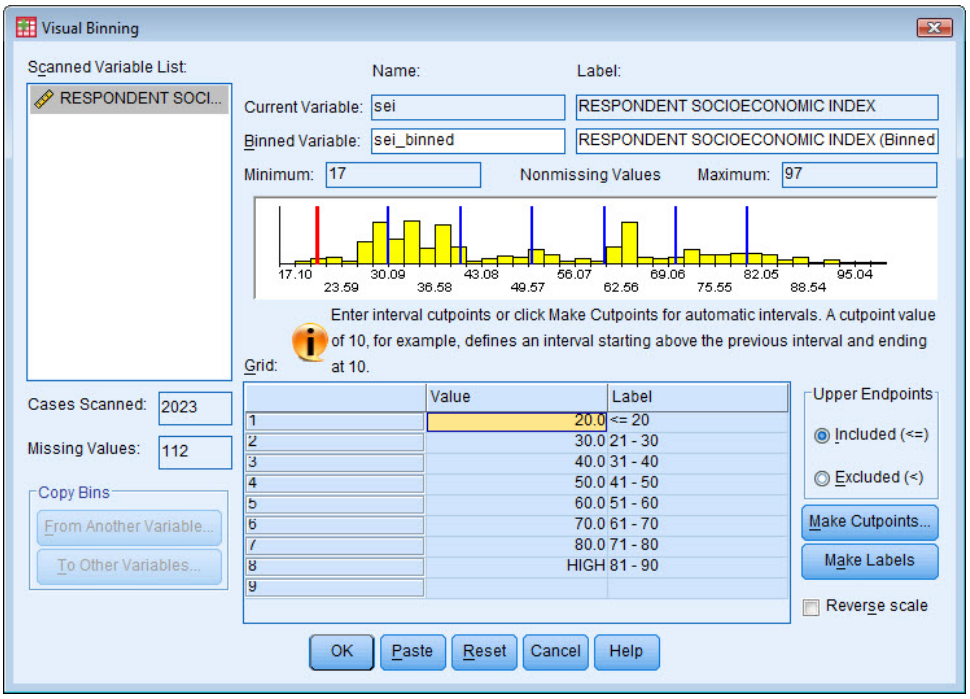
Creating categories based on standard deviations may result in some defined bands outside the actual data range and even outside the range of possible data values (for example, a negative salary range).

When you have selected a method of making cutpoints, you are returned to the main dialog box, where the bins are now represented with vertical bars in the histogram. By default, the upper endpoint is included in the bin, i.e., 0-20 includes 20 in the bin.

The *Make Labels* button allows you to create automatic labels that list the lower and upper endpoints of each bin, or the range for the lower and upper bins. Alternatively, you can add your own labels directly.

Note that missing values are not included in the binning.

Figure 7.3 Variable with Bins Added



7.7 *Demonstration: Visual Binning*

We will work with the PASW Statistics data file *census_small.sav* in this lesson. As an example, we will create a new variable, *age_grouped*, grouping *age* into 3 categories: up to age 35 (including 35), 35-64 (excluding 35, including 64) and 65 and over.

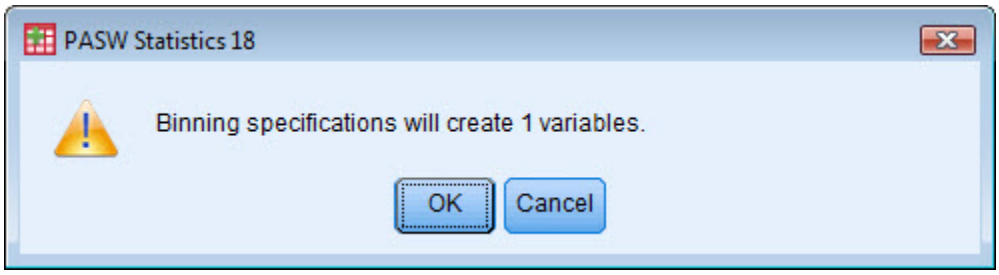
Detailed Steps for Visual Binning

1) Place the variable ***age*** in the Variables to Bin box
2) Select ***age*** in the Visual Binning dialog
3) Name the new binned variable ***age_grouped***
4) Label the new binned variable ***AGE - Grouped***

5) In the Make Cutpoints dialog, input **35** in the First Cutpoint Location box
6) Input **2** in the Number of Cutpoints box
7) Input **29** in the Width box
8) Select **Make Labels** to create automatic labels

PASW Statistics will display a message box that 1 new variable will be created.

Figure 7.4 Message about New Variable to be Created



Results from Visual Binning

The new variable *age_grouped* is added to the end of the data file and is displayed in the far right column in Data Editor, Data View tab. To see its distribution, we can use the **Frequencies** procedure, available from the Analyze…Descriptive Statistics…Frequencies menu.

1) Place the variable ***age_grouped*** in the Variable(s) box

Figure 7.5 Frequencies for Binned Age

AGE - Grouped

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<= 35	521	25.8	28.7	28.7
	36 - 64	953	47.1	52.5	81.2
	65+	342	16.9	18.8	100.0
	Total	1816	89.8	100.0	
Missing	REFUSED TO ANSWER	207	10.2		
Total		2023	100.0		

Apply Your Knowledge

1. Which binning method creates bins with equal numbers of cases?
   a. Equal widths
   b. Percentiles
   c. Mean and standard deviation

2. Suppose that you want to bin a variable *income* into a new variable, but it doesn't appear in the Visual Binning dialog box. Why?
   a. *Income* is defined as nominal in level of measurement
   b. *Income* is defined as ordinal in level of measurement
   c. *Income* is defined as scale in level of measurement

7.8 *Recode into Different Variables*

For some analyses you may need to collapse or group category values of a nominal or ordinal variable into a smaller number of categories. Examples include creating broad job categories from specific job position codes, grouping postal codes into larger regional categories, or collapsing five marital status categories (married, widowed, divorced, separated, never married) into two categories (married, not married).

The Recode features on the Transform menu allow you to easily create new categories for a nominal or ordinal variable. You can choose to **Recode into Same Variables** which changes the values of the existing (source) variable or **Recode into Different Variables** which creates a new variable with the new categories and leaves the existing (source) variable unchanged. Because we most often want to retain both versions of the variable, the Recode into Same Variables option is not often used. It can be useful for changing values when cleaning your data.

Specifications for Recode into Same Variables are almost identical to those for Recode into a Different Variables. Given the situation that **Recode into Same Variables** is less often used, we will discuss **Recode into Different Variables** in this lesson.

7.9 *Requesting Recode into a Different Variable*

Requesting a **Recode into Different Variables** is completed with these steps:

1) Select the variable to recode
2) Name the new Output Variable and add an optional label
3) Specify how old values should be grouped into new categories

7.10 *Recode Output*

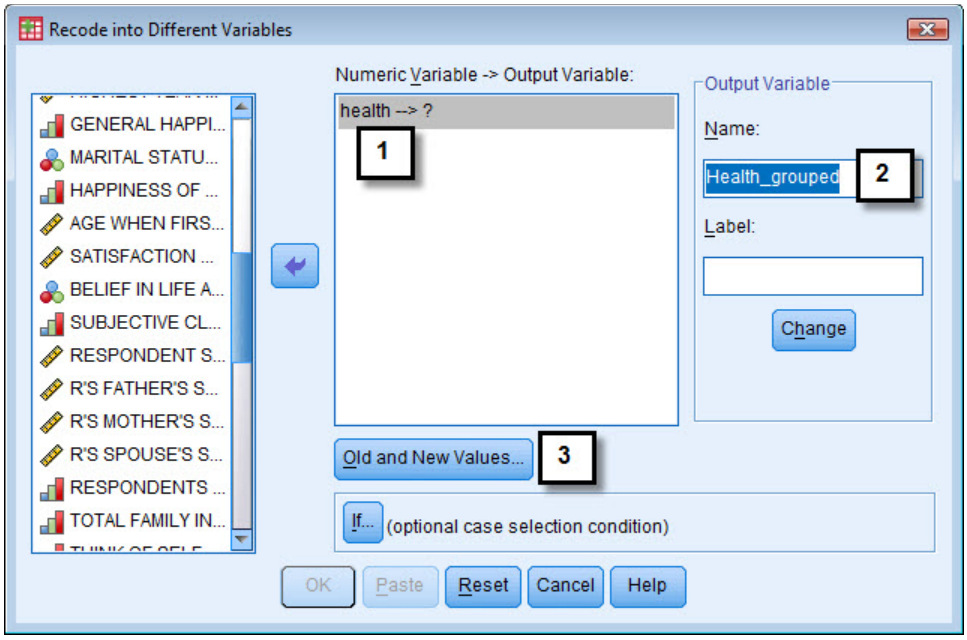
Recode into Different Variables is a transformation, so it doesn't create any output as such, just a new variable. To see the result, we need to use a table or graph to display the distribution of the new binned variable, as we do below.

7.11 *Procedure: Recode into Different Variables*

The **Recode into Different Variables** transformation is accessed from the *Transform…Recode into Different Variables* menu choice. In the opening dialog box

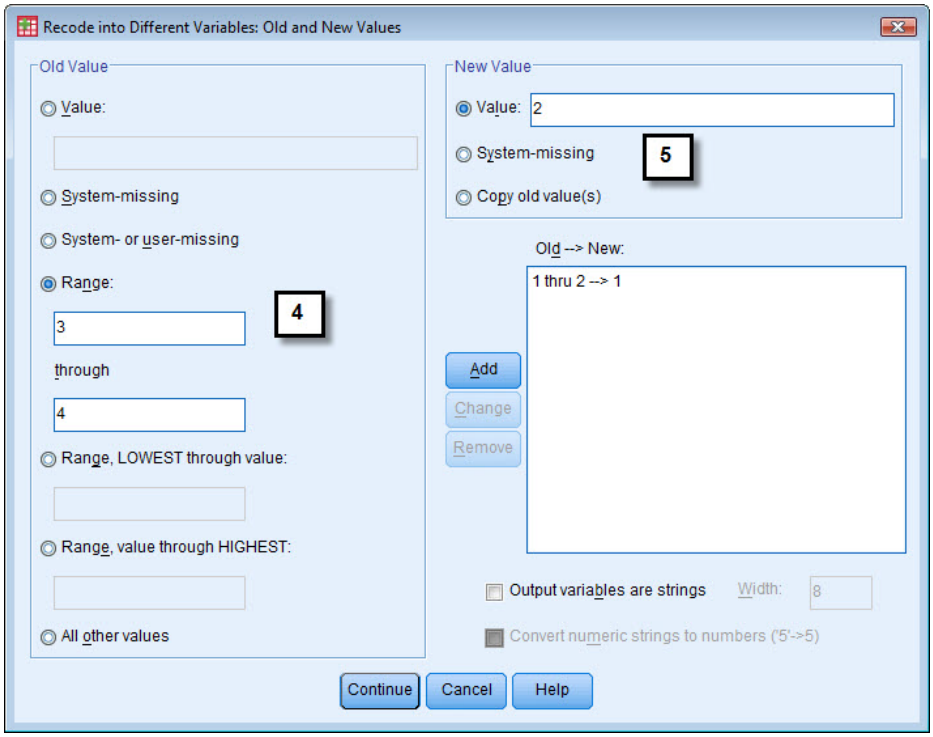
1) Place a variable to be recoded in the Numeric Variable box
2) Name the new variable and select the *Change* button to add to the Output Variable list
3) Select *Old and New Values* button to input recode specifications

Figure 7.6 Recode Into Different Variables Dialog



4) In the Old and New Values dialog, one or more old (existing) values are specified on the left in the Old Value area
5) A single new value is specified on the right in the New Value area.
6) Select the *Add* button to create the new category

Figure 7.7 Old and New Values Dialog



In specifying the old values, there are several options:

- Specify a single value
- Specify a range of values, in three ways:
    - With lower and upper endpoint user-specified
    - Starting with the lowest value encountered in the data through a user-specified upper endpoint
    - Starting with a user specified lower endpoint through highest value encountered in the data
- Specify missing values, in two ways:
    - System-missing
    - System- or user-missing
- All other values: all values not specified

Note

When using the Range options, user-missing values are included if they fall within a range, unless they are specifically recoded with the missing value option, or recoded explicitly otherwise.

In specifying the new values, there are three options:

- Value: specify a single value
- System-missing: recode to system-missing
- Copy old value

None of the variable descriptive information from the input variable will be copied to the output, recoded variable. No value labels are defined, no user-missing values will be declared, and the measurement level is seta to scale. Before using this variable, you will need to add the appropriate variable definitions.

Important

The order of defining the categories matters. For instance, first recoding 3 into 1, and then recoding all other values to 0, gives a different result from that of first recoding all other values to 0, and then recoding 3 into 1. In the latter case, actually “all other values” includes all values of the variable at the moment that no value is specified, so actually includes all values. So, in general, the first hit counts.

Demonstration: Recode into Different Variables

To demonstrate recoding, we will create a new variable from the variable *MARITAL STATUS [marital*]. The variable currently has the following coding:

Value	Value label
1	Married
2	Widowed
3	Divorced
4	Separated
5	Never married
9	Refused to answer

Our task is to create a new variable, called *marital_3* with these three categories:

1 Married
2 Never Married
3 Other

Detailed Steps for Recode into Different Variables

1) Place the variables ***marital*** in the Input Variable box
2) Name the new variable ***marital_3***
3) Select the **Change** button
4) Select the **Old and New Values** button
5) In the Old Value area, select **Value** and specify **1**
6) In the New Value area, select **Value** and specify **1**
7) Select **Add**
8) In the Old Value area, select **Value** and specify **5**
9) In the New Value area, select **Value** and specify **2**
10) Select **Add**
11) In the Old Value area, select **Range** and specify **2** through **4**
12) In the New Value area, select **Value** and specify **3**
13) Select **Add**

The initial value for all new numeric variables in PASW Statistics is the system-missing value. Cases containing values that have not been specified will have the system-missing value on the new variable. At this point, we have specified recodes for the five valid values, but since we have not mentioned the value 9, cases containing that user-missing value will be system-missing on *marital_3*. This could be acceptable, but if we want to label this category as “Refused to Answer,” we need to retain the value of 9.

14) In the Old Value area, select **All other values**
15) In the New Value area, select **Copy old values**
16) Select **Add**

Results from Recode into Different Variables

The new variable *marital_3* is added to the end of the data file and is displayed in the far right column in Data Editor, Data View tab. To see its distribution, we can use the **Frequencies** procedure. We will request frequencies for both *marital* and *marital_3*.

It is very easy to make a logical error, or typing error, when recoding variables. It is thus very important that you check your work.

Important

1) In the Frequencies dialog, place ***marital*** and ***marital_3*** in the Variable(s) box

When we view the frequency tables, we want to match frequency counts to insure that the recode was successful. For example, there are 164 respondents who are widowed, 281 who are divorced, and 70 who are separated. That is a total of 515 respondents who should have a value of 3 for *marital_3*.

Note that we should assign value labels to the new variable, declare 9 as user-missing, set the number of decimals to 0, and set the measurement level to nominal. It is common to have to do this additional work after recoding.

Figure 7.8 Frequencies for Marital Status and Recoded Marital Status

MARITAL STATUS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	MARRIED	972	48.0	48.2	48.2
	WIDOWED	164	8.1	8.1	56.3
	DIVORCED	281	13.9	13.9	70.2
	SEPARATED	70	3.5	3.5	73.7
	NEVER MARRIED	531	26.2	26.3	100.0
	Total	2018	99.8	100.0	
Missing	REFUSED TO ANSWER	5	.2		
Total		2023	100.0		

marital_3

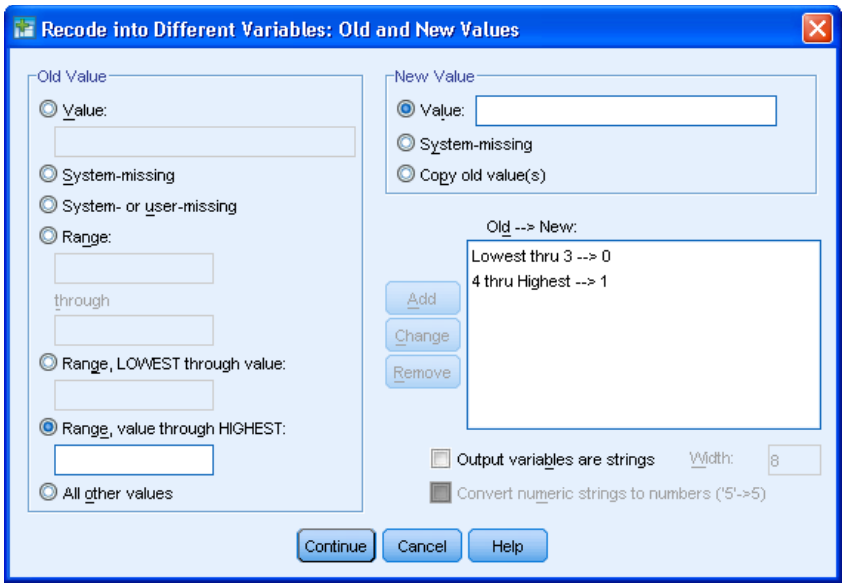
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	972	48.0	48.0	48.0
	2.00	531	26.2	26.2	74.3
	3.00	515	25.5	25.5	99.8
	9.00	5	.2	.2	100.0
	Total	2023	100.0	100.0	

Apply Your Knowledge

1. Look at the table and figure below, which contain a Frequencies table for the number of persons in a household aged 18 years and above, and a recoding scheme for this variable in the Recode dialog box. What will the user-missing value of 99 (REFUSED TO ANSWER) be after the recode?
    a. System-missing
    b. 0
    c. 1
    d. 99

HOUSEHOLD MEMBERS 18 YRS AND OLDER

		Frequency	Valid Percent
Valid	1	668	33.1
	2	1049	52.0
	3	227	11.2
	4	52	2.6
	5	18	.9
	6	4	.2
	Total	2018	100.0
Missing	99 REFUSED TO ANSWER	5	
Total		2023	



2. True or false? An existing value in a variable can be recoded to more than one value in a new variable?

7.12 *Automatic Recode of Variables*

One common task in data transformation is to convert string variables into numeric variables. This may be necessary because some statistical procedures require numeric variables, or because you want to combine categories of string variables (e.g., product type) and it is easier done with numeric codes. To make this conversion easy, PASW Statistics provides the **Automatic Recode** feature, which creates a new numeric variable, assigning a unique numeric value for each different string value in the original variable. The value labels—or if there are no value labels, the original data values—of the string variable become the value labels of the newly created numeric variable.

Automatic Recode can also be used with numeric values.

When recoding, these options are available:

- Sort the values starting from lowest value (ascending order) or from highest value (descending order). Default is starting from lowest value, in which case the alphabetically first string value will be assigned a value of 1.
- Treat blank string values as user-missing, so blank strings are automatically recoded into a user-missing value (and its value will be one higher than the highest non-missing value).
- Save the autorecoding scheme in a template file and then apply it to other variables and other data files. For example, you may have a large number of alphanumeric product codes that you automatically recode into integers every month, but some months new product codes are added that change the original autorecoding scheme. If you save the original scheme in a template and then apply it to the new data that contain the new set of codes, any new codes encountered in the data are automatically recoded into values higher than the last value in the template, preserving the original autorecoding scheme of the original product codes.

7.13 *Requesting Automatic Recode*

Requesting an **Automatic Recode** is accomplished with these steps:

1) Select a variable to recode

2) Provide a new name for the output variable
3) Optionally, select how to sort string values
4) Optionally, select how to handle blanks
5) Optionally, apply an existing recoding template or save the current recoding scheme as a template

7.14 *Automatic Recode Output*

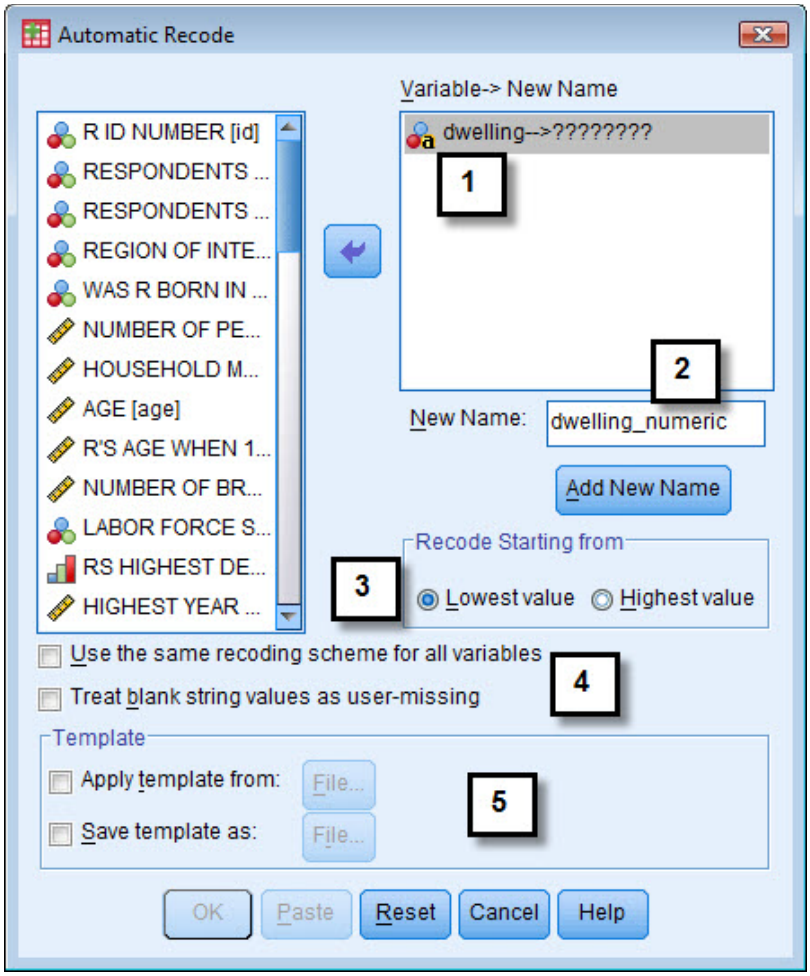
Automatic Recode creates a table of the old and new values in the Viewer window, and you should check this. In addition, it is a good idea to request Frequencies for the new recoded variable.

7.15 *Procedure: Automatic Recode*

The **Automatic Recode** transformation is accessed from the *Transform…Automatic Recode* menu choice. With the dialog box open:

1) Add one or more variables to recode to the Variable->New Name list
2) Enter a name for the recoded variable in the New Name box, and add it
3) You can change the sorting order of string values in the Recode Starting from area; the default of *Lowest value* begins in alphabetical order, with upper case letters preceding lower case
4) If there is more than one variable being recoded, you can select *Use the same coding scheme for all variables*; you can also treat blank values as user-missing
5) The Template area check boxes allow you to apply an existing template for recoding or create a new one

Figure 7.9 Automatic Recode Dialog



7.16 *Demonstration: Automatic Recode*

In this example, we will recode the variable *TYPE OF STRUCTURE [dwelling]*. Suppose that we want to recode this variable into a smaller number of categories. The frequency table below shows the current values and distribution of cases.

Notice that this variable is defined as alphanumeric, and attempting to recode it by typing in each of the values (which are case sensitive) is quite demanding. What we would like to have is a variable, say *dwelling_numeric*, with numeric values and the string values of *dwelling* as value labels. We then can perform the Recode operation on this numeric variable.

Note that 31 respondents have left this variable blank. We will treat this blank value as a missing value.

Figure 7.10 Frequencies for Type of Structure

TYPE OF STRUCTURE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		31	1.5	1.5	1.5
	2 UNITS SIDE BY SIDE	62	3.1	3.1	4.6
	2 UNITS-ONE ABOVE	25	1.2	1.2	5.8
	3-4 FAM HOUSE	34	1.7	1.7	7.5
	APARTMENT HOUSE	224	11.1	11.1	18.6
	APARTMENT-4 STORIES	73	3.6	3.6	22.2
	APARTMENT-COMMERCIAL	8	.4	.4	22.6
	DETACHED 1-FAM HOUSE	1336	66.0	66.0	88.6
	OTHER	23	1.1	1.1	89.8
	ROW HOUSE	89	4.4	4.4	94.2
	TRAILER	118	5.8	5.8	100.0
	Total	2023	100.0	100.0	

Detailed Steps for Automatic Recode

With the **Automatic Recode** dialog box open:

1) Place the variable ***dwelling*** in the Variable->New Name list
2) Name the new variable ***dwelling_numeric***
3) Select the option to **Treat blank string values as user-missing**

Results from Automatic Recode

The output generated from **Automatic Recode** is a table of the old and new values for the variable. Notice that each value of *dwelling* is assigned a unique numeric value, in alphabetical order (numbers come before letters), on *dwelling_numeric* and the data values of *dwelling* become the value labels for *dwelling_numeric*. The blanks have been autorecoded to the value 11 and this value is declared as a user-missing value.

Figure 7.11 Recoding Scheme for Type of Structure

dwelling into dwelling_numeric (TYPE OF STRUCTURE)

Old Value	New Value	Value Label
2 UNITS SIDE BY SIDE	1	2 UNITS SIDE BY SIDE
2 UNITS-ONE ABOVE	2	2 UNITS-ONE ABOVE
3-4 FAM HOUSE	3	3-4 FAM HOUSE
APARTMENT HOUSE	4	APARTMENT HOUSE
APARTMENT-4 STORIES	5	APARTMENT-4 STORIES
APARTMENT-COMMERCIAL	6	APARTMENT-COMMERCIAL
DETACHED 1-FAM HOUSE	7	DETACHED 1-FAM HOUSE
OTHER	8	OTHER
ROW HOUSE	9	ROW HOUSE
TRAILER	10	TRAILER
M	11M	

Apply Your Knowledge

1. True or False? Automatic Recode can only be used for string variables?

7.17 *Lesson Summary*

Lesson Objectives Review

Students who have completed this lesson should be able to:

- Use various methods to group values of variables

And, they should also be able to:

- Use the features of Variable Binning to group a scale variable
- Use the features of Recode into a Different Variable for categorical variables
- Use the features of Automatic Recode to create a numeric variable from a string variable

7.18 *Learning Activity*

In this exercise you will use the PASW Statistics data file *employee data.sav*.

Supporting Materials

The data file *employee data.sav* contains information on employees of a major bank. Included is data on beginning and current salary position, time working, and demographic information.

1. Use **Visual Binning** on the variable *salary* to create a new variable that groups it into 5 categories (call it *salgrp*). To decide which range of values to group together, you should look at the histogram. You can request equal Cut Points for 5 groups by clicking the Make Cutpoints button and requesting 5 cutpoints with equal percentiles. Or, you can choose another method of binning. Have **Visual Binning** make labels. After the binning is complete, check your work by running frequencies of *salgrp*.

2. Use **Recode into Different Variables** to group the variable *jobcat* into two categories, 1= Manager and 2= All Others. Name the new variable *manager*. Add value labels to *manager*. Use Frequencies to check your work.

3. The variable *gender* is a string variable (with values f and m). Use **Automatic Recode** to create a numeric variable *gender_numeric* for *gender*. Make the value of m equal to 1. Use Frequencies to check your work.

4. *For those with more time:* Use **Visual Binning** on the variable *prevexp* (Months Previous Experience). Given its distribution, what type of binning would you recommend? Try at least two different types of binning schemes and compare them. Which one would you prefer and why?

Lesson 8: Modifying Data Values: Compute

8.1 *Objectives*

After completing this lesson students will be able to:

- Use the **Compute Variable dialog** to create new variables

To support the achievement of this primary objective, students will also be able to:

- Describe the features of **Compute Variable**
- Create new variables with numeric expressions
- Create new variables with conditional numeric expressions

8.2 *Introduction*

A common step in data preparation is to create a new variable with a mathematical expression from one or more existing variables. As one example, we might want to add two sources of customer revenue together to obtain total revenue. Or, we might need to use a special date function to extract portions of a date for a report, e.g., month.

These types of transformations can be completed with the **Compute Variable** feature, which allows you to construct a simple, or complex, expression based on one or more existing variables or numeric values. The expression can be applied to every case in the file, or to only a subset of cases (e.g., females).

In this lesson we will provide several examples of using **Compute Variable**.

Business Context

The topics in this lesson can help in answering questions such as:

- How can I create a variable with the mean of three variables?
- How can I create a new variable for only a subset of cases?

Supporting Materials

The file *census_small.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

8.3 *Computing a Variable*

Compute Variable allows the user to create variables using a numeric expression. For instance, if we have the variable *income*, measuring income of the respondent, and *fincome*, measuring family income, we can derive the percentage of family income that comes from the respondent's income with the following simple equation:

```
pct_income = (income/fincome) * 100.
```

The variable *pct_income* is the target of the expression. In the numeric expression, variables and arithmetic operators can be used, in addition to functions. The usual order of operations applies, so expressions in parentheses are evaluated first.

Compute Variable can also be used to create, or modify, string variables, typically with the use of string functions.

Functions
The expression can include a function, and arithmetic, statistical, logical and string functions are available. Functions require one or more *arguments*, specified in parentheses. For instance, if a variable *X* has negative and positive values and all scores need to be transformed into positive values, the function ABS (absolute value) is the right choice, in this expression:

```
X_positive  = ABS(X).
```

Expressions can include more than one function, and, of course, other operators.

Conditional Computing
Conditional computing makes processing the expression dependent on a logical true/false condition. For instance, suppose we want to compute a variable *tax*, which is calculated as 10% of income for those having an annual income up to 50,000 and as 5,000 + 15% of income above 50,000 for those having an annual income greater than 50,000. So, the computations are:

```
if  income <= 50000  Tax =  income*0.10
if  income > 50000    Tax =  5000 + (income – 50000)*0.15
```

The computation is not the same for all cases, but depends on whether a certain condition is satisfied. In this example the condition involves the same variable as used in the numeric expression (*income*), but this need not to be the case, and the condition can include any variable in the data (or a combination of variables, functions, and so on).

Note

Existing variables can be used as a target variable in **Compute Variable**.

8.4 *Requesting Compute Variable*

The steps for **Compute Variable** are:

1) Specify the name of the target variable to create or modify
2) Set the variable type and label, if necessary or desired
3) Build the expression
4) Optionally, specify a logical condition to perform the transformation, in the *If Cases* subdialog

8.5 *Compute Variable Output*

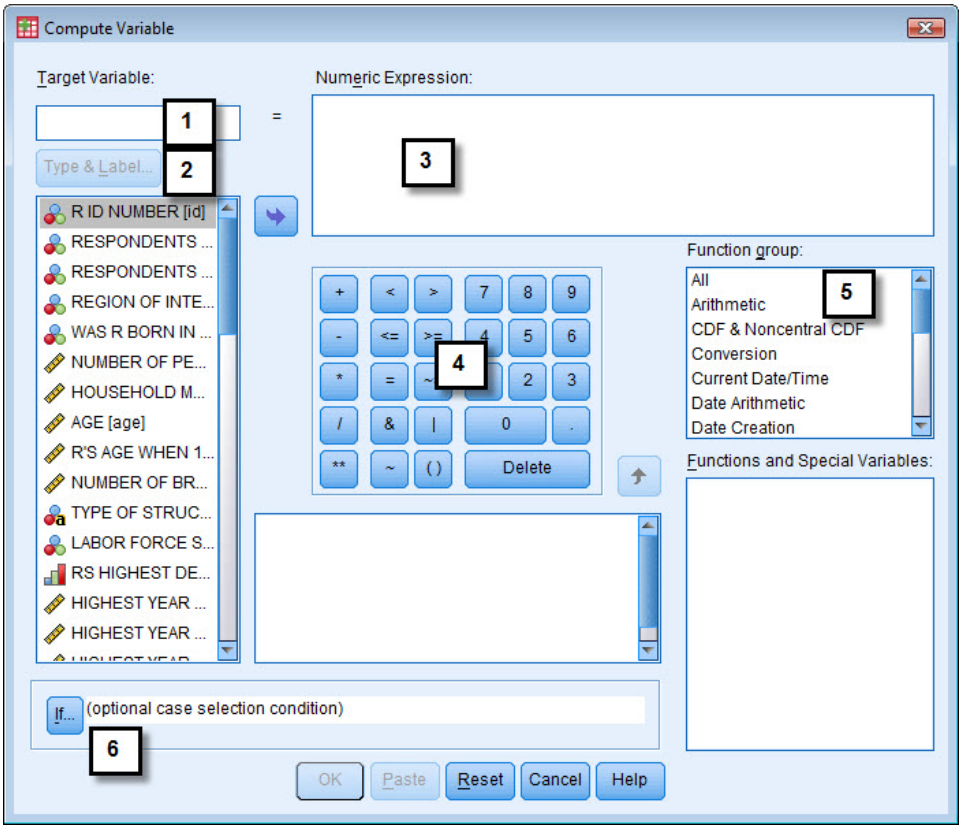
Compute Variable is a transformation, not a procedure. This means that it doesn't create any output as such, just the new variable. To see the result, we need to switch to the Data Editor, or use a table or graph to display the distribution of the new variable.

8.6 *Procedure: Compute Variable*

The **Compute Variable** transformation is accessed from the *Transform…Compute Variable* menu choice. In the main dialog box

1) Name the new variable in the Target Variable box
2) Set the type and add a label for the new variable in the *Type & Label* subdialog
3) Create the expression in the Numeric Expression box
4) Use the calculator pad for operators
5) The functions are available in 18 Function groups
6) The *If* button opens the If Cases subdialog to create a logical condition for processing

Figure 8.1 Compute Variable Dialog

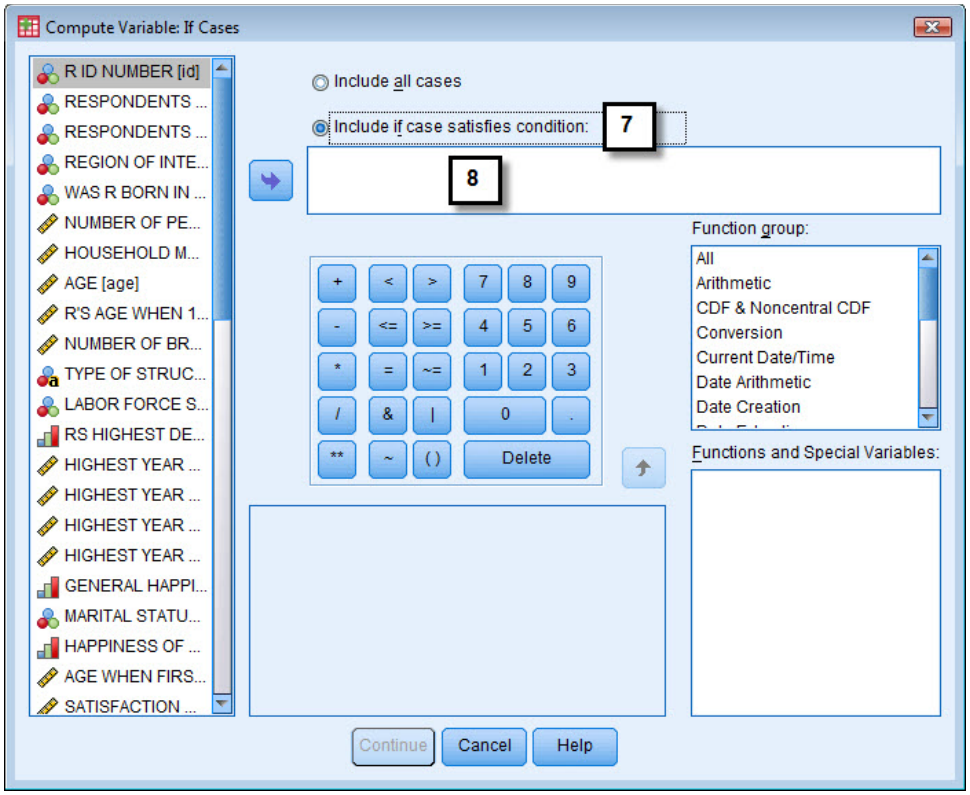


In the If Cases subdialog:

7) Select the *Include if case satisfies condition* button to activate the expression box
8) Create an expression to be evaluated for each case; if true, the compute is done; if false, the compute is not done

All the variables, operators, and functions are available to create the logical expression.

Figure 8.2 Compute Variable If Cases Subdialog



8.7 *Demonstration: Compute Variable*

We will use the data file *census_small.sav* in these examples. The data contain a variable for the respondent's current age (*age*) and a variable for the age of the respondent when their first child was born (*R'S AGE WHEN 1ST CHILD BORN [agekdbrn]*). For respondents with no children, *agekdbrn* is system-missing. The data file does not contain a variable for age of the oldest child, which is what we would like to know.

We can create a new variable, *age_1stBornChild*, that is the difference between a respondent's current age and the age when their first child was born, which should be the approximate age of their oldest child.

Detailed Steps to Compute Age of Oldest Child

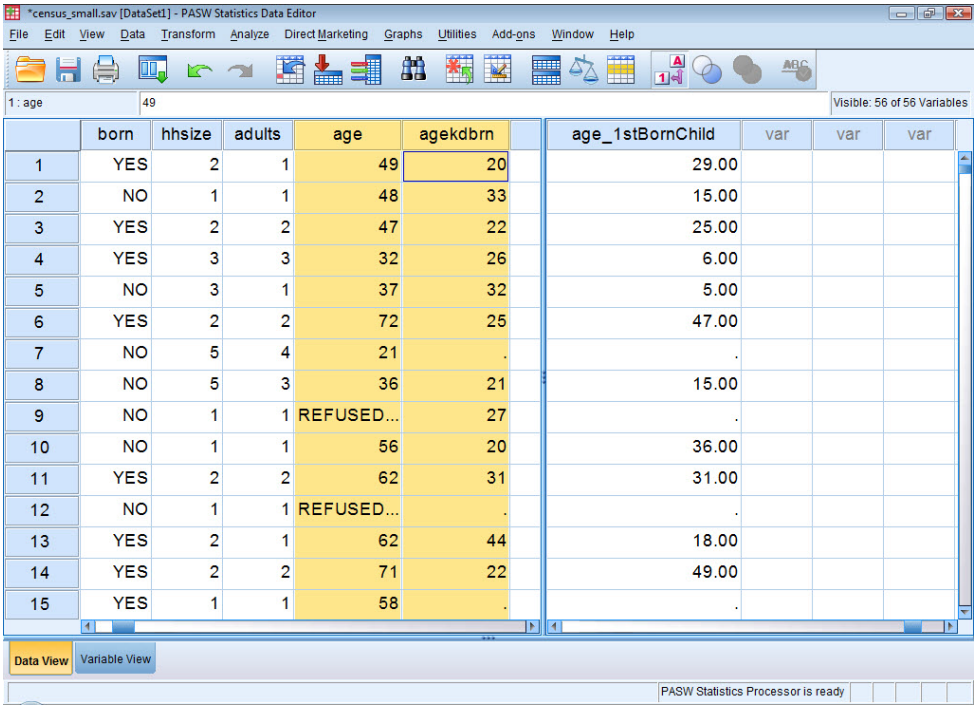
1) In the Variable Name box enter the name ***age_1stBornChild***
2) In the Type & Label dialog, enter the label **Age of first born child**
3) In the Numeric Expression area, construct the expression **age - agekdbrn**

Results to Compute Age of Oldest Child

When **Compute Variable** is executed, the new variable is created and placed at the end of the Data Editor. We see that some cases have a system-missing value for *age_1stBornChild*. This is because one of the age variables, or both, has a missing value. The result of a numeric expression is missing when one or more of the variables in the expression are missing. When computing variables with missing values, it's imperative that you check the results to ensure that the missing data are handled in a way that is acceptable.

Otherwise, the new variable has been created with the approximate age (probably accurate within one year) of the age of a respondent's eldest child.

Figure 8.3 Variable *age_1stBornChild* Added to Data Editor



	born	hhsize	adults	age	agekdbrn	age_1stBornChild	var	var	var
1	YES	2	1	49	20	29.00			
2	NO	1	1	48	33	15.00			
3	YES	2	2	47	22	25.00			
4	YES	3	3	32	26	6.00			
5	NO	3	1	37	32	5.00			
6	YES	2	2	72	25	47.00			
7	NO	5	4	21	.	.			
8	NO	5	3	36	21	15.00			
9	NO	1	1	REFUSED...	27	.			
10	NO	1	1	56	20	36.00			
11	YES	2	2	62	31	31.00			
12	NO	1	1	REFUSED...	.	.			
13	YES	2	1	62	44	18.00			
14	YES	2	2	71	22	49.00			
15	YES	1	1	58	.	.			

Important

It is easy to make a logical error, or typing error, when computing new variables. It is thus very important that you check your work.

Apply Your Knowledge

1. True or false? If a case has a user-missing value 99 for variable X, and a new variable Y= X * 2 is computed, Y has the value 198 for this case?

2. True or False? Can **Compute Variable** be used to create a string variable?

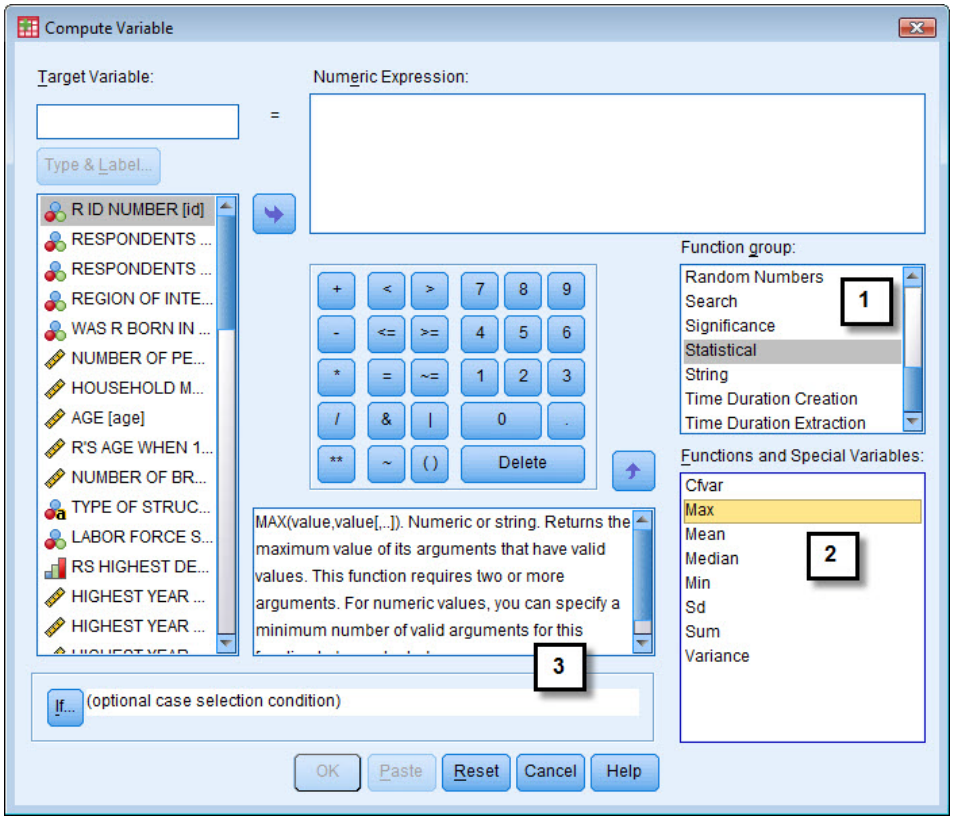
8.8 *Using Functions in Compute Variable*

PASW Statistics provides many functions for use in transformations.

1) The *Function group* area in the **Compute Variable** dialog groups functions into areas of similar use or action. For example, the Statistical group contains functions that compute statistical measures across variables, while the String functions manipulate string variable values.
2) Once you have selected a function group, all of the functions in that group are listed in the *Functions and Special Variables* list box.

3) When you select a function from that list, a brief description of the function appears to the left of the functions list. This is a very useful way to review all of the functions available in PASW Statistics as well.

Figure 8.4 Functions in the Compute Variable Dialog



8.9 *Demonstration: Computing with Functions*

In these examples we will compute the mean of a set of variables. There are two different ways to compute the mean of two or more variables.

- The first is to create a numeric expression which sums together variables and then divides by the number of variables in the set. However, the presence of missing data complicates the calculation.
- The second approach is to use the MEAN statistical function, which calculates the mean of a set of one or more arguments (variables).

To demonstrate, we will calculate the mean socioeconomic index for the set of four variables *RESPONDENT SOCIOECONOMIC INDEX [sei], R'S SPOUSE'S SOCIOECONOMIC INDEX [spsei], R'S FATHER'S SOCIOECONOMIC INDEX [pasei]*, and *R'S MOTHER'S SOCIOECONOMIC INDEX [masei]*. Socioeconomic index values vary from 1 to 100.

We will calculate the mean with and without the Mean function to demonstrate the difference.

Detailed Steps for Computing the Mean with and without a Function

1) In the Variable Name box enter the name ***sei_mean1***
2) In the Numeric Expression area, construct the expression **(*sei* + *spsei* + *pasei* + *masei*)/4**

After *sei_mean1* is created, return to the **Compute Variable** dialog.

3) In the Variable Name box enter the name ***sei_mean2***
4) In the Numeric Expression area, insert the **Mean** function
5) Then construct the expression **Mean(*sei,spsei,pasei,masei*)**

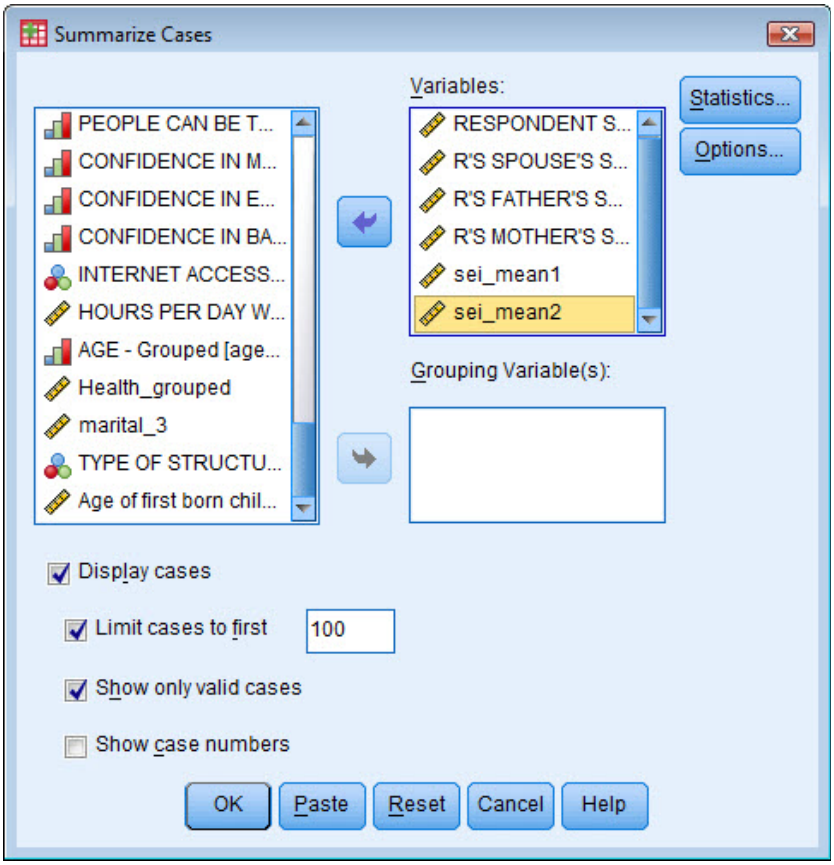
Note that a comma should be inserted between each argument of a function.

Results of Computing the Mean with and without a Function

The two new variables are added to the end of the Data Editor. An effective method to check on the results of a Compute is to use the **Case Summaries** procedure, available in the *Analyze…Reports…Case Summaries* menu. This procedure lists out the values of variables on a case-by-case basis. We simply provide the steps here to create the necessary output but don't detail the options in **Case Summaries**.
With the **Case Summaries** dialog open:

1) Place the variables ***sei***, ***spsei***, ***pasei***, ***masei***, ***sei_mean1*** and ***sei_mean2*** in the Variable(s) box

Figure 8.5 Summarize Cases Dialog



We see from the Case Summaries table, a portion of which is shown in the next figure, that any case that has a missing value on at least one of the four index variables is assigned the system-missing value for *sei_mean1*. (The value of -1 has been assigned the label "IAP" for user-missing values on the four sei variables.) Thus, the variable *sei_mean1* was calculated only for a subset of the cases.

Figure 8.6 Case Summaries Report on New Computed Mean Variables

Case Summaries[a]

	RESPONDENT SOCIOECONOMIC INDEX	R'S SPOUSE'S SOCIOECONOMIC INDEX	R'S FATHER'S SOCIOECONOMIC INDEX	R'S MOTHER'S SOCIOECONOMIC INDEX	sei_mean1	sei_mean2
1	28	IAP	28	33	.	29.97
2	76	63	IAP	28	.	55.70
3	38	IAP	IAP	38	.	37.90
4	35	IAP	IAP	38	.	36.40
5	29	IAP	34	26	.	29.70
6	88	79	36	IAP	.	67.60
7	33	IAP	29	29	.	30.33
8	76	IAP	37	IAP	.	56.85
9	81	IAP	37	IAP	.	59.10
10	37	IAP	72	IAP	.	54.25
11	74	73	28	28	50.63	50.63
12	73	IAP	36	23	.	44.07
13	80	IAP	30	38	.	49.50
14	63	69	46	73	62.88	62.88
15	51	IAP	50	IAP	.	50.35
16	87	33	IAP	IAP	.	60.10
17	55	IAP	85	IAP	.	69.60

Unlike the first numeric expression, the numeric expression with the Mean function removes variables containing missing values from the equation and computes the mean based on the remaining variables' values. If only one variable has a valid value for a case, the computed variable will have the value of that variable. There is no missing data for *sei_mean2* in the file because the variable *sei* has a valid value for every respondent.

Note

You may have noticed that the mean doesn't always equal the value you might have expected, as for the third case in the table. This is because the socioeconomic index variables are measured to greater precision than integers, but the display format for these variables is set to show no decimal digits. Try increasing the number of decimal digits as a check.

Specifying Minimum Number of Valid Values for a Function

You can modify the behavior of the MEAN function, and many other functions, to specify the minimum number of valid values required in order to calculate the mean. For example, it would make some sense to calculate the mean socioeconomic index based on a minimum of 3 variables since spouse's sei is missing for all unmarried respondents.

The numeric expression would then be:

MEAN.3 (educ, maeduc, paeduc, speduc).

where a number is added, with a period, after the function name and before the left parenthesis. The 3 means that the calculation will only be done when at least 3 values are valid.

All of the functions in the Statistical Function group have this option to control the calculations when missing data is present.

Apply Your Knowledge

1. True or false? Specifying MEAN X) for a Numeric Expression in the **Compute Variable** dialog box gives the same result as running Frequencies and requesting the mean for X?

8.10 *Conditional Compute*

By default, **Compute Variable** will include all cases in the calculation. There are situations, however, when the computation needs to be different for separate groups of cases. For instance, suppose we want to compute the difference between the number of years of education of the respondent and that of his father or mother. Now, one could argue that a fair way to compute this difference is to compare the educational years of a male respondent with his father's years of education, while the difference for a female respondent should be computed looking at her mother's years of education. So, the computations are:

educ_difference = educ – paeduc if sex = 1 (computation for males)
educ_difference = educ – maeduc if sex = 2 (computation for females)

Thus, the expression is not the same for all cases.

8.11 *Demonstration: Conditional Compute*

Requesting a conditional **Compute Variable** requires first specifying the normal elements for the computation (variable name and expression). Then we specify the logical condition. We will create the education difference variable discussed above.

Detailed Steps to do a Conditional Compute

1) In the Variable Name box enter the name ***educ_difference***
2) In the Numeric Expression area, construct the expression **educ – paeduc**
3) Select the **If** button
4) Select **Include if case satisfies condition**
5) Create the expression ***sex*=1**

We must complete the **Compute Variable** to create the new variable for males. We cannot create both conditions for males and females simultaneously in the dialog box.

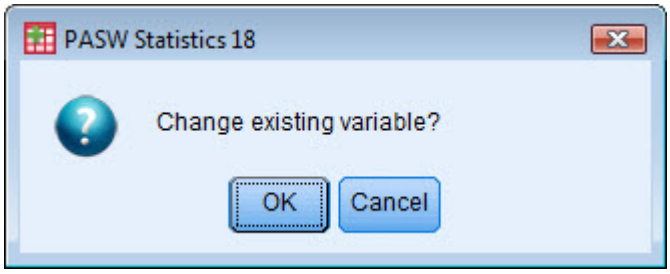
Then we do a similar calculation for females.

6) In the Numeric Expression area, construct the expression **educ – maeduc**

7) Select the **If** button
8) Create the expression ***sex*=2**

When the **Compute Variable** dialog box for females has been completed and the user selects OK, a Warning box appears asking the user if an existing variable can be changed. You can answer OK.

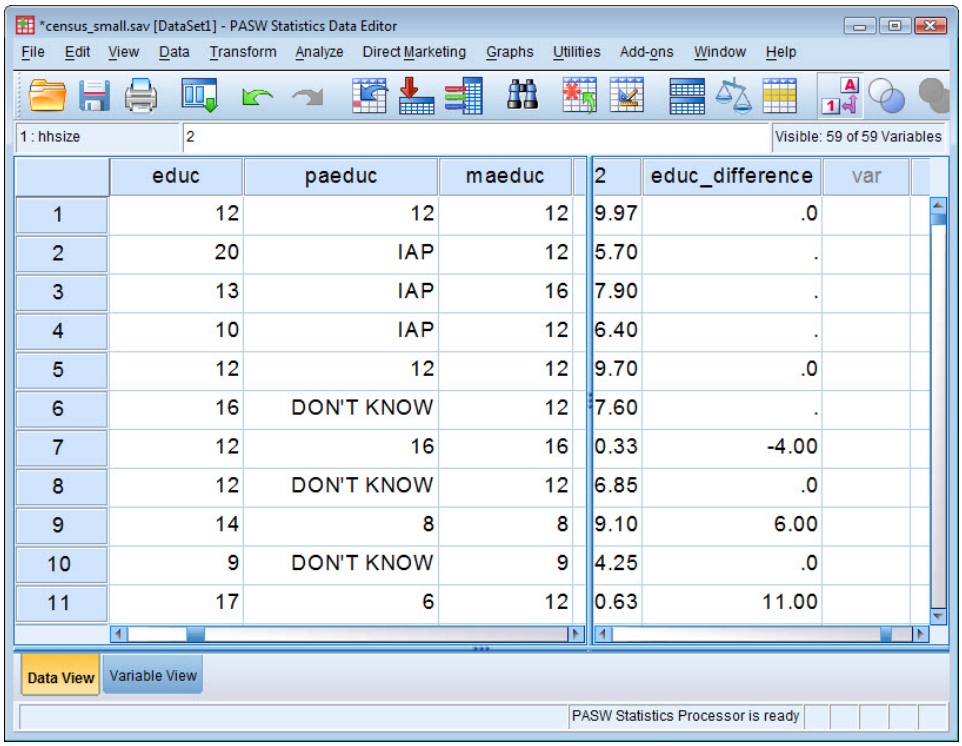
Figure 8.7 Compute Variable Warning to Change Existing Variable



Results from Conditional Compute

We can view the results of the computes in the Data Editor. Note that for the case in row 11, the value of *educ_difference* is 11, which means this respondent must be a male, since the difference between *educ* and *paeduc* is 11.
As usual, the computation can only be done when data are not missing for both variables in the expression. But for the case in row 8, the value of *paeduc* is missing, but the value of *educ_difference* is 0, so this must be a female.

Figure 8.8 New Variable *educ_difference* in Data Editor



	educ	paeduc	maeduc	2	educ_difference	var
1	12	12	12	9.97	.0	
2	20	IAP	12	5.70	.	
3	13	IAP	16	7.90	.	
4	10	IAP	12	6.40	.	
5	12	12	12	9.70	.0	
6	16	DON'T KNOW	12	7.60	.	
7	12	16	16	0.33	-4.00	
8	12	DON'T KNOW	12	6.85	.0	
9	14	8	8	9.10	6.00	
10	9	DON'T KNOW	9	4.25	.0	
11	17	6	12	0.63	11.00	

8.12 *Lesson Summary*

Lesson Objectives Review

Students who have completed this lesson should be able to:

- Use the **Compute Variable dialog** to create new variables

And, they should also be able to:

- Describe the features of **Compute Variable**
- Create new variables with numeric expressions
- Create new variables with conditional numeric expressions

8.13 *Learning Activity*

In this exercise you will use the PASW Statistics data file *census_small.sav*.

Supporting Materials

The file *census_small.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

1. Use **Compute Variable** to create a new variable that contains the difference between the age when a respondent's first child was born (*agekdbrn*) and the age when the respondent was first married (*agemar*). Call it *agediff.*

2. After the variable is created, run Frequencies for *agediff*. How many valid values does *agediff* have? What is the smallest value? The largest? What do negative values imply?

3. There are four variables in the data that ask the respondent about support for national spending on various social issues or problems, such as mass transportation. They are *natroad*, *natmass*, *natsci*, and *natchld*. They are measured on three-point scale, from 1 to 3. Calculate the mean of these four variables using an equation. Call the new variable *nat4*.

4. Check your work on this variable by using the **Case Summaries** procedure. Then run Frequencies. How many valid values are there for *nat4*? What is the mean of *nat4*?

5. Now calculate the mean using the Mean function, requesting that the mean be calculated if only 2 values are valid. Call the new variable *nat2*.

6. Run Frequencies for *nat2*. How many valid values does it have, and what is its mean? How close are the averages for *nat4* and *nat2*?

7. In the lesson we calculated the age of a respondent's oldest child using the variables *age* and *agekdbrn*. We did this for everyone in the file. Now we want you to make this calculation only for respondents who have never been married. Marital status is recorded in the variable *marital*, and the value 5 indicates "Never married." Name the new variable *childage*.

8. After you've computed the new variable, run Frequencies for *childage*. How many respondents have had a child but have never been married?

Lesson 9: Describing Relationships Between Variables

9.1 *Objectives*

After completing this lesson students will be able to:

- Analyze relationships between categorical variables and between categorical and scale variables

To support the achievement of this primary objective, students will also be able to:

- Select the appropriate procedure to summarize the relationship between two variables
- Use the Crosstabs procedure to summarize the relationship between categorical variables
- Use the Means procedure to summarize the relationship between a scale and a categorical variable

9.2 *Introduction*

We have already covered techniques for describing individual variables in the lesson *Summarizing Individual Variables*. In most studies we will also be interested in summarizing and describing relationships *between* variables. For example, we can ask:

- *Does a certain type of customer buy a certain type of product?*
- *Is people's trust in others related to their age?*
- *Does the number of hours spent watching TV differ by degree of education or by age?*

These questions can only be answered by looking at the relationship between variables. There are a variety of techniques that allow us to summarize and answer questions of relationship and statistical significance for multiple variable relationships. In this lesson we will look at two of the most common techniques: crosstabulation and subgroup mean summary tables. Crosstabulations (CROSSTABS procedure) are used to describe the relationship between two or more categorical variables. Mean summary tables (MEANS procedure) are used to describe the relationship of a scale variable for groups defined by a category variable.

Business Context

Studying relationships between variables is often the central focus of an analysis. As examples:

- We may wish to study how customer type is related to revenue
- We may wish to study which students are more likely to complete a degree program
- We may wish to study whether males or females are more likely to offer support for a government policy

Supporting Materials

The *Census_Small.sav* PASW Statistics data file is used in this lesson. These data are a subset of demographic and attitudinal variables from a survey conducted in 2008 of a sample of the general population.

9.3 *Requesting a Crosstabulation*

Requesting **Crosstabs** for two categorical variables is completed with these steps:

1) Select a variable for the rows of the table
2) Select a variable for the columns of the table
3) Select appropriate percentages or other statistics
4) Select statistics to test or summarize the relationship between the two variables
5) Optionally, add variables to the layers of the table to create multi-way crosstabulations

9.4 *Output from a Crosstabulation*

Crosstabulation tables (also called contingency tables) display the relationship between two or more categorical (nominal or ordinal) variables. The size of the table is determined by the number of distinct values for each variable, with each cell in the table representing a unique combination of values. For instance, the crosstabulation of *HAPPINESS OF MARRIAGE* by *RESPONDENTS SEX* is shown in the figure below.

Here we have *HAPPINESS OF MARRIAGE* in the row crossed with *RESPONDENTS SEX* in the column. The Total column and Total row, often referred to as the Marginal Totals, give the total frequency counts for each variable.

Figure 9.1 Sample Crosstabulation

HAPPINESS OF MARRIAGE * RESPONDENTS SEX Crosstabulation

Count

		RESPONDENTS SEX		
		MALE	FEMALE	Total
HAPPINESS OF MARRIAGE	VERY HAPPY	297	299	596
	PRETTY HAPPY	160	183	343
	NOT TOO HAPPY	13	17	30
Total		470	499	969

Note

Actually, the counts in the table are based on the number of cases with a valid value on both variables, so the marginal totals may, or may not, be equal to the same counts from the Frequencies procedure.

It is often difficult to analyze a crosstabulation simply by looking at the counts in each cell. In order to interpret the relationship, we need to examine percentages. In the table above, the fact that there are more NOT TOO HAPPY women than NOT TOO HAPPY men may not mean much (or anything) since there are also more women in the data (499 women against 470 men). In order to interpret the relationship, we need to examine the percentages within each gender, as we want to make statements like: *"…% of the women are not too happy in their marriage, while …% of the men are not too happy*". In other words, we need to compute the percentages columnwise, within each category of sex.

The next figure shows the table with column percentages added. Studying the percentages in the NOT TOO HAPPY category, there is not much difference between men and women (2.8% versus 3.4%) and the same goes for the other categories.

Figure 9.2 Crosstabulation with Column Percentages

HAPPINESS OF MARRIAGE * RESPONDENTS SEX Crosstabulation

			RESPONDENTS SEX		
			MALE	FEMALE	Total
HAPPINESS OF MARRIAGE	VERY HAPPY	Count	297	299	596
		% within RESPONDENTS SEX	63.2%	59.9%	61.5%
	PRETTY HAPPY	Count	160	183	343
		% within RESPONDENTS SEX	34.0%	36.7%	35.4%
	NOT TOO HAPPY	Count	13	17	30
		% within RESPONDENTS SEX	2.8%	3.4%	3.1%
Total		Count	470	499	969
		% within RESPONDENTS SEX	100.0%	100.0%	100.0%

Statistical Tests

The statement “not too much difference” can be formalized by performing a statistical test (considering the data to be a random sample of married people from our population of interest). The most popular test in a crosstabulation is the Pearson Chi-Square test, shown in the next figure.

The Pearson Chi-Square tests the hypothesis that the row variable and column variable are independent (unrelated). The actual value of the statistic isn't very informative. The significance value (Asymp. Sig.(2-sided)) has the information we're looking for. Loosely speaking, this gives the probability that the variables are independent (unrelated). The lower this probability, the less likely it is that the two variables are independent (unrelated). Many analysts use a cut off for significance values less than 0.05 to reject the null hypothesis of independence and consider the differences seen in the table as statistically significant. In this case, the significance value is higher than 0.05, which means it would appear that the two variables are unrelated and so there are no differences between men and women with respect to happiness in marriage.

Figure 9.3 Chi-Square Test for Crosstabulation

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.216[a]	2	.545
Likelihood Ratio	1.217	2	.544
Linear-by-Linear Association	1.214	1	.271
N of Valid Cases	969		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 14.55.

Further Information

The course *Statistical Analysis using PASW Statistics* has a detailed discussion of this test (and statistical testing in general).

Layer Variables

A layer variable can be added to create a three-way table in which categories of the row and column variables are further subdivided by categories of the layer variable. This variable is sometimes referred to as the control variable because it may reveal how the relationship between the row and column variables changes when you "control" for the effects of the third variable. In the table below the control variable *WAS R BORN IN THIS COUNTRY* was added.

Figure 9.4 Layer Variable Added to Crosstabulation

HAPPINESS OF MARRIAGE * RESPONDENTS SEX * WAS R BORN IN THIS COUNTRY Crosstabulation

WAS R BORN IN THIS COUNTRY				RESPONDENTS SEX		
				MALE	FEMALE	Total
YES	HAPPINESS OF MARRIAGE	VERY HAPPY	Count	265	260	525
			% within RESPONDENTS SEX	64.8%	61.6%	63.2%
		PRETTY HAPPY	Count	132	148	280
			% within RESPONDENTS SEX	32.3%	35.1%	33.7%
		NOT TOO HAPPY	Count	12	14	26
			% within RESPONDENTS SEX	2.9%	3.3%	3.1%
	Total		Count	409	422	831
			% within RESPONDENTS SEX	100.0%	100.0%	100.0%
NO	HAPPINESS OF MARRIAGE	VERY HAPPY	Count	32	39	71
			% within RESPONDENTS SEX	52.5%	50.6%	51.4%
		PRETTY HAPPY	Count	28	35	63
			% within RESPONDENTS SEX	45.9%	45.5%	45.7%
		NOT TOO HAPPY	Count	1	3	4
			% within RESPONDENTS SEX	1.6%	3.9%	2.9%
	Total		Count	61	77	138
			% within RESPONDENTS SEX	100.0%	100.0%	100.0%

Looking at the percentages (now within each combination of *RESPONDENTS SEX* and *WAS R BORN IN THIS COUNTRY* as the base for computing percentages) there does not appear to be a difference between men and women, so the location of a respondent's birth doesn't seem to have an effect on the relationship between sex and happiness in marriage. Again, this could be studied more formally by carrying out the Chi-Square test (or more advanced techniques) for each of the subtables.

Apply Your Knowledge

1. The variable *sex* has two values and the variable *region* has four values. If we request a crosstabulation of *sex* with *region*, how many cells will the table have?
    a. 4

b. 6
c. 8

2. Can you explain why it is better to use percentages in a crosstabulation than counts for studying whether two variables are related?

9.5 *Procedure: Crosstabulation for Categorical Variables*

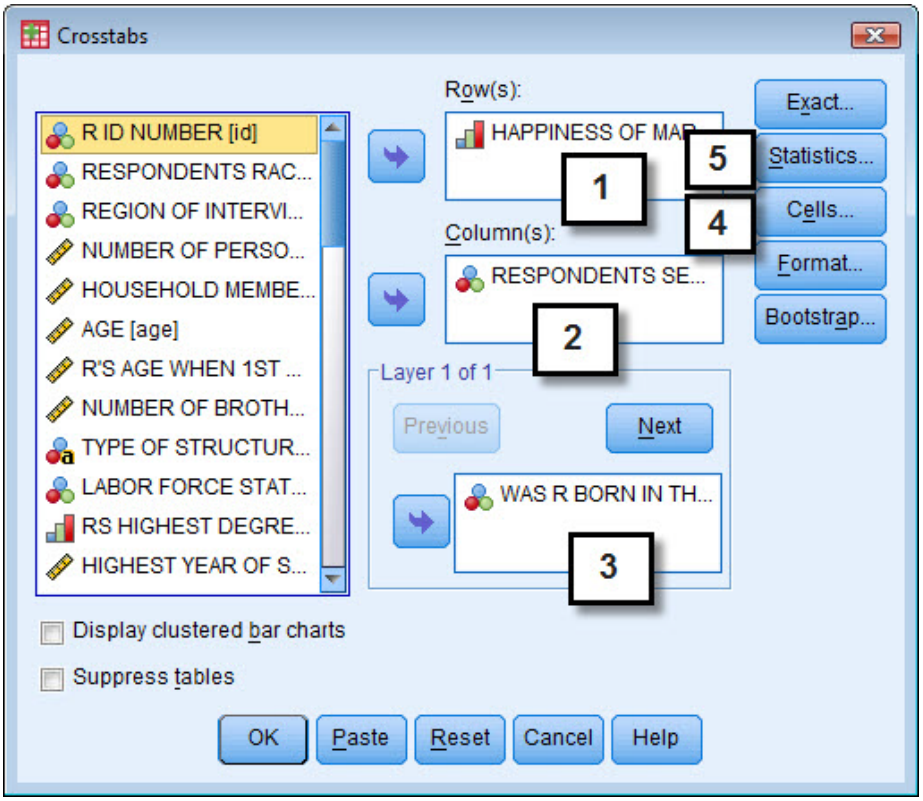
The **Crosstabs** procedure is accessed from the *Analyze…Descriptive Statistics…Crosstabs* menu choice. The minimum specification is one variable in the Row(s) box and one variable in the Column(s) box. In the opening dialog box:

1) Place one or more variables in the Row(s) box
2) Place one or more variables in the Columns(s) box
3) Optionally, place one or more variables in the Layer(s) box
4) Select the *Cells* button to request percentages and related statistics
5) Select the *Statistics* button to request tests of the relationship (such as Chi-Square)

Further Information

Additional layer variables can be added to the table by clicking the Next button. However, larger tables are difficult to interpret and even with large data sets, your data may be so sparsely distributed among the cells that the relationship can no longer be easily interpreted.

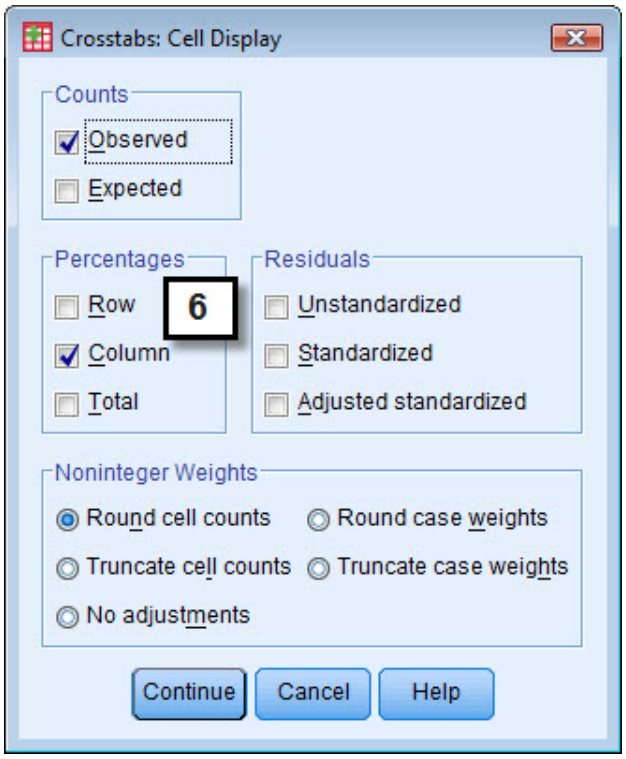
Figure 9.5 Crosstabs Main Dialog



Different types of percentages can be selected in the Cell Display subdialog box. Although all three types of percentages can be requested, this is normally not done, as it makes the table more difficult to read. Normally, you would select the one percentage that answers the question of interest.

6) Select the appropriate percentage in the Percentage group area

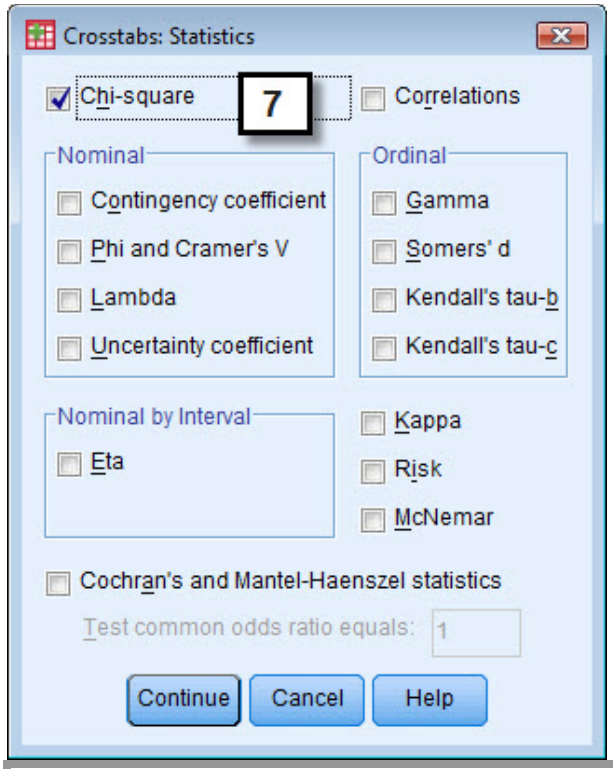
Figure 9.6 Crosstabs Cell Display Subdialog



The Statistics subdialog box provides a number of measures, including the Chi-square statistic. Note that the statistics are categorized according to the level of measurement of the variables involved.

7) Select *Chi-square* check box

Figure 9.7 Crosstabs Statistics Subdialog



Further Information

Discussion of the many statistics is beyond the scope of this lesson. See *Statistical Analysis using PASW Statistics* for a review. Note that again the concept of level of measurement is essential for making the right choice.

9.6 *Demonstration: Crosstabs*

We will use the data file *census_small.sav* in these examples. What factors affect whether people trust their fellow human beings? Do women believe that people can be trusted more than their male counterparts do? Are older people more trusting, or those who are better off financially? In this example, we will use Crosstabs to examine the relationship between respondents' belief that people can be trusted and their subjective social class. As we are interested in comparing the trust of working class respondents with those of middle class respondents, etc., we ask for percentages by subjective social class. We will include a statistical test to see if the differences are statistically significant.

Detailed Steps for Crosstabs

1) Place the variable ***PEOPLE CAN BE TRUSTED OR CANT BE TOO CAREFUL [cantrust]*** in the Row(s) box
2) Place the variable ***SUBJECTIVE SOCIAL CLASS [class]*** in the Column(s) box
3) Select the **Cells** button
4) Select **Column** check box
5) Select the **Statistics** button
6) Select **Chi-square** check box

Results for Crosstabs

The first table displayed is the Case Processing Summary. This table shows the number and percentage of cases that have a valid value on both variables (here: 1348), number and percentage of cases that have a missing value on at least one of the variables (675) and the total number of cases in the data set (2023)

Only the cases with valid answers to both questions are included in the crosstabulation table. It is important to review this table, and if there is a substantial amount of missing data, think about how this might affect the analysis.

Figure 9.8 Case Processing Summary Table

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
PEOPLE CAN BE TRUSTED OR CANT BE TOO CAREFUL * SUBJECTIVE CLASS IDENTIFICATION	1348	66.6%	675	33.4%	2023	100.0%

The cells of the crosstab table show the count or number of cases for each joint combination of values. For example, 81 respondents who rate themselves as middle class believe people can always be trusted. Looking at the percentages, there does seem to be some differences by social class when it comes to trusting people. Generally, those of the middle class or upper class appear to be more trusting, and those of the lower or working class appear less so.

Figure 9.9 Crosstab for Trust in People by Subjective Social Class

PEOPLE CAN BE TRUSTED OR CANT BE TOO CAREFUL * SUBJECTIVE CLASS IDENTIFICATION Crosstabulation

			SUBJECTIVE CLASS IDENTIFICATION				
			LOWER CLASS	WORKING CLASS	MIDDLE CLASS	UPPER CLASS	Total
PEOPLE CAN BE TRUSTED OR CANT BE TOO CAREFUL	ALWAYS TRUSTED	Count	9	49	81	6	145
		% within SUBJECTIVE CLASS IDENTIFICATION	6.8%	7.7%	15.1%	13.0%	10.8%
	USUALLY TRUSTED	Count	27	194	205	26	452
		% within SUBJECTIVE CLASS IDENTIFICATION	20.5%	30.6%	38.2%	56.5%	33.5%
	USUAL NOT TRUSTED	Count	68	297	206	10	581
		% within SUBJECTIVE CLASS IDENTIFICATION	51.5%	46.8%	38.4%	21.7%	43.1%
	ALWYS NOT TRUSTED	Count	28	94	44	4	170
		% within SUBJECTIVE CLASS IDENTIFICATION	21.2%	14.8%	8.2%	8.7%	12.6%
Total		Count	132	634	536	46	1348
		% within SUBJECTIVE CLASS IDENTIFICATION	100.0%	100.0%	100.0%	100.0%	100.0%

The Chi-Square test answers the question whether the observed differences can be attributed to random variation (random variation arising from sampling).

In this case, the significance value is so low that it is displayed as .000, which means that the two variables are, indeed, related. The relationship is as described above.

Figure 9.10 Chi-Square Test for Crosstab

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	66.962[a]	9	.000
Likelihood Ratio	67.293	9	.000
Linear-by-Linear Association	56.089	1	.000
N of Valid Cases	1348		

a. 1 cells (6.3%) have expected count less than 5. The minimum expected count is 4.95.

Note

There is an important distinction between statistical and *substantive* significance. For the current table, we might decide that the percentage differences are not large enough to be important in the real-world.

Apply Your Knowledge

1. True or false? We should use both row and column percentages in a crosstab to study the relationship between the variables?

9.7 *Relationships between Scale and Categorical Variables*

Crosstabulation is appropriate for studying the relationship between two categorical variables. If the interest is in studying how a categorical variable relates to a scale variable, the **Means** procedure is used. **Means** is used to display descriptive statistics, such as means and standard deviations, of a scale variable for subgroups of cases as defined by the values of one or more categorical variables.

9.8 *Requesting a Means Procedure*

Requesting **Means** for a scale and a categorical variable is completed with these steps:

1) Select one or more scale variables on which to calculate summary statistics
2) Select one or more categorical grouping variables
3) Select appropriate percentages or other statistics

9.9 *Output from Means*

As an example, if we want to see how age varies by political position—from liberal to conservative—we would use the **Means** procedure.

By default, the mean, number of cases and standard deviation are displayed for each subgroup (category) of the categorical variable, as well as for the total. In the table below, 1,734 people answered both questions and appear in the table. We

can see that on average, those with a CONSERVATIVE political position were the oldest (51.57), while the youngest group (41.51) were those who are SLIGHTLY LIBERAL. There is no obvious trend in age as we move from liberal to conservative, though conservatives seem to be a few years older.

Statistical tests are available in the Means Statistics dialog to formally test the statistical significance of the observed differences.

Figure 9.11 Sample Means Table

Report

AGE

THINK OF SELF AS ···	Mean	N	Std. Deviation
EXTREMELY LIBERAL	47.19	62	17.808
LIBERAL	45.60	214	17.181
SLIGHTLY LIBERAL	41.51	201	15.505
MODERATE	48.06	648	17.567
SLGHTLY CONSERVATIVE	48.88	247	16.948
CONSERVATIVE	51.57	302	17.400
EXTREMELY CONSERVATIVE	49.92	60	16.676
Total	47.76	1734	17.357

Similar to Crosstabs, additional independent (categorical) variables can be added as layers in the same table; for instance, we might want to see how men and women differ by age within each political group.

In the table below, differences in age by gender are not large, and they don't seem to have any pattern.

Figure 9.12 Sample Means Table with Gender Added As Layer

Report

AGE

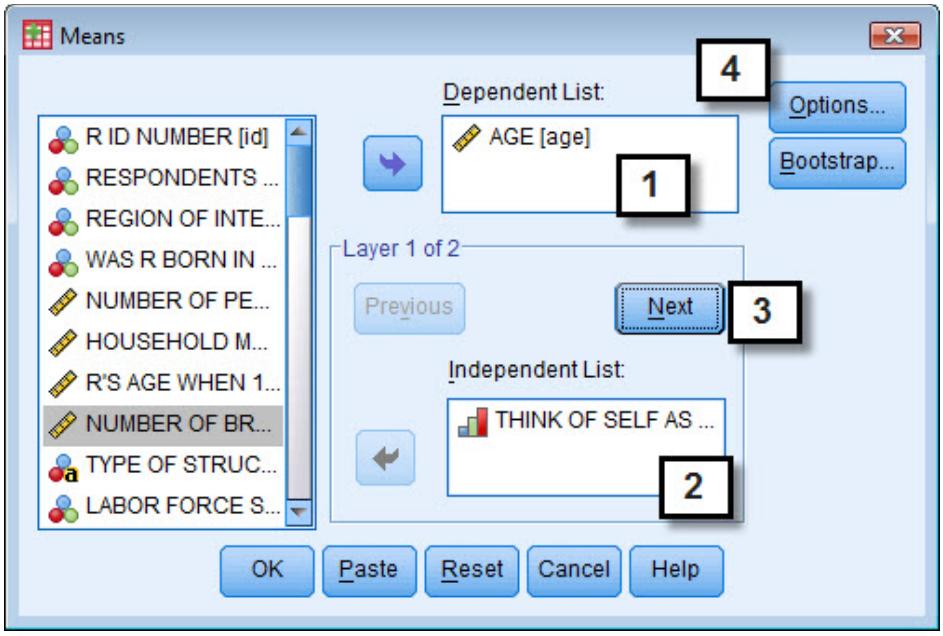
THINK OF SELF AS ···	RESPONDENTS SEX	Mean	N	Std. Deviation
EXTREMELY LIBERAL	MALE	44.63	30	15.410
	FEMALE	49.59	32	19.738
	Total	47.19	62	17.808
LIBERAL	MALE	46.33	90	15.673
	FEMALE	45.07	124	18.241
	Total	45.60	214	17.181
SLIGHTLY LIBERAL	MALE	40.34	90	14.977
	FEMALE	42.45	111	15.924
	Total	41.51	201	15.505
MODERATE	MALE	47.55	304	16.527
	FEMALE	48.52	344	18.449
	Total	48.06	648	17.567
SLGHTLY CONSERVATIVE	MALE	47.51	119	16.821
	FEMALE	50.16	128	17.031
	Total	48.88	247	16.948
CONSERVATIVE	MALE	53.25	143	17.538
	FEMALE	50.06	159	17.191
	Total	51.57	302	17.400
EXTREMELY CONSERVATIVE	MALE	48.36	33	15.908
	FEMALE	51.81	27	17.685
	Total	49.92	60	16.676
Total	MALE	47.54	809	16.738
	FEMALE	47.95	925	17.887
	Total	47.76	1734	17.357

9.10 ***Procedure: Means***

The **Means** procedure is accessed from the *Analyze…Compare Means…Means* menu choice.

1) Place one or more variables in the *Dependent List* box
2) Place one or more variables in the *Independent List* box
3) Optionally, place one or more variables in the Layer(s) by selecting the *Next* button
4) Select the *Options* button to request additional statistics and a statistical test of the mean differences

Figure 9.13 Means Main Dialog



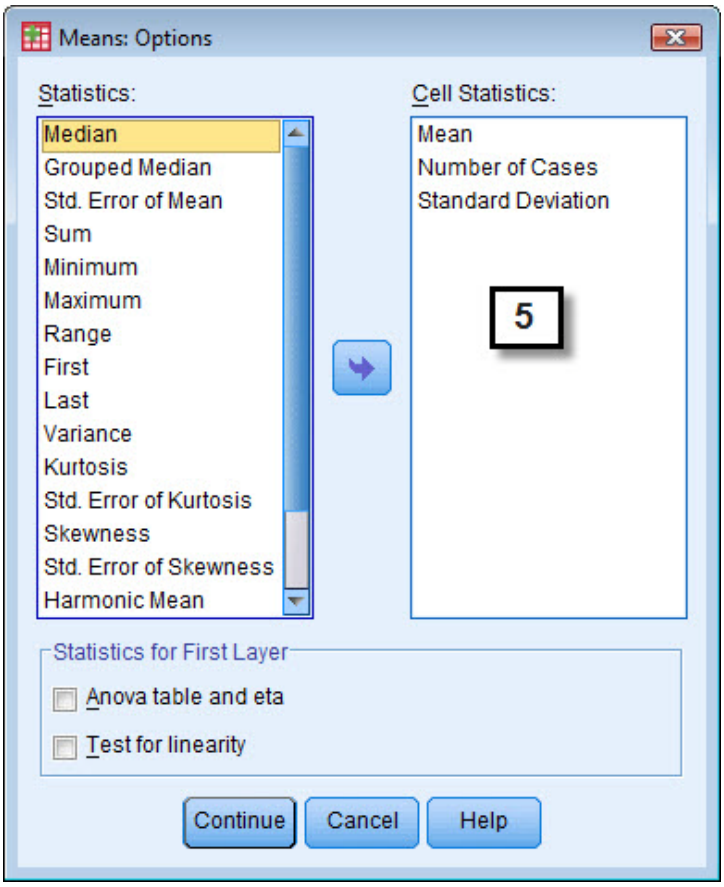
Further Information

Many layer variables can be added to the table by clicking the Next button. However, larger tables are difficult to interpret and even with large data sets, there may be so few cases in categories that the relationship can no longer be assessed.

The statistics reported can be customized in the Options subdialog box.

5) Select appropriate statistics in the Statistics: list

Figure 9.14 Means Options Subdialog



An ANOVA table can be requested for the first layer variable to see whether the mean differences across the categories are statistically significant.

Further Information

Discussion of the ANOVA test is beyond the scope of this lesson. See *Statistical Analysis using PASW Statistics* for an explanation and examples.

9.11 *Demonstration: Means*

For an example of the Means procedure, we will look at mean differences in hours watching television per day by labor force status. Does having more time available—for those retired or unemployed—lead to more television viewing? We will use the default statistics.

Detailed Steps for Means

1) Place the variable ***HOURS PER DAY WATCHING TV [tvhours]*** in the Dependent List box
2) Place the variable ***LABOR FORCE STATUS [wrkstat]*** in the Independent List box

Results for Means

The first table displayed is Case Processing Summary (not shown); this table displays the number of included and excluded cases and is analogous to the same table discussed with Crosstab.

The table labeled Report shows the means.

Figure 9.15 Mean Hours Watching Television per Day by Labor Force Status

Report

HOURS PER DAY WATCHING TV

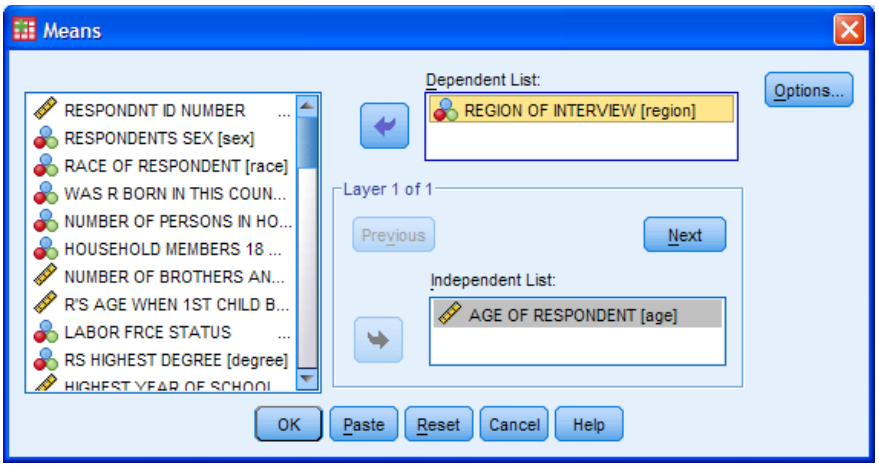
LABOR FORCE STATUS	Mean	N	Std. Deviation
WORKING FULLTIME	2.37	673	1.924
WORKING PARTTIME	2.76	140	2.865
TEMP NOT WORKING	3.71	38	3.952
UNEMPL, LAID OFF	4.06	48	4.159
RETIRED	3.93	202	2.919
SCHOOL	2.50	36	2.793
KEEPING HOUSE	3.88	149	2.807
OTHER	4.50	36	3.598
Total	2.98	1322	2.661

By default the mean, number of cases and standard deviation are displayed for each category of *wrkstat* as well as for the total. In this table, 1322 people answered both questions and appear in the table. We can see that on average, the group OTHER watched the most TV (4.50) while the second highest average was for the UNEMPLOYED, LAID OFF (4.06). And those RETIRED also watched almost four hours of TV per day (3.93).

In contrast, the WORKING FULLTIME group spent the fewest number of hours (2.37) watching TV. It seems reasonable that this group is more likely to have other responsibilities that afford them less time in front of the TV, and in general, those with a job or going to school watch less TV than others.

Apply Your Knowledge

1. True or false? Will this Means dialog box below give us a table of the mean age by region?



9.12 *Lesson Summary*

In this lesson we discussed and demonstrated how to examine the relationship between categorical variables, and between a scale and a categorical variable, using the **Crosstabs** and **Means** procedures, respectively.

Lesson Objectives Review

Students who have completed this lesson should now be able to:

- Analyze relationships between categorical variables and between categorical and scale variables

And, they should also be able to:

- Select the appropriate procedure to summarize the relationship between two variables
- Use the Crosstabs procedure to summarize the relationship between categorical variables
- Use the Means procedure to summarize the relationship between a scale and a categorical variable

9.13 *Learning Activity*

In this exercise you will use the PASW Statistics data file *employee data.sav* to explore the relationships between several variables.

Supporting Materials

The data file *employee data.sav* contains information on employees of a major bank. Included are data on beginning and current salary position, time working, and demographic information.

1. Run a crosstabulation of the variables *gender* and *jobcat*. Request percentages based on *gender*. What percentage of men are managers? What percentage of females? What percentage of men and women are custodial workers?

2. Request a chi-square test for this table? Is it significant at the .05 level? What does that tell us about how job category is related to gender?

3. Now rerun the crosstab table, requesting percentages based on *jobcat*. What percentage of managers is female? What percentage is male? Think about the different questions being answered depending on how a table is percentaged.

4. Examine the relationship between *jobcat* and minority status (*minority*). Percentage the table by *minority*. Is there a significant relationship as assessed by chi-square? How would you describe the relationship?

5. Use the Means procedure to study the relationship between *salary* and *jobcat*, and *salary* by *gender* (separate analyses). Request the median, minimum, and maximum statistics in addition to the defaults. Are the patterns the same for both the mean and median? How much more salary do men make than women, on the average?

6. *For those with extra time:* Create a three-way crosstab table of *gender*, *minority*, and *jobcat*. Use *gender* as the layer variable. Does the relationship between *minority* and *jobcat* vary by *gender*? How would you describe it? Add a chi-square test to the analysis. Does this support your interpretation of the subtables?

7. *For those with more extra time:* Add the variable *minority* as a second layer to the Means table for *salary* and *jobcat*. Are there differences in average salary by minority status within job category?

Lesson 10: Selecting Cases

10.1 *Objectives*

After completing this lesson students will be able to:

- Select cases in a data file using various methods

To support the achievement of this primary objective, students will also be able to:

- Describe and use the features of the Select Cases dialog
- Describe and use the features of the Split File dialog

10.2 *Introduction*

In most analyses, there comes a point at which reports or charts must be created on subgroups of cases. Or, we might need to repeat the same analysis on several subgroups. For example, you might want to produce a chart displaying only the students in one specific school, or you might want to produce
the same chart for each of the schools in the school district. PASW Statistics has two features, **Select Cases** and **Split File**, which provide these functions.
We will demonstrate both in this chapter.

Business Context

The topics in this lesson can help in answering questions such as:

- How can I temporarily select/filter cases?
- How can I create a dataset with only a subset of cases?
- How can I repeat an analysis for subgroups, without having to select each subgroup separately?

Supporting Materials

The file *census_small.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

10.3 *Selecting Cases*

PASW Statistics can filter cases, retain only selected cases in a data file (in a Data Editor), or write out the selected cases to a new PASW Statistics dataset.

- When filtering is invoked, all cases remain in the Data Editor window, but only those meeting the criteria are used when running statistical or graphical procedures.
- If deletion is requested, only selected cases remain in the Data Editor window.
- Or, you can write the selected cases to a new PASW Statistics dataset (Data Editor window) which leaves the original dataset unchanged.

If you want to perform an analysis or two on a subgroup, use filtering; if you will be doing extensive analyses with selected cases, deletion or writing the selected cases to a new dataset is usually more efficient.

10.4 *Requesting Select Cases*

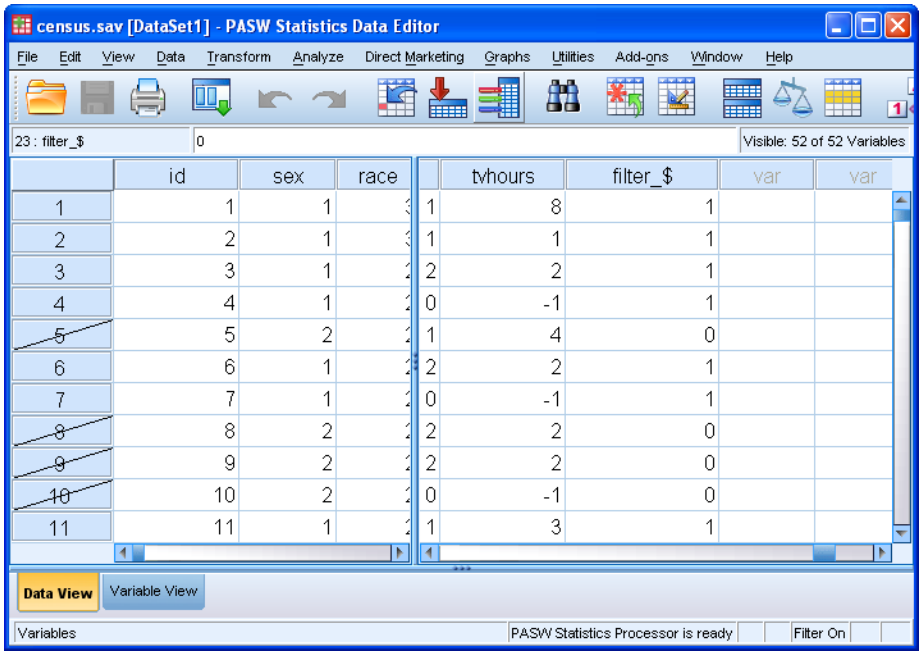
Selecting cases is accomplished with these steps:

1) Specify the subset of cases, by specifying a condition that has to be satisfied, by drawing a sample of cases, by specifying a case range, or by using a filter variable.
2) Specify the output destination for the selected cases.
3) Run the analyses for the selected cases.
4) Optionally, undo the selection.

10.5 *Select Cases Output*

The result of the selection depends on what output destination is selected. PASW Statistics can filter cases, retain only selected cases, or write out the selected cases to a new PASW Statistics dataset. For example, if cases are filtered, all cases remain in the Data Editor window, but only those meeting the criteria are used when running statistical or graphical procedures. The unselected cases are marked in the Data Editor. As well, the message Filter On is displayed in the information area of the Data Editor. Finally, a variable named *filter_$* will be added to the active dataset, with value 1 for the selected cases and value 0 for the unselected cases.

Figure 10.1 Filtering Cases



Tip

Results of analyses are the same for the filtered, copy into new dataset, and deleted choices. If you want to perform an analysis or two on a subgroup, use filtering; if you will be doing extensive analyses with selected cases, deletion or writing the selected cases to a new dataset is usually more efficient.

Warning

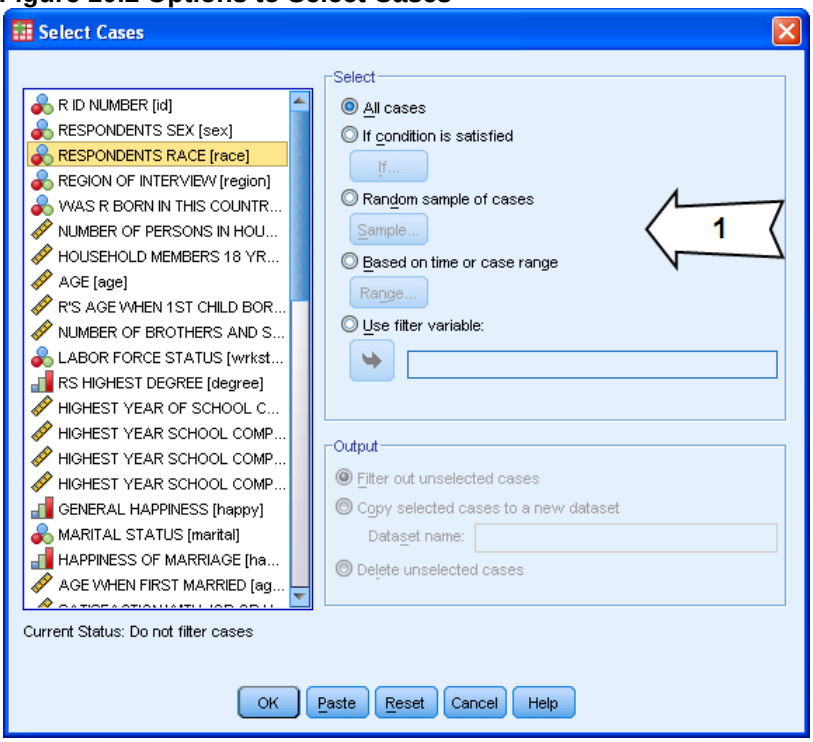
When cases are selected, as with filtering, the tables and charts in the Output Viewer will not indicate that they have been created on a subset of cases. Therefore, it is important to note this or mark the output accordingly.

10.6 *Procedure: Select Cases*

The **Select Cases** procedure is accessed from the **Data…Select Cases** menu choice.

1) In the Select area in the **Select Cases** dialog, there are five choices of how to select cases.

Figure 10.2 Options to Select Cases

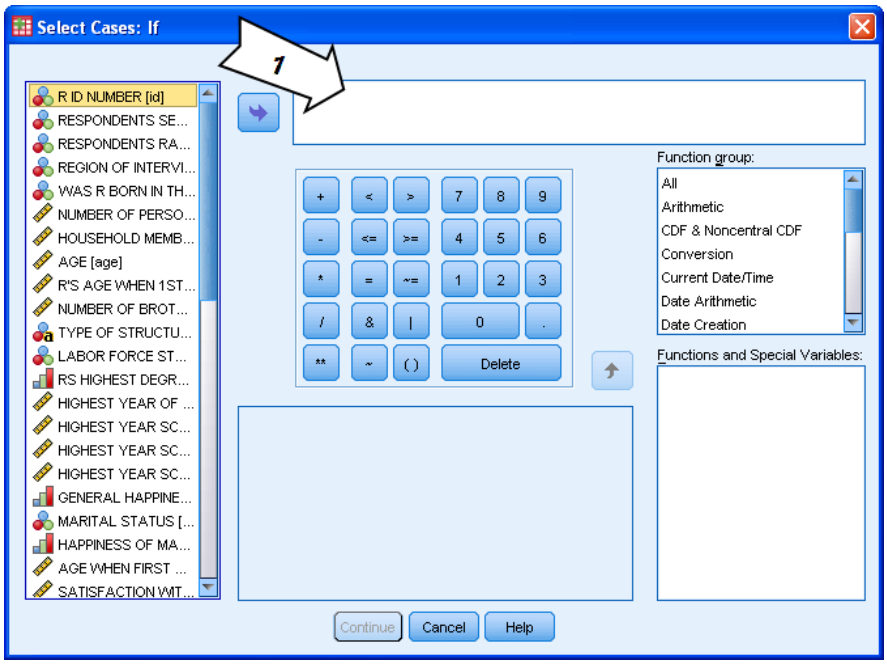


Option 1: All cases. By default, all cases are included in analyses. If cases have been selected, choosing this option will undo the selection and all cases will be included again.

Option 2: If condition is satisfied. Only cases meeting the specified condition will be selected. The condition is constructed by:

1) Selecting the *If condition is satisfied* button
2) Creating an expression in the Select Cases: If dialog

Figure 10.3 Select Cases: If Dialog to Specify a Condition for Case Selection



Warning

Variable names and not variable labels are used in the selection expression text box.

The selection rule can be quite elaborate, involving multiple variables, relational operators (<. <=, =, ~=, >, >=), logical operators (&, !, ~), arithmetic operations (+, -, etc), and a broad range of functions.

Tip

Very useful functions in building the condition are the ANY and RANGE function. These functions provide an easy way to evaluate a list of values for a variable

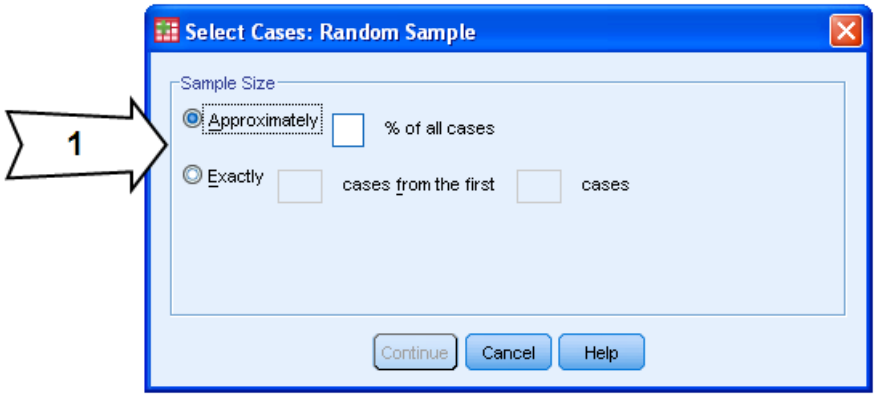
Tip

In general, a case with a missing value (either system- or user missing) will not satisfy a condition. There are special missing value functions, however, to deal with this situation. Refer to the *Data Management and Manipulation* course for a detailed discussion.

Option 3: Random sample of cases. Either a random sample based on an approximate percentage or an exact number of cases can be chosen. Sampling is performed without replacement, so the same case cannot be selected more than once.

1) Specify the method of sampling.

Figure 10.4 Select Cases: Random Sample Dialog

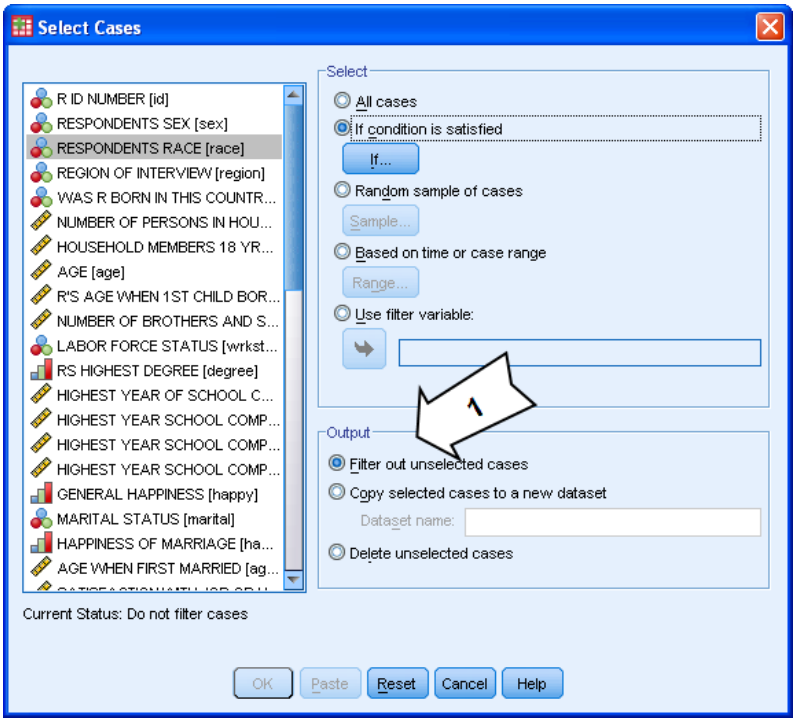


Option 4: Based on time or case range. Selects cases based on a range of case (row) numbers or a range of dates/times.
Option 5: Use filter variable. Based on the values of a numeric variable; cases with values other than 0 or missing will be selected.

How unselected or selected cases are handled is set in the Output area of the dialog box. Three options are available.

1) Specify the output destination for the selected cases.

Figure 10.5 Select Cases Output Options



Option 1: Filter out unselected cases. Unselected cases remain in the Data Editor but will not be used in analyses.

Option 2: Copy selected cases to a new dataset. Selected cases are written to a new Data Editor window. The original dataset is unchanged.

The focus will not automatically be on the new dataset, so switch to this dataset to make it the active window.

Tip

Option 3: **Delete unselected cases**. Unselected cases are deleted from the Data Editor. Thus the Data Editor window retains only the selected cases. This avoids opening a second Data Editor window.

Be careful not to save this selected subset of cases under the same file name as your original data file.

Warning

10.7 *Demonstration: Select Cases with If Condition*

We will work with the *census_small.sav* data file in this lesson. We want to select the subset of respondents who are not currently married *and* are living in a single person household. Those respondents not currently married have values not equal to 1 (MARRIED) for *MARITAL STATUS [marital]*; the household size variable is *NUMBER OF PERSONS IN HOUSEHOLD [hhsize]*.

Detailed Steps for Select Cases with If Condition

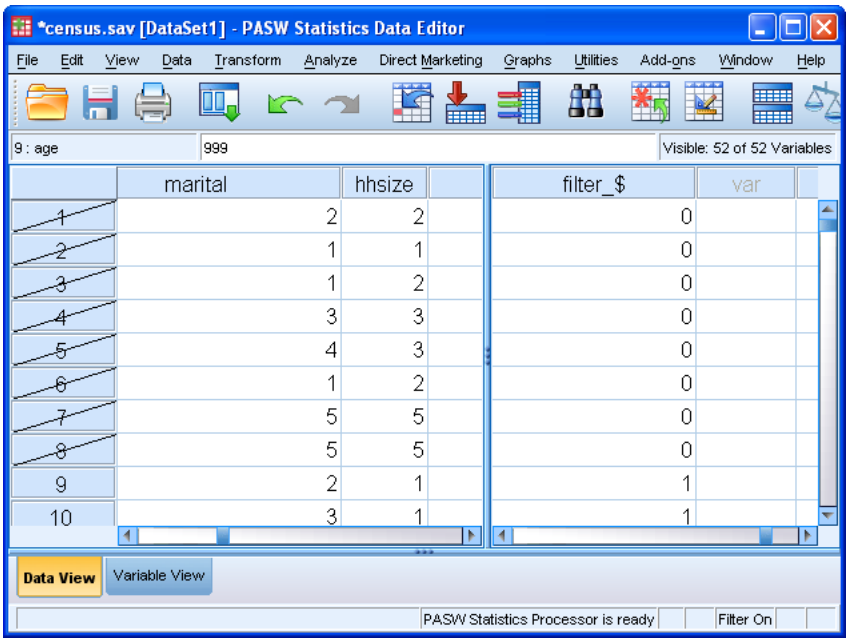
With the **Select Cases** dialog open:

1) Select **If condition is satisfied** button
2) Specify the condition as ***marital* NE 1 and *hhsize*=1**
3) By default, the unselected cases are filtered out. We will leave that choice for this example.

Results from Select Cases with If Condition

A variable named *filter_$* is created, coded 1 or 0 depending on whether or not a case meets the specified criteria (1 indicates it did). The selection rule is included in the variable label of the filter variable. Each unselected case has a slash through its row number to indicate that the data have been filtered. The first cases in the file did not meet the criteria; case 9 and 10 are single and unmarried. The message Filter On appears on the status bar of the Data Editor.

Figure 10.6 Data Editor for Select Cases with If Condition



Tip

If you are creating complicated selection rules, it can be a good idea to double-check that the selection was done correctly. To determine if the selection worked properly for this demonstration, we can crosstabulate *marital* by *hhsize*.

The filter will be in effect until we cancel it by selecting the All Cases option button in the Select Cases dialog box. If the option All Cases is selected, the slashes through the row numbers will disappear and the Filter On message in the Status Bar is removed. The *filter_$* variable remains in the data. This variable could be renamed and the file saved so the variable can be used in the future as a filter variable to select this same subset of cases (choose the Use filter variable option in the Select area).

Note

If a filter is in effect and the data file is saved (File…Save or File…Save as) all cases will be selected and not only the selected cases.

Apply Your Knowledge

1. True or false? If we have selected cases by filtering, we could now delete those cases from the data file without having to respecify the filtering condition?

2. Which of these is not an option for case selection?
   a. A logical condition
   b. A random sample of cases

c. A case range
d. Comparing values of a variable in two files
e. Using a filter variable

10.8 *Split File Processing*

If you need to repeat the same analysis on separate subgroups of cases, one method is to apply a series of filters to the data, each followed by an analysis request. If the subgroups are mutually exclusive, split file processing makes this task much easier. The data file is split into separate groups based on the values of the variable (or variables) that you specify as the grouping variable(s), and results will be presented for each group. So, split file processing streamlines what would otherwise be a tedious set of operations.

There are two choices in split file processing: the results for the groups are presented together in the same table or with separate tables for each group.

10.9 *Requesting Split File*

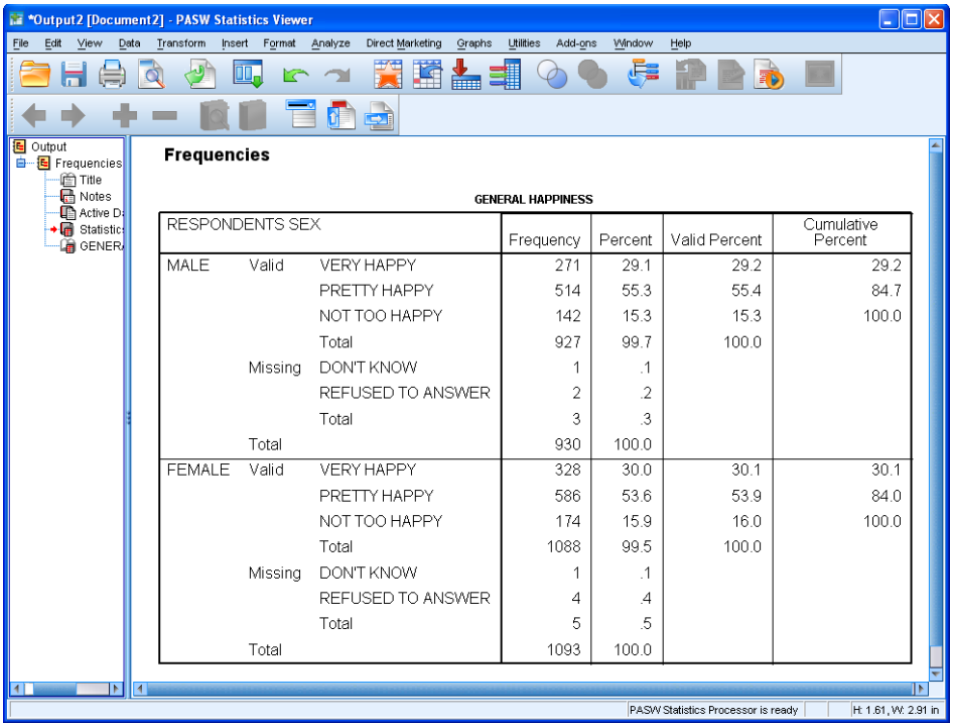
Split File processing is accomplished with these steps:

1) Define the split groups. The split groups can be based on more than one variable. If you specify multiple variables, each combination of values will make up a group. For example, specifying *sex* with 2 categories and *region* with 8 categories will create 16 groups.
2) Select the way output should be organized, output in one table in order to compare groups or output for each group in a separate table.
3) Run the analyses.
4) Optionally, undo split file processing.

10.10 *Split File Output*

The **Split File** procedure allows for two options of organizing the output. In the first option, output for all groups is presented together for comparison purposes, where here we have split on sex.

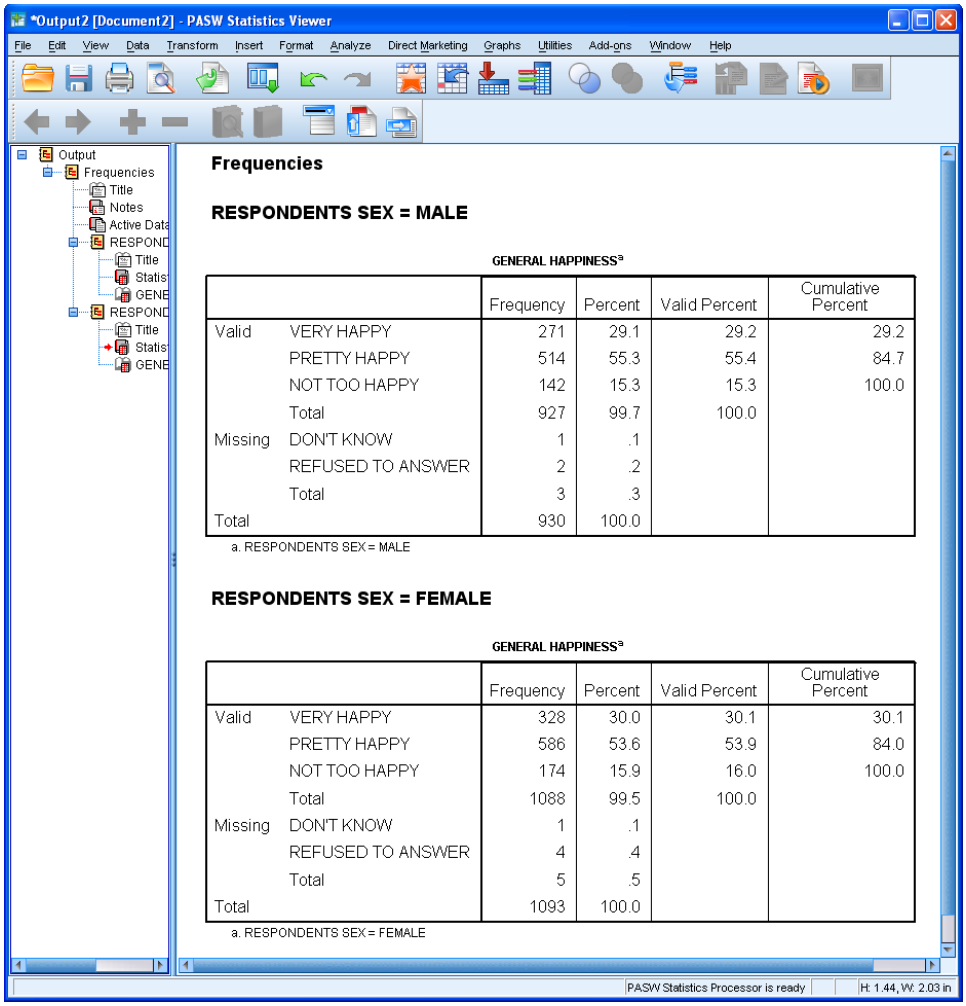
Figure 10.7 Split File, Compare Groups Output



RESPONDENTS SEX			Frequency	Percent	Valid Percent	Cumulative Percent
MALE	Valid	VERY HAPPY	271	29.1	29.2	29.2
		PRETTY HAPPY	514	55.3	55.4	84.7
		NOT TOO HAPPY	142	15.3	15.3	100.0
		Total	927	99.7	100.0	
	Missing	DON'T KNOW	1	.1		
		REFUSED TO ANSWER	2	.2		
		Total	3	.3		
	Total		930	100.0		
FEMALE	Valid	VERY HAPPY	328	30.0	30.1	30.1
		PRETTY HAPPY	586	53.6	53.9	84.0
		NOT TOO HAPPY	174	15.9	16.0	100.0
		Total	1088	99.5	100.0	
	Missing	DON'T KNOW	1	.1		
		REFUSED TO ANSWER	4	.4		
		Total	5	.5		
	Total		1093	100.0		

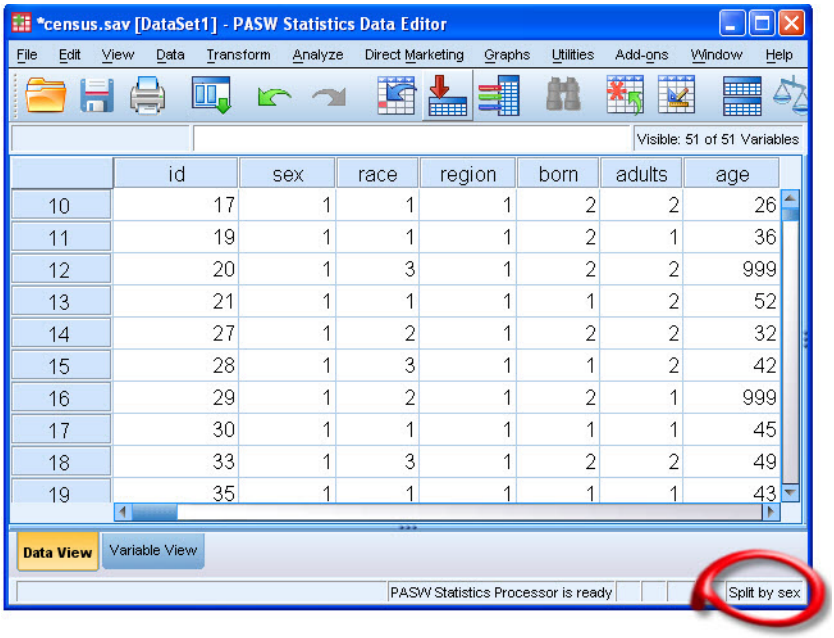
The second option displays the results from each procedure separately for each split-file group.

Figure 10.8 Split File, Organize Output by Groups Output



The status bar of the Data Editor will reflect that split file processing is on, and indicates the split file variable.

Figure 10.9 Split File Message in Status Bar

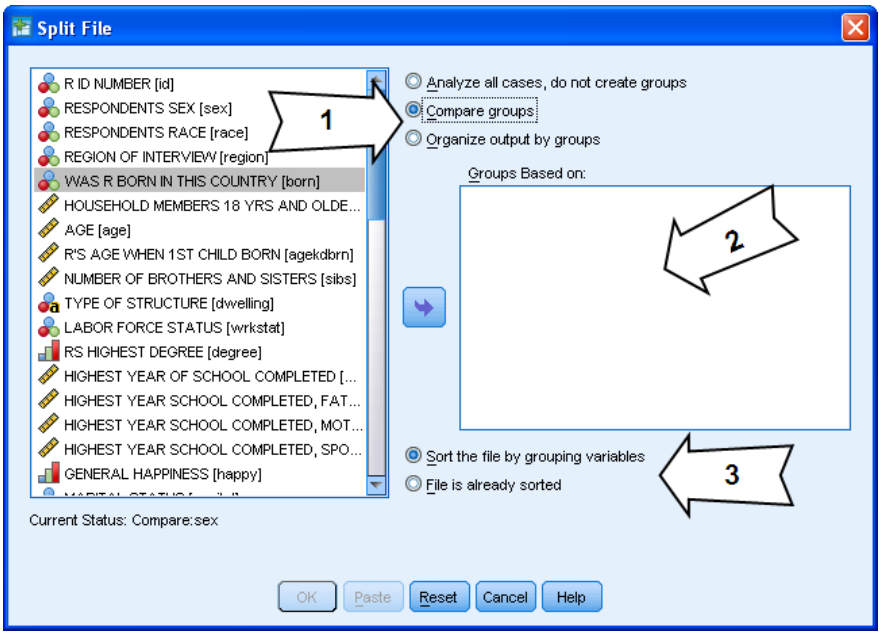


10.11 *Procedure: Split File*

Split File is invoked from the **Split File** choice on the **Data** menu.

1) Specify the way output must be organized.
2) Select variable (s) to split on.
3) Optionally, specify if the data must be sorted. Sorting by the split file variables is the default.

Figure 10.10 Split File Dialog



Tip

If the data are already sorted, select the option *File is already sorted* for efficiency.

Note

Split file processing is turned off by selecting *Analyze all cases, do not create groups* in the Split File dialog box. This will not resort the data to its original order, so use Sort Cases to sort the cases back to the order before Split File (or open the original data file).

10.12 *Example: Split File*

We will work with the *census_small.sav* data file in this lesson. We want to create a report of respondent's education, spouse's education, father's education, and mother's education by marital status, but we want these statistics for each marital group separately. We use the **Descriptives** procedure with **Split File**.

Note

If we want the statistics in one table we can use the **Split File**, option Compare Groups, but the **Means** procedure does exactly that, so that procedure is preferred above **Split File** in this situation.

Detailed Steps for Split File

With the **Split File** dialog open:

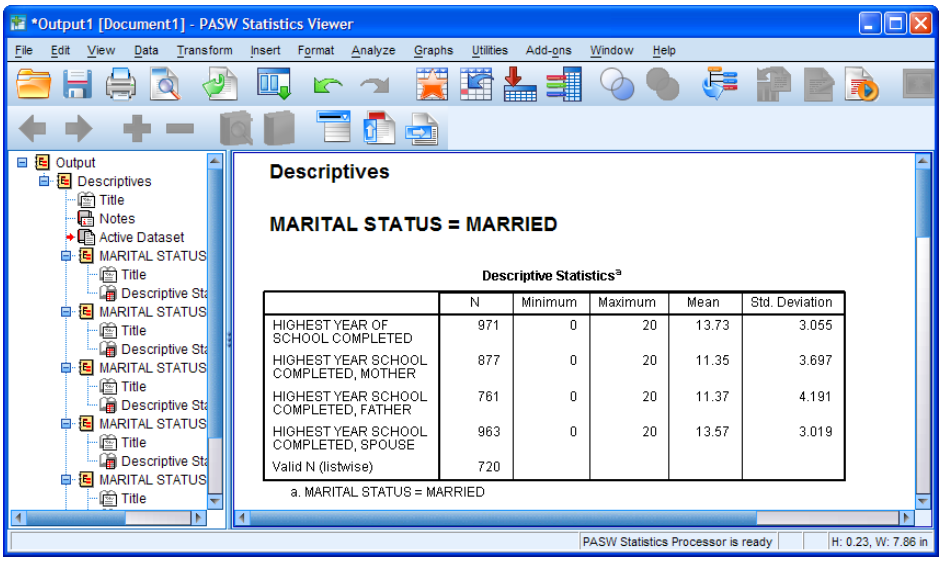
1) Select option **Organize output by groups**.
2) Select ***MARITAL STATUS [marital]*** as the variable to base the groups on.
3) Select **Analyze…Descriptive Statistics…Descriptives**
4) Place the variables ***HIGHEST YEAR OF SCHOOL COMPLETED [educ]*, HIGHEST YEAR SCHOOL COMPLETED, FATHER [paeduc]**, ***HIGHEST YEAR SCHOOL COMPLETED, MOTHER [maeduc]***, and ***HIGHEST YEAR SCHOOL COMPLETED, SPOUSE [speduc]*** in the Variable(s) box

Results from Split File

The data will be sorted on the split variable (here *marital*) and the status bar in the Data Editor will display a message that split file processing is on.

The **Descriptives** procedure on the education variables will give a separate table for each of the categories of marital status.

Figure 10.11. Split File with Descriptives, Table for Married Respondents



Notice that split file identification information appears both in the title and as a footnote for each table. Each subgroup descriptive table is also identified in the Outline portion of the Viewer.

Apply Your Knowledge

1. True or False? Split File creates a filter variable for the purposes of splitting the file?

10.13 *Lesson Summary*

Lesson Objectives Review

Students who have completed this lesson should be able to:

- Select cases in a data file using various methods

And, they should also be able to:

- Describe and use the features of the Select Cases dialog
- Describe and use the features of the Split File dialog

10.14 *Learning Activity*

The overall goal of this learning activity is to select cases and use split file processing. You will use the PASW Statistics data file *employee data.sav*.

Supporting Materials

The data file *employee data.sav* contains information on employees of a major bank. Included is data on beginning and current salary position, time working, and demographic information.

1. Select (with filtering) employees with a current salary less than 10,000 and create a new dataset named *salary_less_than_10000* for the selected cases.
   How many cases are in this dataset?

2. Return to the original dataset and select those cases whose job category (*jobcat*) is manager and current salary is greater than 20,000. Filter out the unselected cases.

3. Check your selection by running a **Means** table on the filtered data requesting summary statistics (including minimum and maximum) for *salary* by *jobcat*. How many cases were selected? Are they all managers?

4. Turn off the filter. Now split the file *gender*, organizing the output by groups rather than comparing groups.

5. Run a crosstab table of *jobcat* by *minority*, requesting percentages based on *minority*.

6. Turn off split file processing.

7. *For those with more time:* Rename the filter variable created in step #2. Then redo the case filtering in that step by using the filter variable instead of constructing a logical condition. Double-check to make certain the filtering worked.

Lesson 11: Creating and Editing Charts

11.1 *Objectives*

After completing this lesson students will be able to:

- Present results with charts

To support the achievement of this primary objective, students will also be able to:

- Use the Chart Builder to create various types of graphs
- Format and edit the graph in the Chart Editor

11.2 *Introduction*

A wide variety of chart types are available in PASW Statistics from the Graphs main menu. In this lesson we will show how to use the Chart Builder to build some basic charts and use basic features of the Chart Editor to edit the charts.

Business Context

The topics in this lesson can help in answering questions such as:

- How can I show a distribution of a variable in a graph?
- How can I show a relationship between variables in a graph?
- What formatting features are available to optimize the lay-out of the graph?

Supporting Materials

The file *census_small.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

11.3 *The Chart Builder*

In PASW Statistics there are several procedures to create graphs:

- The Chart Builder. The Chart Builder allows you to build charts from predefined galleries. You first select the chart type and then you select the variables to fit in this chart type.
- The Graphboard Template Chooser. The Graphboard Template Chooser works the other way around than the Chart Builder. In the Graphboard Template Chooser, you first select your variables and then given this selection, Graphboard suggests appropriate charts for the type of data.
  The Graphboard Template Chooser provides integration with PASW Viz Designer, a PASW Statistics add-on module
- Legacy Dialogs. The Legacy Dialogs provide access to the graph interfaces referred to as Standard Charts in previous versions of the software.
- Additional statistical charts are available from the Analyze menu, e.g. Quality Control Charts, ROC Curves, and Forecasting Charts.

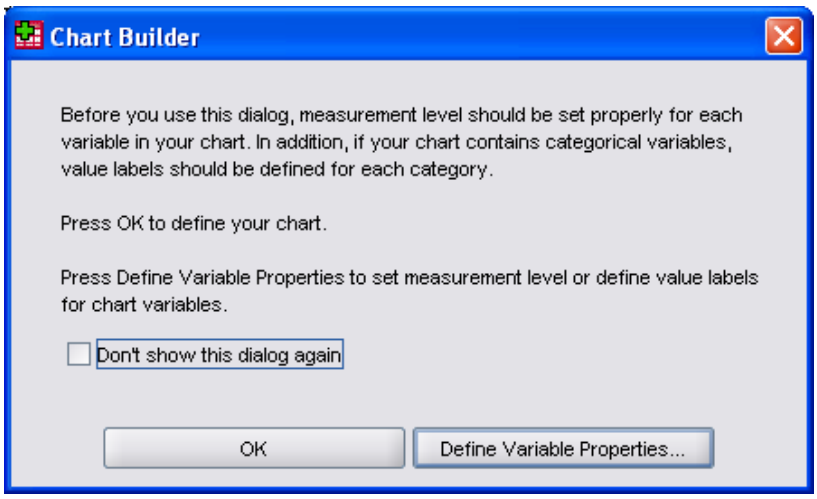
In this lesson, we will show how to use the Chart Builder to create graphs, as it is the easiest way to create a graph and provides many features for creating simple to customized charts.

11.4 *Procedure: Use the Chart Builder*

The Chart Builder is invoked from the Graphs menu. The first time it is opened after the software is installed, an information box is opened which reminds us that the measurement level for the variables to be used in the chart should be properly set because it will affect the way variable can be used in the Chart Builder. It also reminds us that we will probably want to define value labels for categorical variables to be used in the chart.

If you need to make changes to the measurement level or add value labels, use the Define Variable Properties dialog, or work directly in the Data Editor, Variable View.

Figure 11.1 Message for Measurement Level and Value Labels in Chart Builder



Further Info

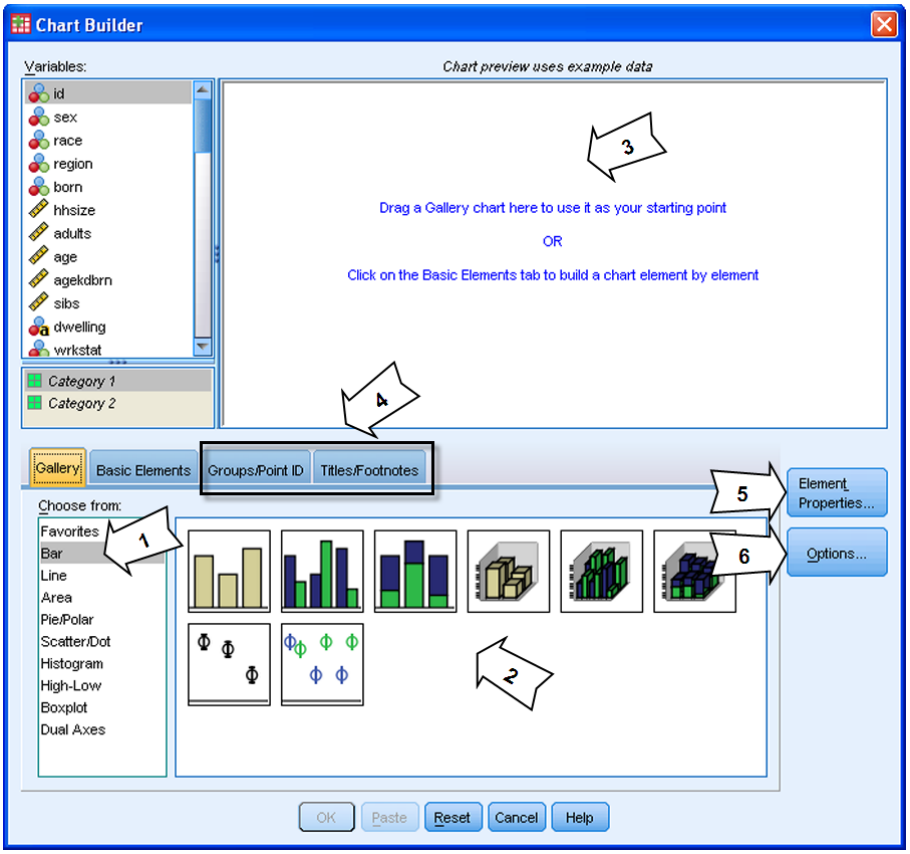
You can also temporarily change measurement level directly in the Chart Builder variable list by right-clicking on a variable and making a selection from the Context menu.

The Chart Builder dialog provides an interactive interface that allows you to preview how the chart looks as you build it in the preview canvas in the upper right corner of the dialog box.

To build a chart:

1) Select from the Gallery of common chart types.
2) Select a specific chart of the selected chart type.
3) Drag and drop this chart to the Preview Canvas.
4) Optionally, define paneled and other special elements under Groups/Point ID and add Titles/Footnotes.
5) Optionally, click the Element Properties button to customize statistics, include/exclude categories, and so forth. Do this after variables have been added to the chart (see below).
6) Optionally, click the Options button for missing value handling.

Figure 11.2 The Chart Builder Dialog

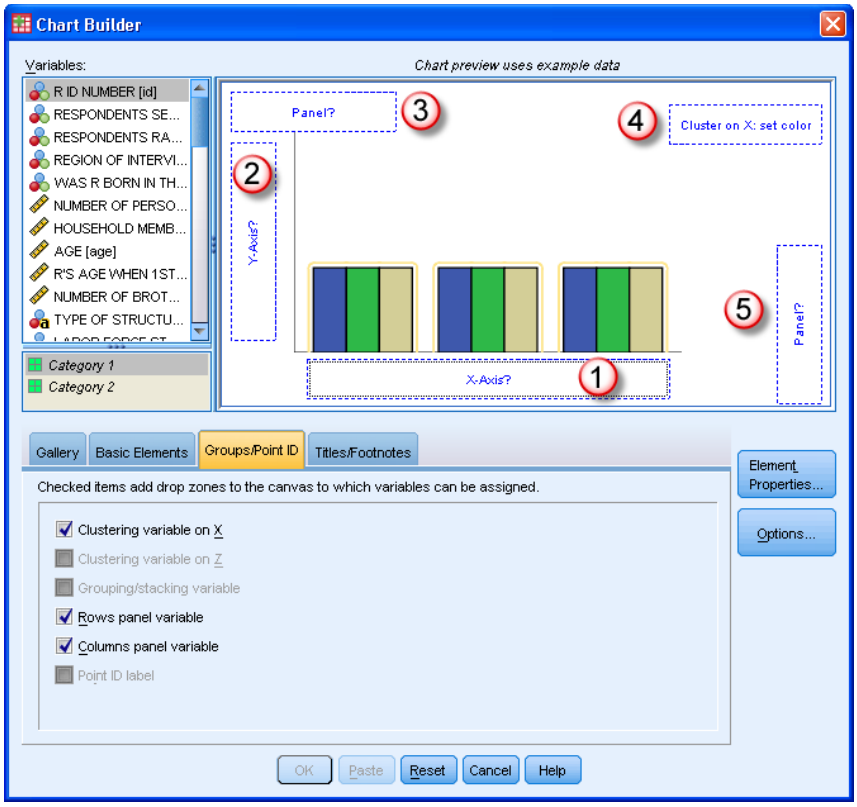


If you drag a chart to the Preview Canvas, the Preview Canvas shows drop zones where the variables should be dragged.

Further Info

Instead of creating a graph from the gallery, you can build the chart element by element from the Basic Elements tab, but this is only recommended for more advanced users.

Figure 11.3 Drop Zones



If you have dragged a particular chart to the Preview Canvas, different drop zones will be displayed. Drop zones are the areas on the canvas to which you drag and drop a variable or a statistic. The basic drop zones include:

- X-Axis drop zone (1) and the Y-Axis drop zone (2).
- Certain charts include more drop zones, like a paneling drop zone (3, 5), and a cluster variable drop zone (4).
- The text in the drop zone describes the purpose of the drop zone.

If a drop zone is included but no element for that drop zone has been specified, the OK button will be grayed out.

Note

The Preview Canvas displays the chart using example data, so the actual chart can differ significantly from that of the preview.

11.5 *Demonstration: Creating a Pie Chart*

We will work with the data file *census_small.sav* data file in this lesson. As a first example, we will create a pie chart for *HAPPINESS OF MARRIAGE*.

Detailed Steps for Creating a Pie Chart

With the Chart Builder dialog open:

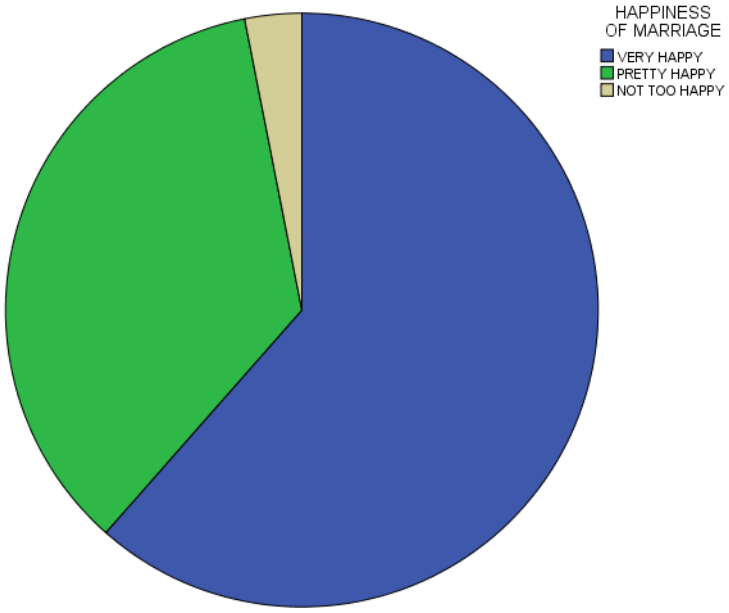
1) From the Gallery Choose from: list, select a **Pie/Polar** graph type.
2) Select the **Pie Chart** and drag it to the Preview Canvas.
3) Drag and drop ***HAPPINESS OF MARRIAGE [hapmar]*** to the X-Axis drop zone (labeled Slice by?)

Notice that Count automatically will be selected as the statistic to report.

Results for Creating a Pie Chart

The figure below depicts the pie chart. It shows that more than half of the people are very happy in their marriage.

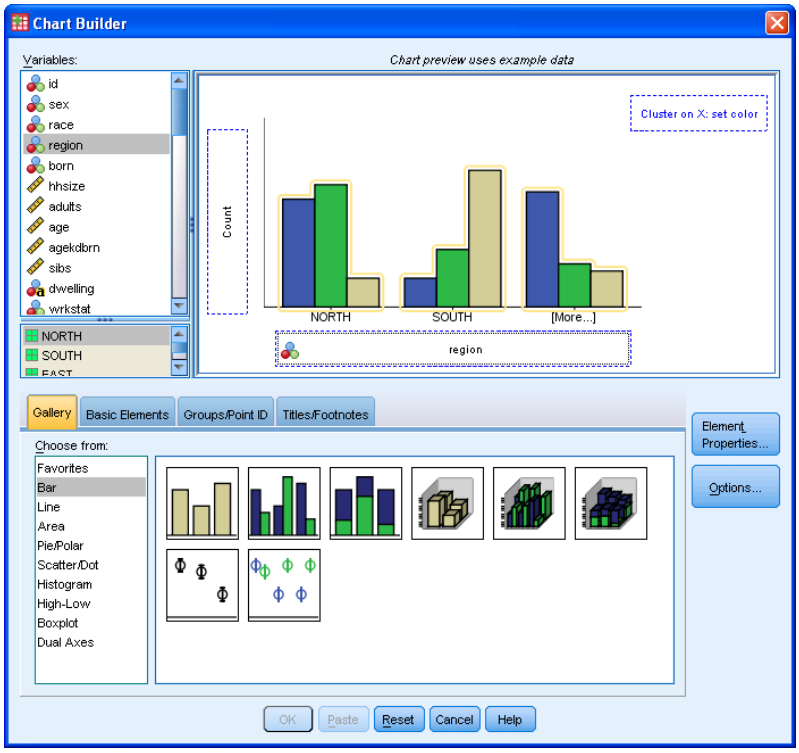
Figure 11.4 Pie Chart of Happiness of Marriage



Apply Your Knowledge

1. True or false? The graph shown in the Preview Canvas is based on data in the active dataset?

2. See the Chart Builder dialog below. What is the reason that the OK button is grayed out?
    a. There are too many categories for *region*
    b. The Cluster: drop zone needs to be specified
    c. Only percentages can be produced in a clustered bar chart, not counts
    d. Count on the Y-Axis should be replaced with a variable



11.6 *Bar Charts*

Bar charts are used to present a variety of data:

- Summarizing categorical variables. This is the common use of a bar chart.
- Summarizing a number of scale variables. For example, showing the mean satisfaction with staff in one bar, next to the mean satisfaction with price in another bar, next to the mean satisfaction for delivery time in another bar
- Summarizing combinations of categorical variables. For example, use a bar chart that shows the percent of cases in each category of highest degree earned by gender.
- Summarizing a scale variable for a grouping (categorical) variable. For example, you can use a bar chart to show the mean salary for men in a bar next to the mean salary for women in the other bar.

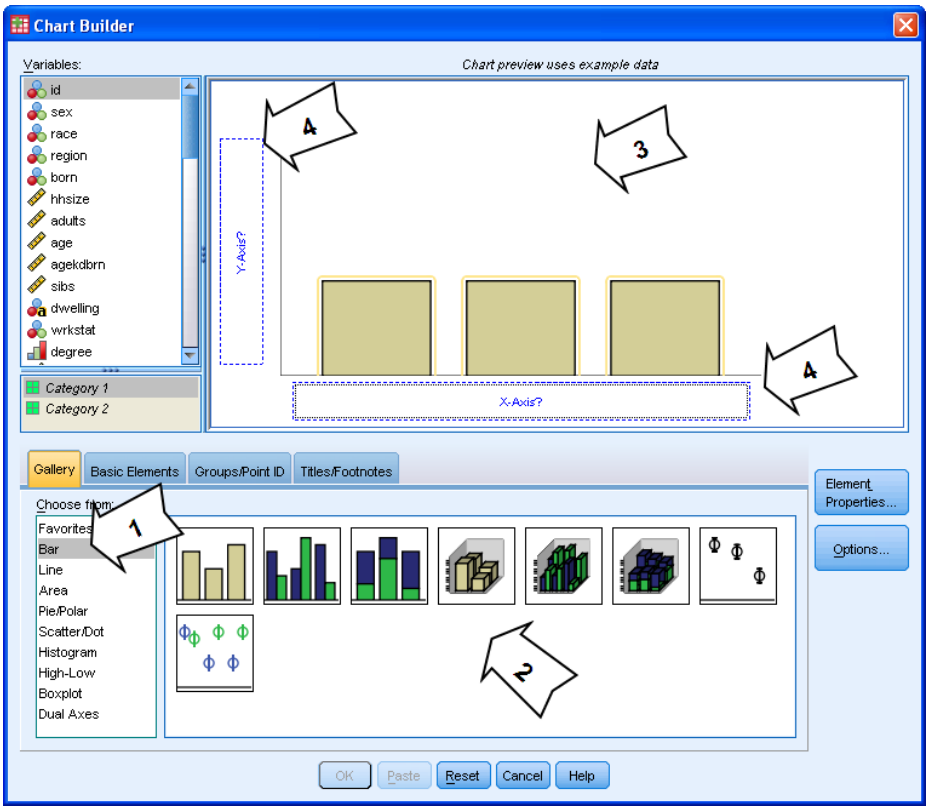
11.7 *Procedure: Creating a Bar Chart*

To create a bar chart, with the Chart Builder dialog open:

1) From the Gallery, select Bar.
2) Select the type of bar chart.
3) Drag the selected chart to the Preview Canvas.
4) Specify the element(s) for the drop zones.

Here, “element(s)” refer to variable(s) or statistic(s) and depend on the chart requested.

Figure 11.5 Chart Builder Dialog Create a Simple Bar Chart



11.8 *Demonstration: A Simple Bar Chart for a Categorical Variable*

In this example we will produce a bar chart for *HAPPINESS OF MARRIAGE*.

Detailed Steps for a Simple Bar Chart

With the Chart Builder dialog open:

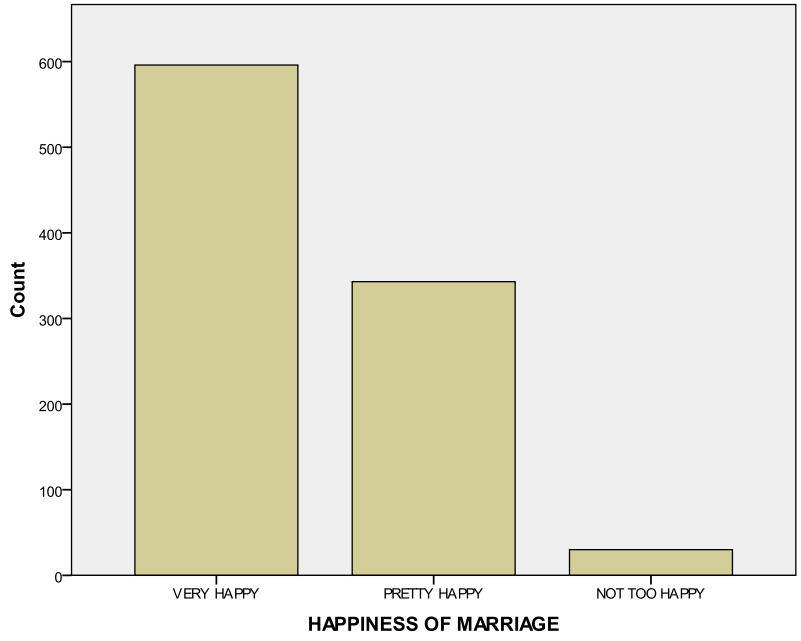
1) Select **Bar** from the Gallery Choose from list.
2) Select the **Simple Bar** (the first icon) and drag it to the Preview Canvas.
3) Drag and drop ***HAPPINESS OF MARRIAGE [hapmar]*** to the X-Axis drop zone.

Notice that automatically the Count statistic is selected as the statistic to report. The reason is that *HAPPINESS OF MARRIAGE* is a categorical variable (that is why measurement level should be set correctly for variables).

Results for a Simple Bar Chart

The figure below depicts the bar chart. The bars show the frequency of each of the categories. As such, it is the graphical equivalent of a table produced by the **Frequencies** procedure.

Figure 11.6 Simple Bar Chart for Happiness of Marriage



11.9 *Demonstration: A Simple Bar Chart for Scale Variables*

There are several variables in the data that record the number of years of education for, respectively, the respondent, his/her spouse, the father, and the mother. In this next example, we want to compare the means for these education variables.

Detailed Steps for a Simple Bar Chart for Scale Variables

With the Chart Builder dialog open:

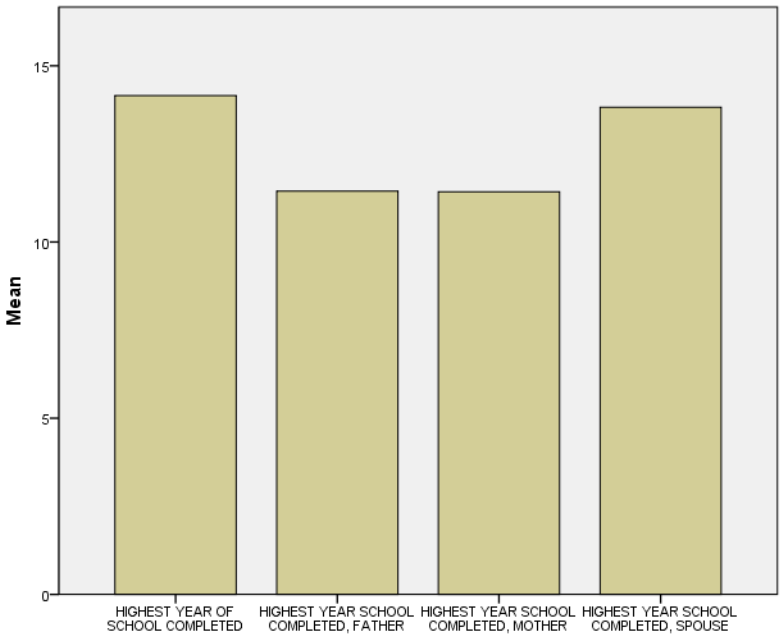
1) Select **Bar** from the Gallery Choose from: list.
2) Select the **Simple Bar** (the first icon) and drag it to the Preview Canvas.
3) Select ***HIGHEST YEAR OF SCHOOL COMPLETED [educ]*, HIGHEST YEAR SCHOOL COMPLETED, FATHER [paeduc]**, ***HIGHEST YEAR SCHOOL COMPLETED, MOTHER [maeduc]***, and ***HIGHEST YEAR SCHOOL COMPLETED, SPOUSE [speduc]*** (use the CTRL key to do a multiple select) and drag and drop these variables to the Y-Axis drop zone.

Notice that automatically the Mean statistic is selected as the statistic to report. This is because these variables have a measurement level of scale.

Results for a Simple Bar Chart for Scale Variables

The figure below depicts the bar chart. The bars represent the mean on each of the variables. The respondent him/herself has the highest mean education. As such, it is the graphical display of a table produced by the **Descriptives** procedure.

Figure 11.7 Simple Bar Chart for Means of Education Variables



11.10 *Demonstration: Bar Chart for the Mean of a Scale Variable by a Categorical Variable*

We often want to display a measure of central tendency, such as the mean, for subgroups of cases. In this demonstration, we'll display the mean number of hours per day watching television (*HOURS PER DAY WATCHING TV*) by categories of *LABOR FORCE STATUS*.

Detailed Steps for a Bar Chart for the Mean of a Scale Variable by a Categorical Variable

With the Chart Builder dialog open:

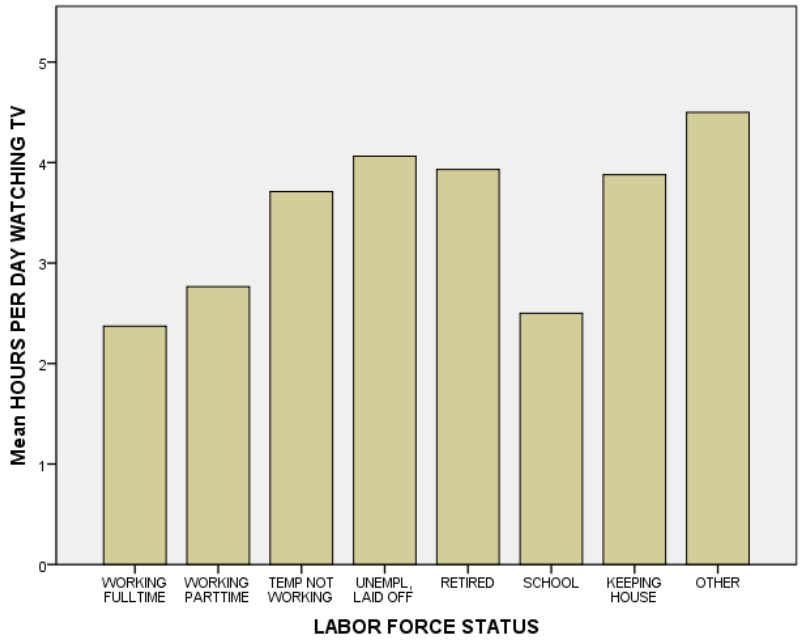
1) Select **Bar** from the Gallery Choose from: list.
2) Select the **Simple Bar** and drag it to the Preview Canvas.
3) Drag ***LABOR FORCE STATUS [wrkstat]*** to the X-Axis drop zone.
4) Drag ***HOURS PER DAY WATCHING TV [tvhours]*** to the Y-Axis drop zone

Notice that automatically the Mean is selected as the statistic to report.

Results for a Simple Bar Chart for the Mean of a Scale Variable for Groups

The chart shows mean watching television each day by working status. This chart is the graphical equivalent of a table produced by the **Means** procedure.

Figure 11.8 Bar Chart for the Mean of Hours Watching Television by Labor Force Status



11.11 *Demonstration: Clustered Bar Chart with a Scale Variable*

In this next example, let's extend the previous chart by adding gender as a clustering variable. This will allow us to display differences in TV viewing by labor force status for men and women separately.

Detailed Steps for a Clustered Bar Chart with a Scale Variable

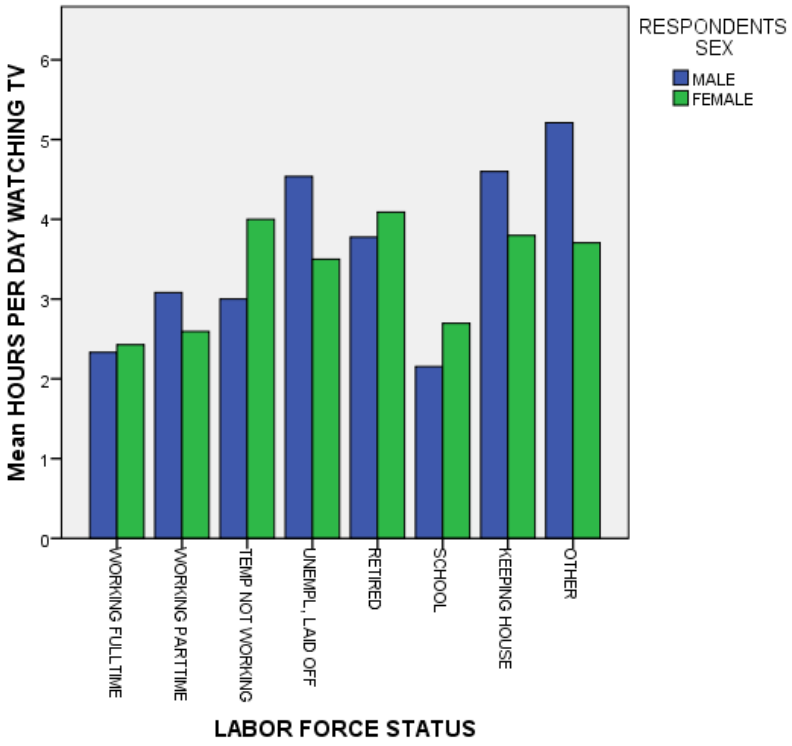
With the Chart Builder dialog open:

1) Select **Bar** from the Gallery Choose from: list.
2) Select the Clustered Bar icon (the second in the top row) and drag it to the Preview Canvas.
3) Drag ***LABOR FORCE STATUS [wrkstat]*** to the X-Axis drop zone.
4) Drag ***RESPONDENTS SEX [sex]*** to Cluster on X: set color drop zone.
5) Drag ***HOURS PER DAY WATCHING TV [tvhours]*** to Y-Axis drop zone

Results for a Clustered Bar Chart with a Scale Variable

The clustered bar chart shows two bars for each category of labor force status, one for males and one for females. We can see that that there are differences by gender, with males higher in some groups, but females higher in others.

Figure 11.9 Clustered Bar Chart with a Scale Variable



11.12 *Histograms*

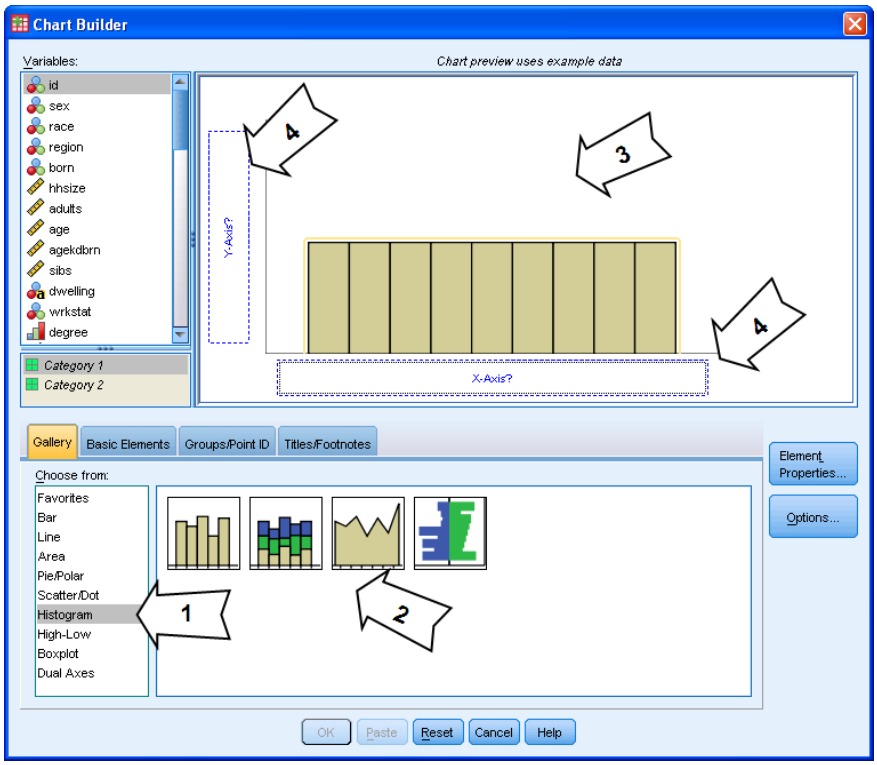
A histogram displays the distribution of a scale variable, as a bar chart does for a categorical variable. A histogram can help you determine whether the distribution is symmetrical or skewed, or if there are outliers. If you want to assess whether a variable is normally distributed, a normal distribution can be superimposed.

11.13 *Procedure: Create a Histogram*

To create a histogram, with the Chart Builder dialog open:

1) From the Gallery, select Histogram.
2) Select the type of histogram.
3) Drag the selected histogram to the Preview Canvas.
4) Specify the element(s) for the different drop zones.

Figure 11.10 Create a Histogram



11.14 *Demonstration: Population Pyramid*

Among the various types of histograms, the Chart Builder provides a **population pyramid** to display the distribution of a scale variable separately for a categorical variable. Normally, population pyramids are used for dichotomous categorical variables, although variables with three or more categories can be used. For example, you could display the distribution of age when first child was born for men and women.

Detailed Steps for a Population Pyramid

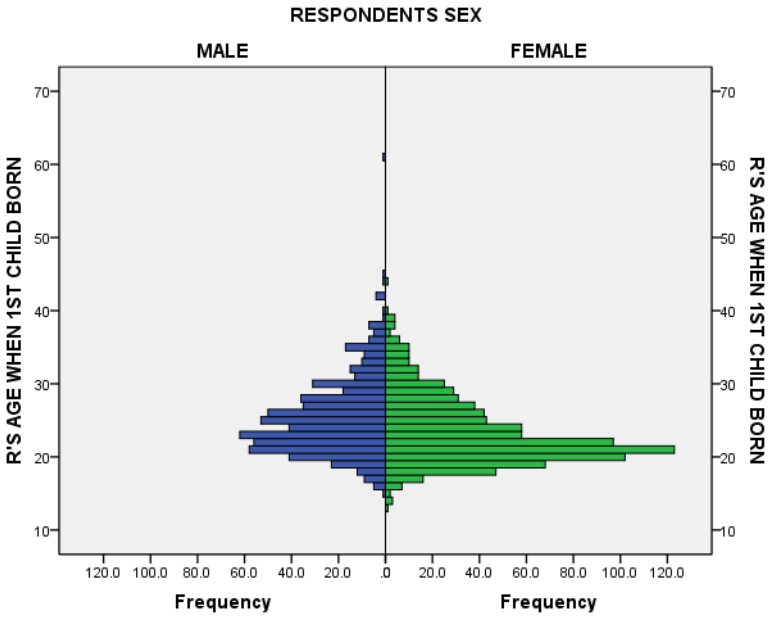
With the Chart Builder dialog open:

1) Select **Histogram** from the Gallery Choose from: list.
2) Select the **Population Pyramid** icon (the last one in the row) and drag it to the Preview Canvas.
3) Drag ***R'S AGE WHEN 1ST CHILD BORN [agekdbrn]*** to the Y Axis drop zone.
4) Drag ***RESPONDENTS SEX [sex]*** to the Split-variable? drop zone.

Results for a Population Pyramid

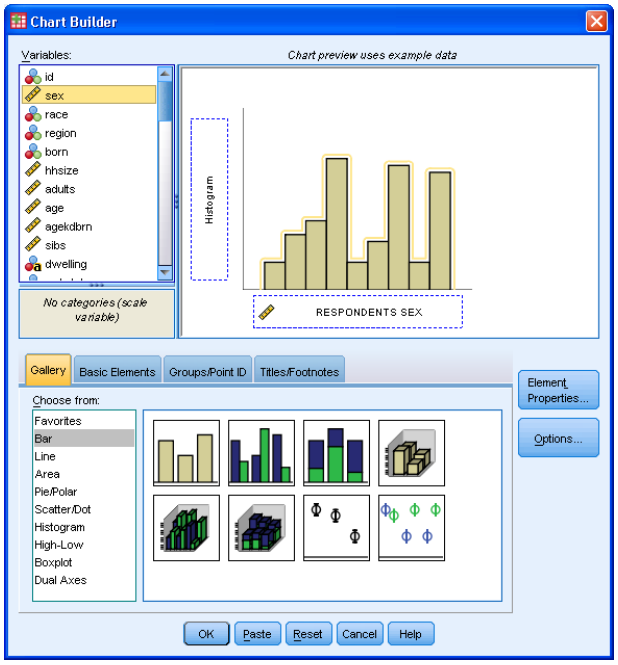
The population pyramid displays the distribution or histogram for males on the left side and the distribution for females on the right side. The scale is the frequency count or number of people in each bin. Comparing the distributions, we see that many females were in their late teens or early 20's when their first child was born. On the other hand, more males were in their mid to late 20's when their first child was born.

Figure 11.11 Population Pyramid for Age When First Child Born by Sex



Apply Your Knowledge

1. See the figure below. Why is a histogram suggested in the Preview Canvas in the Chart Builder for the variable *sex*?
    a. A bar chart will always give a histogram
    a. The variable *sex* is defined as nominal in measurement level
    b. The variable *sex* is defined as ordinal in measurement level
    c. The variable *sex* is defined as scale in measurement level



11.15 *Editing Charts*

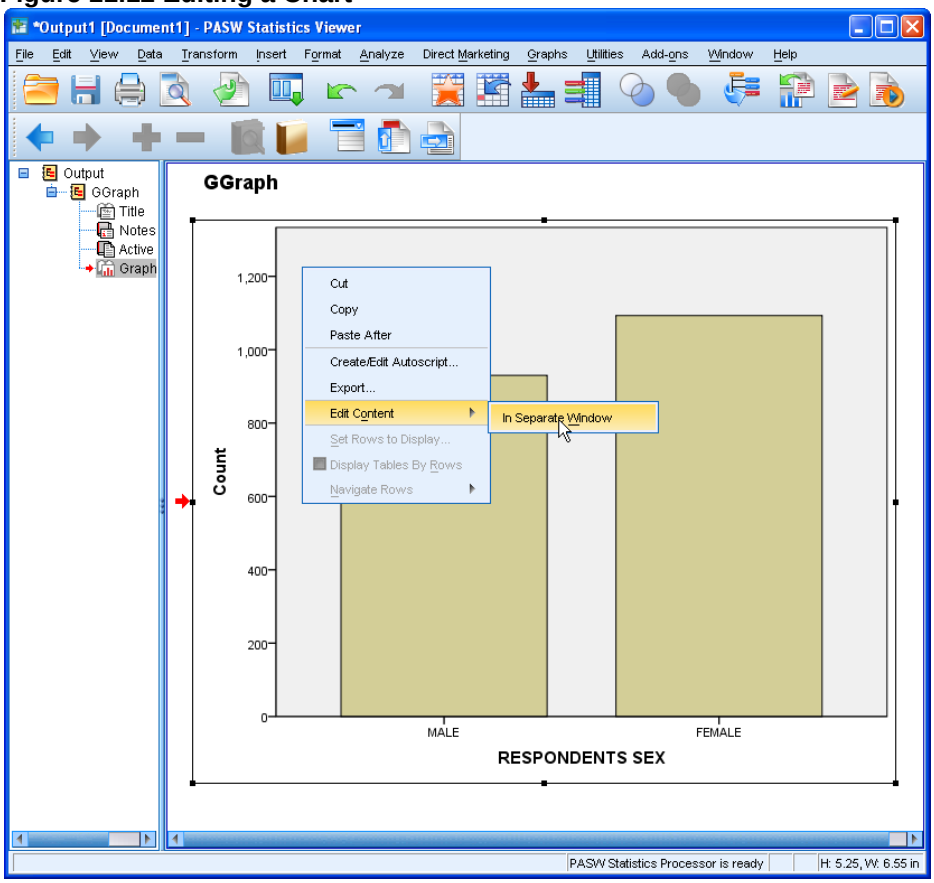
You can edit many attributes of charts built with the Chart Builder using the Chart Editor. The chart opens in a separate window, the Chart Editor window. Most attributes of charts are edited from the Properties dialog box.

11.16 *Procedure: Editing Charts*

To edit a chart:

1) Right-click the chart in the Output Viewer and select Edit Content…In Separate Window (alternatively, double-click the chart).

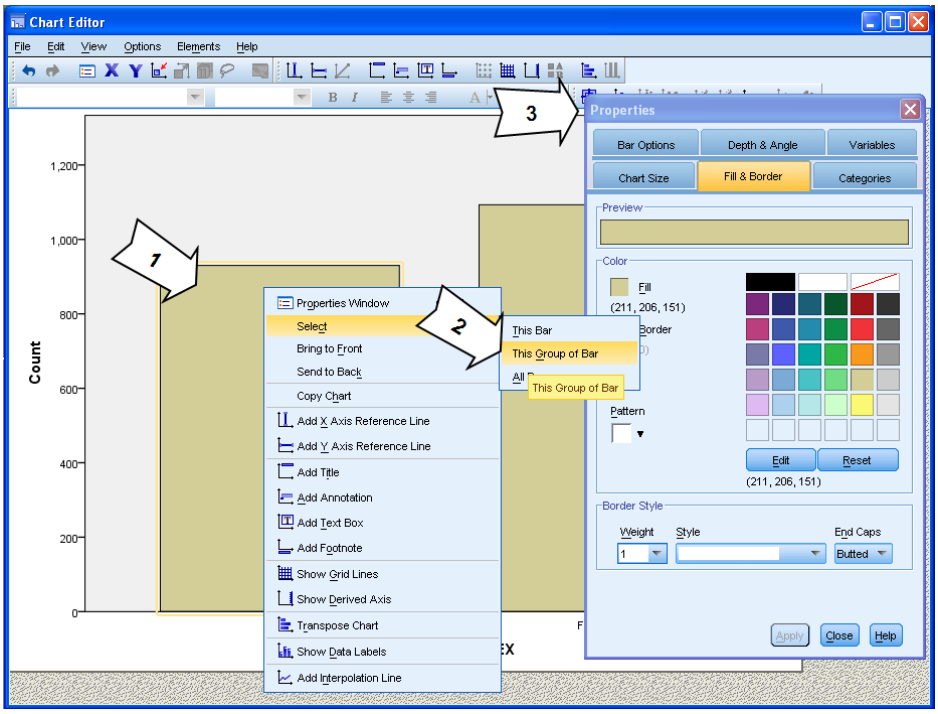
Figure 11.12 Editing a Chart



The chart will be opened in the Chart Editor. To edit elements in the chart:

1) Select the element you want to change. (You can use the Control key to select multiple groups of elements.)
2) Select the specific one that you want to edit (this is the general "drill-down" rule of selecting elements in most charts).
3) Use the tabs in the Properties dialog box to change the properties for the selected element.

Figure 11.13 Chart Editor Dialog



11.17 *Demonstration: Editing a Bar Chart*

For a bar chart, changing the fill color of the bars is often necessary. A nice effect as well is to display the data labels so it is easy to read the value represented by a chart element, such as a bar. Additionally, often bar charts are transposed so that the bars are horizontal. We'll make all these changes to the bar chart for the mean hours watching television by labor force status.

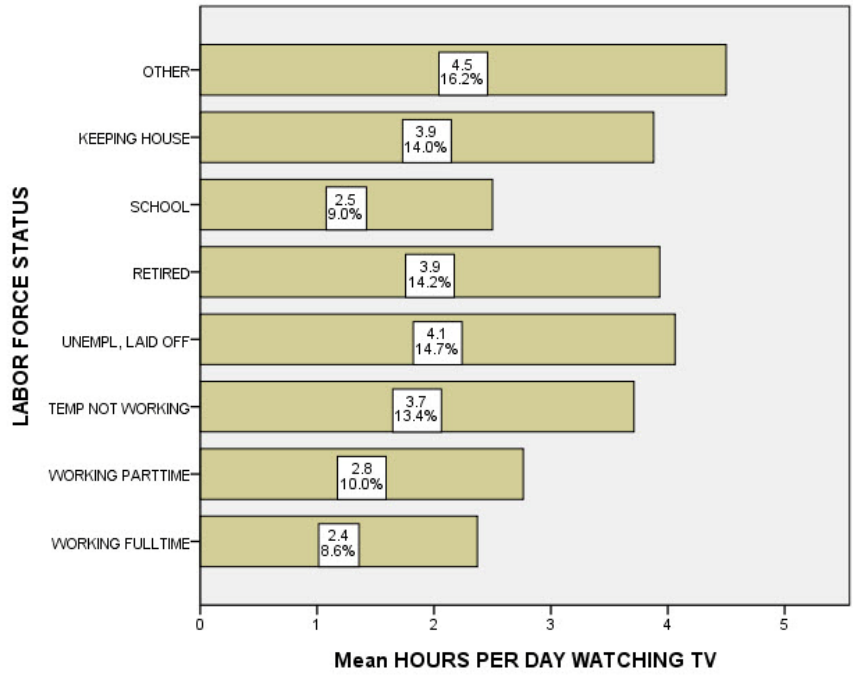
Detailed Steps for Editing a Bar Chart

1) Open the bar chart in the Chart Editor
2) Right-click a bar and choose **Select…All Bar** from the context menu.
3) Select **Elements…Show Data Labels**.
4) In the Properties dialog, select the **Data Value Labels** tab and move **Percent** from Not Displayed to **Displayed**
5) Select **Apply**.
6) In the Properties dialog, select the **Number Format** tab
7) Change the number of Decimal Places to **1**
8) Select **Apply**
9) Select **Options…Transpose Chart**.

Results for Editing a Bar Chart

The figure below depicts the edited bar chart.

Figure 11.14 Edited Bar Chart



Note

On the File menu of the Chart Editor there is an option to save characteristics of the edited chart in a chart template which can then be applied to other charts. In this way, you can carry over to other charts the characteristics that you have edited in the chart to create your own custom look.
We discuss this feature in detail in the *Data Management and Manipulation* course.

11.18 *Lesson Summary*

Lesson Objectives Review

Students who have completed this lesson should be able to:

- Present results with charts

And, they should also be able to:

- Use the Chart Builder to create various types of graphs
- Format and edit the graph in the Chart Editor

11.19 *Learning Activity*

The overall goal of this learning activity is to create and edit graphs. You will use the PASW Statistics data file *employee data.sav*.

Supporting Materials

The data file *employee data.sav* contains information on employees of a major bank. Included is data on beginning and current salary position, time working, and demographic information.

1. Open *employee data.sav*.
2. Produce a simple bar chart that displays the distribution of job categories (*Employment Category*).
3. Create a clustered bar chart showing *Employment Category* by *Gender*. How would you describe the distributions?
4. Change various aspects of the chart in the Chart Editor. For example, change the color of selected bars, add data value labels, and change the scale values for the Y-axis.
5. Create a population pyramid chart that displays the distribution of *Current Salary* by gender. Add a distribution curve to the chart (Hint: Element Properties dialog or Elements menu in Chart Editor). Which distribution is closer to normal?
6. Edit other chart attributes as you desire.

Lesson 12: Output in the Viewer

12.1 *Objectives*

After completing this lesson students will be able to:

- Use the Output Viewer and its editing features

To support the achievement of this primary objective, students will also be able to:

- Navigate through the Viewer
- Customize a pivot table
- Create and apply a template for a pivot table
- Export output to other applications

12.2 *Introduction*

As we have seen, the results of running a statistical or graphical procedure are displayed in the Output Viewer window. The output produced can be statistical tables, charts or graphs, or text, depending on the choices you make when you run the procedures. In the Output Viewer window, you can easily browse the output by navigating to individual tables or charts. You can also manage and manipulate the output to create a document that contains the required results, arranged and formatted appropriately.

In this lesson we demonstrate features of the Output Viewer that can be used to manage the results, the Pivot Tables editor, and exporting results into other applications.

Business Context

When you have run procedures in PASW Statistics, you will often want to share your findings with others. This can involve:

- Formating tables as cleanly as possible for clarity of presentation and ease of viewing
- Exporting output to a file format that can be read by any user, e.g., Portable Document Format (pdf).
- Doing post-processing on pivot tables in other applications, e.g., Microsoft Excel.
- Creating a template for pivot tables to match a preexisting format.

Supporting Materials	• The file *census_small.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics. • The PASW Statistics output file *Sample_Viewer.spv* which contains output to be viewed and used to illustrate various features available in the Viewer window.

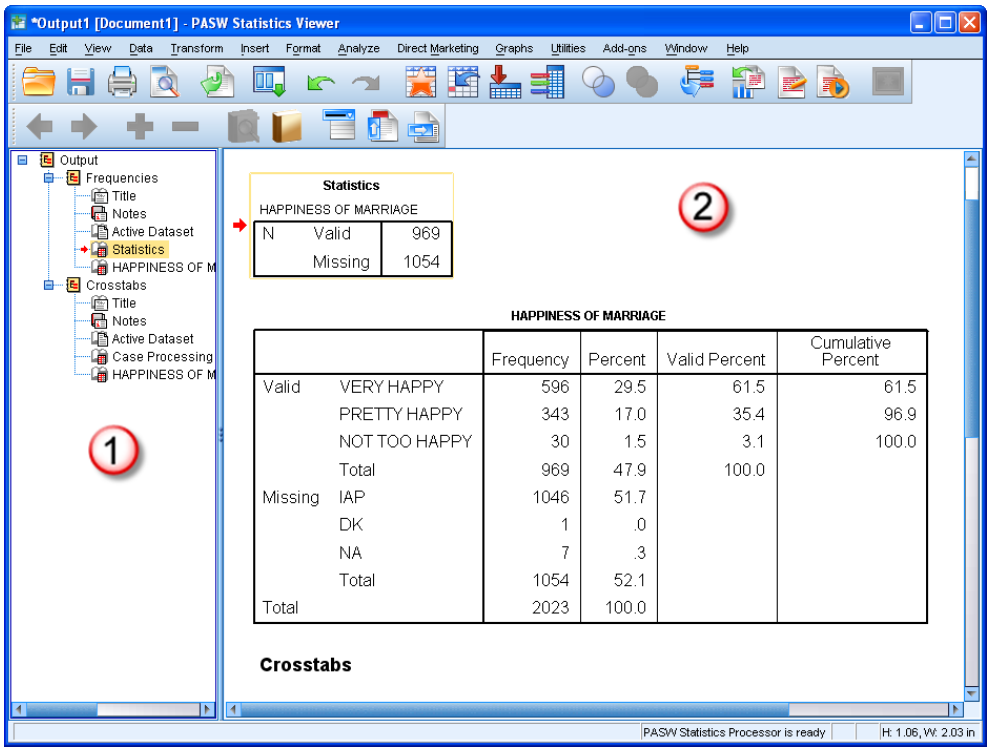
12.3 *The Viewer*

PASW Statistics presents and stores results in a separate window, the Viewer. Tables and graphs will be collected in this document in the order they are created, so understanding the features available in this window is key to customizing output and exporting output to other applications.

The Viewer window is divided into two panes, as in the figure below.

- The outline pane (1) contains an outline of all of the information stored in the Viewer.
- The content pane (2) contains the statistical tables, charts, and text output. The divider between the two can be moved (left click and drag) to allocate more or less space to each pane.

Figure 12.1 The Viewer Window



In the outline pane, an open book icon appears in front of the items visible in the content pane and a closed book icon in front of those that are not shown in the content pane. You can hide an object in the content pane by double-clicking on the open book icon in the outline pane or selecting the item and using the hide tool on the toolbar. Or, you can use the View menu to Hide or Show items. Hiding items without actually deleting them allows the user to more easily concentrate on the results of interest while retaining all of the results.

Navigating in the Viewer

To navigate in the Viewer it is important to point out that the items in the Viewer are a series of objects organized in hierarchical associated blocks, each defined by a procedure. Each block is identified by a name and the marker. For example, Frequencies defines all of the results associated with running the **Frequencies** procedure. Within a block of output there are several objects, such as tables, charts, titles, notes, and so forth.

Tip

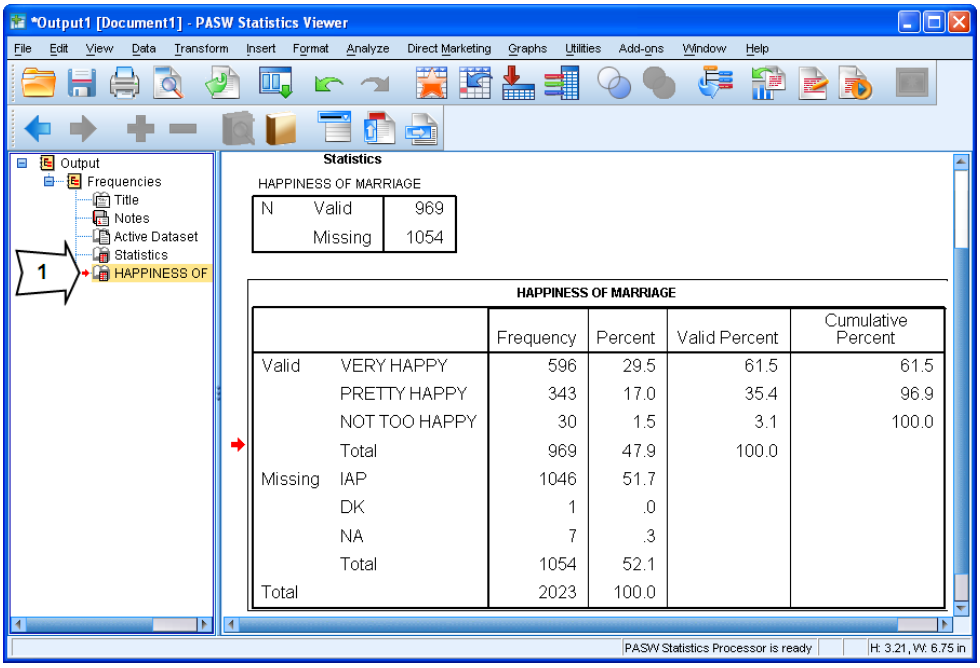
You can change several default settings for the Output Viewer in the Viewer tab of the **Edit…Options** menu. Ten output item types can be shown or hidden. Typically, you will want at least the Title, Pivot Table, Chart, Trees, Text, and Model Viewer Output to be shown initially. We also recommend showing the Warnings item so that you are aware of any unusual conditions.

12.4 *Procedure: Navigating in the Viewer*

You can scroll through the output objects in the content pane. But, navigation is easier from the outline pane.

1) Click on an item in the outline pane to select it in the content pane.

Figure 12.2 Selecting an Item (Object) in the Viewer Outline Pane



A red arrow highlights the selected item in both the outline pane and the content pane.

Tip

Select multiple items with Ctrl and/or Shift keys. Select all items with Ctrl+A.

Standard Windows functionality allows you to move, copy or delete an item:

- To move a selected item (or selected items): drag the selected object to the desired new location. In this way, you can organize output generated by different statistical and graphical procedures together for your presentation or report.

- To copy a selected item (or selected items): select **Edit…Copy** or from the popup menu (right-click), select **Edit…Paste After**.
- To delete a selected item (or selected items), simply press the Delete key or select **Edit…Delete**.

You can hide all of the output from a particular statistical procedure or all of the output in the Viewer. You can click the box with the minus sign (-) to the left of the procedure to hide all of the results from that procedure. The outline collapses, visually indicating that these results are hidden.

Tip

You can insert a heading, title and text in the Viewer. To insert a heading or other item, select the item before the desired location in the outline pane of the Viewer, and select the desired option (heading, title, and so forth) from the Insert menu. These objects can be edited directly by double-clicking the item, and then formatting can be done as is normally performed in a word processor.

12.5 *Demonstration: Editing Objects in the Viewer*

We will work with the *census_small.sav* data file in this example and the Viewer file *Sample_Viewer.spv*. The output file contains a frequency table of *HAPPINESS OF MARRIAGE [hapmar]* and a crosstabulation of *hapmar* by *sex*. To provide examples of editing, we will delete all the Title objects, move the crosstab table, and hide the Crosstabs command block.

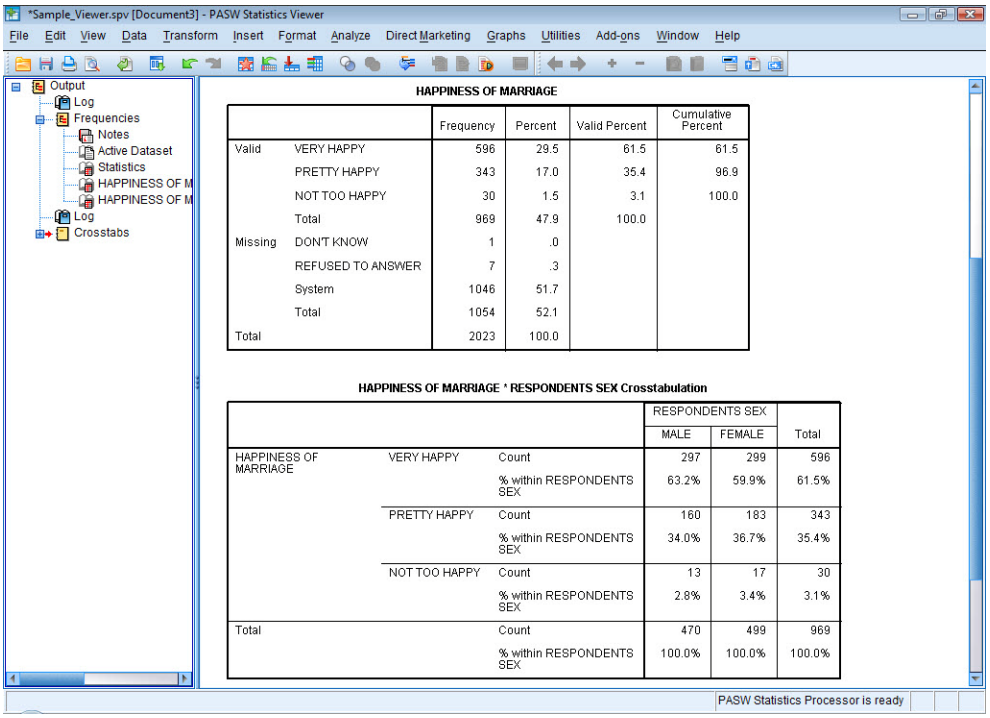
Detailed Steps for Editing Objects in the Viewer

1) Select **Edit…Select…All Titles**.
2) Select **Edit…Delete**
3) Select the *HAPPINESS OF MARRIAGE * RESPONDENTS SEX* crosstabulation in the outline pane
4) Drag and drop the selected table to the frequency table
5) Click the minus sign ⊟ before the block of **Crosstabs** output to hide the **Crosstabs** output

Results from Editing Objects in the Viewer

The figure below depicts the results after manipulating the items.

Figure 12.3 Viewer Output Edited



Apply Your Knowledge

1. True or false? If you want to create an empty Viewer document, you have to exit PASW Statistics and start over?

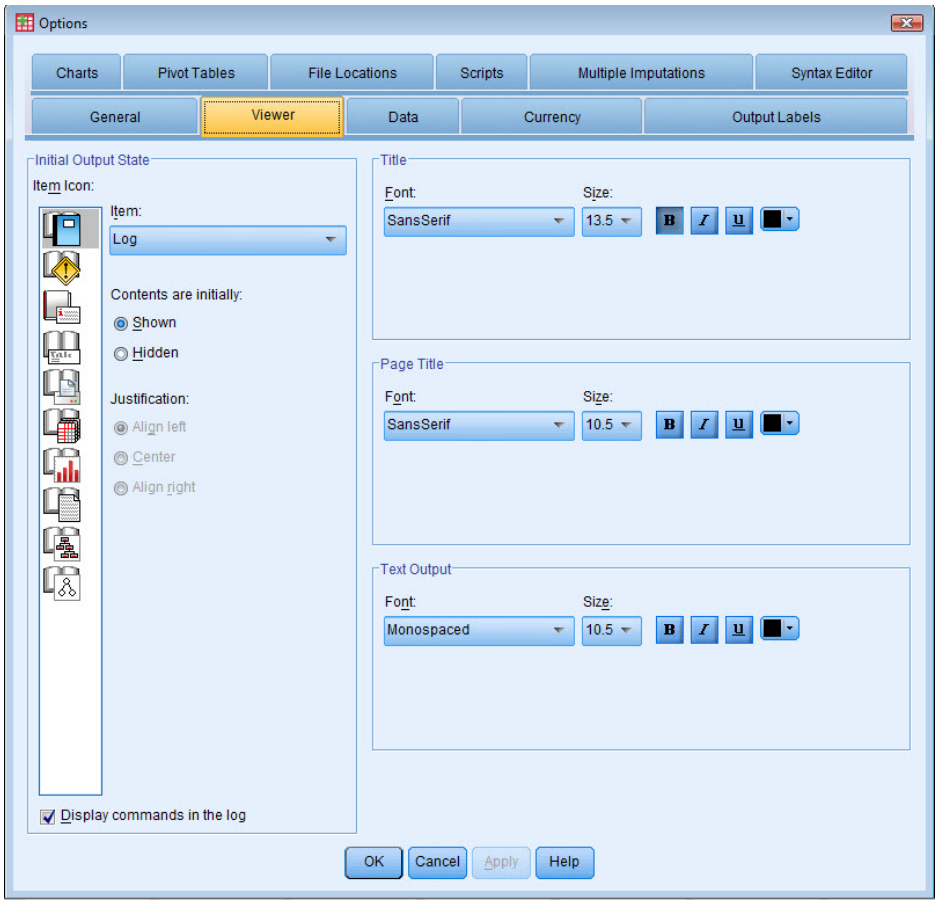
12.6 *Setting Default Viewer Options*

You can change several default settings for the Output Viewer in the **Viewer** tab of the **Edit…Options** menu.

Options in the *Initial Output State* group control which items of each procedure's output are either initially displayed or hidden. For example, Log items contain the syntax generated by a PASW Statistics command (see Lesson 12 Syntax Basics) and are displayed by default (note the *Display commands in the log* check box).

You can change the settings for any of these items by selecting the item from the pull-down list or selecting its icon, and then selecting *Shown* or *Hidden*.

Figure 12.4 Viewer Tab of Options Dialog



12.7 *The Pivot Table Editor*

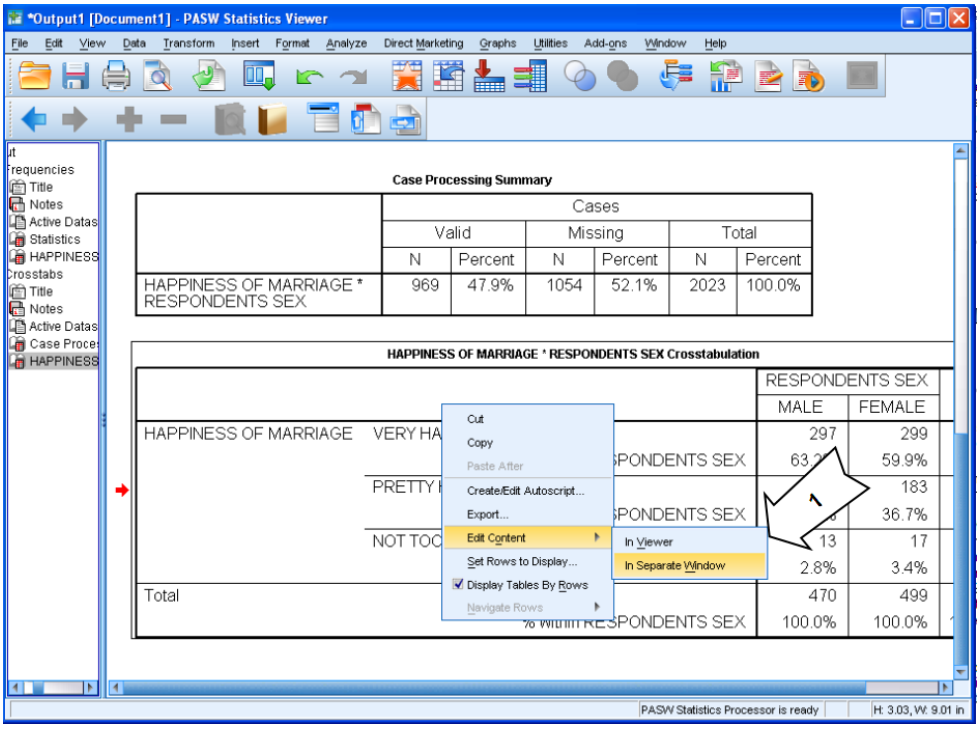
Procedures in PASW Statistics display the numeric output in pivot tables which are composed of objects and elements that can be rearranged and edited using the Pivot Table Editor. You can pivot the table (change rows to columns and columns to rows, or move individual elements, such as a variable or set of statistics), change cell properties such as fonts, colors and alignment of data in individual cells, as well as change table properties associated with areas of the table such as border lines and style. Changing the layout of the table does not affect the results. Instead, it's a way to display your information in a different or more desirable manner.

12.8 *Procedure: Editing a Pivot Table*

You can edit a pivot table in the Viewer window or open it to a separate window called the Pivot Table Editor. Often times a larger table is better edited in its own window. To edit a pivot table in a separate window:

1) Right-click the pivot table and select **Edit Content…In Separate Window**.

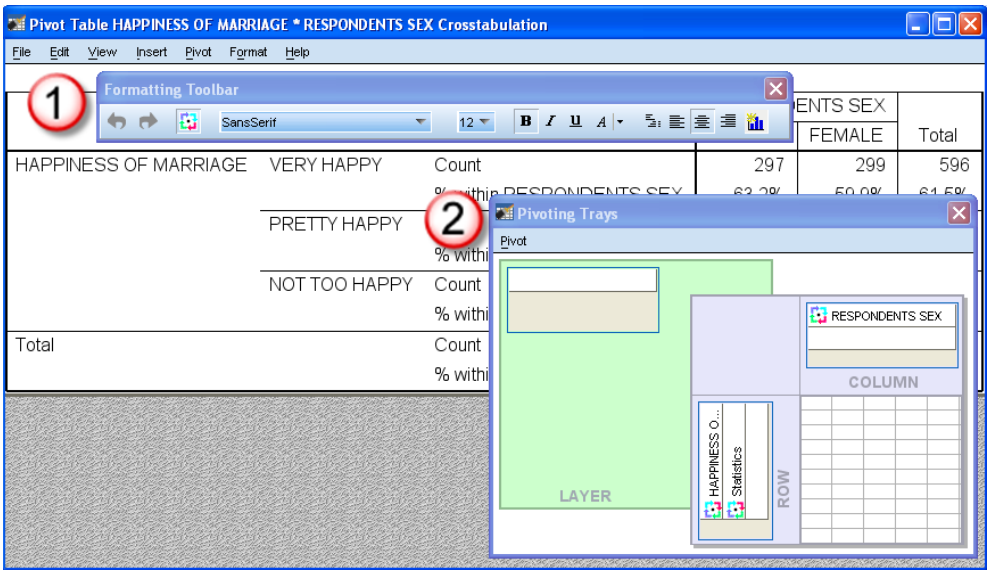
Figure 12.5 Request to Edit Pivot Table in a Separate Window



By default, the Pivot Table Editor will show the:

- Formatting Toolbar (1)
- Pivoting Trays window (2).

Figure 12.6 The Pivot Table Editor



Note

If the Formatting Toolbar does not display in the Pivot Table Editor, select **View…Toolbar**. If the Pivoting Trays window does not display, select **Pivot…Pivoting Trays**.

The Formatting Toolbar allows the user to customize the text properties (select the element you want to edit and then use the toolbar to format it).

The Pivoting Trays window visualizes each data element in the table. Here, in the figure above, the Column area at the top of the pivoting tray contains RESPONDENTS SEX which is displayed in the column dimension in the table. The Row area at the left contains the elements displayed across the rows of the table, in this case HAPPINESS OF MARRIAGE and the Statistics (here: Count and % within RESPONDENTS SEX). Note that the Pivoting Trays window shows the statistics nested within the categories of HAPPINESS OF MARRIAGE, which is in agreement with the table layout. So, the order of items in a dimension reflects the order of the elements in the table.
The Layer area in the Pivoting Trays window is shown on the upper left of the pivoting tray (empty here). Layers can be useful for large tables with nested categories of information. By creating layers, you simplify the look of the table, making it easier to read. Layers work best when the table has at least three variables.

Another way to customize the layout of the table is via the Cell properties dialog (Format…Cell Properties). Fonts, background colors, formatting of numbers and alignment of cells can be changed here.

Figure 12.7. Cell Properties Dialog to Format Cells

Further formatting is provided by the Table Properties dialog (Format…Table Properties menu) that works on areas of the table all at once.

12.9 *Demonstration: Editing a Pivot Table*

In this demonstration we will customize the crosstabs of sex and happiness of marriage in our modified Viewer window.

Detailed Steps for Editing a Table

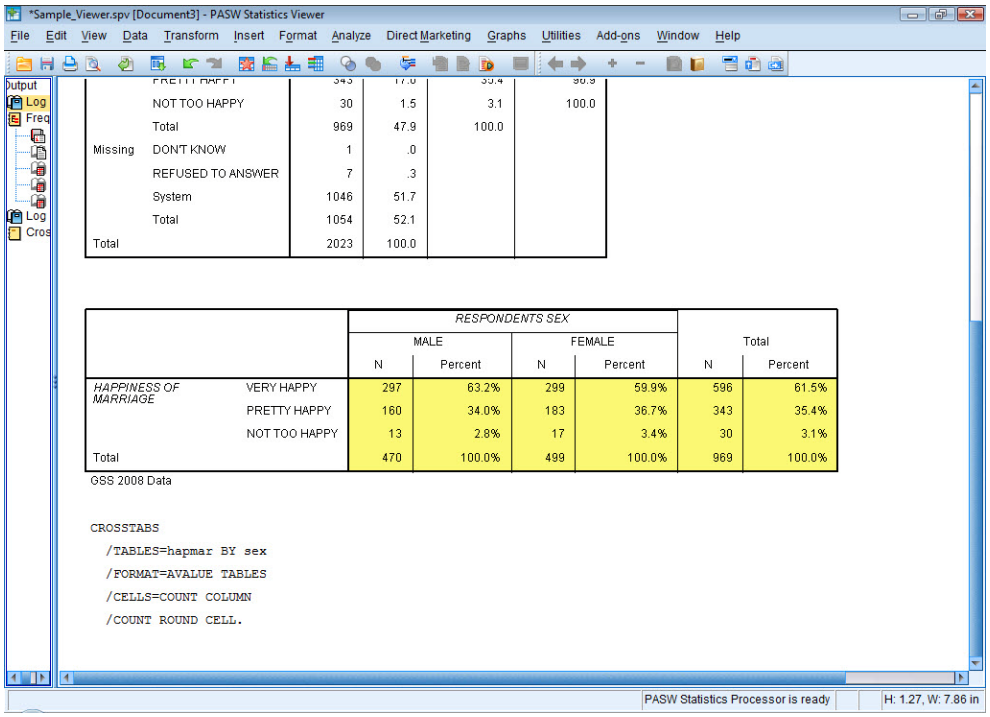
With the table opened in the Pivot Table Editor:

1) Drag the **Statistics** from the Row area to the bottom (below RESDONDENT'S SEX) in the **Column** area
2) Use the mouse to reduce the **column width** of the columns with the label **% within RESPONDENT'S SEX**
3) Select the label **HAPPINESS OF MARRIAGE** and use the Formatting Toolbar to put the text in **italics**.
4) Select the label **RESPONDENT'S SEX** and use the Formatting Toolbar to put the text in **italics**.
5) Select the statistic **Count** and change its label to **N**.
6) Select the statistic **% within RESPONDENTS SEX** and change its label to **Percent**.
7) Select all data cells (select and drag with the mouse) and then select **Format…Cell Properties**
8) Select the **Font and Background** tab and then the **Background** box
9) Set the background color to **yellow**.
10) Select **Insert…Caption** and enter the text **GSS 2008 Data** for the caption
11) Close the Pivot Table Editor, so the table is updated in the Viewer.

Results from Editing a Table

Editing the table as described above results in the table shown below.

Figure 12.8 Edited Crosstab Table



Apply Your Knowledge

1. True or false? Any pivot table can be edited?

2. True or false? Editing tables can change the data in the table?

12.10 *TableLooks®*

The look and feel of your tables is a critical part of providing clear, concise, and meaningful results. For example, if your table is difficult to read, the information contained within that table may not be easily understood.

PASW Statistics TableLooks allow you to customize the look and feel of the tables you create. TableLooks are templates that can be applied to any pivot table.

Although you can change the format of a table after it has been created, it will be more efficient to change the default TableLook so that you do not have to change the format every time you create a table. This is especially useful if you have several tables that should have the same style.

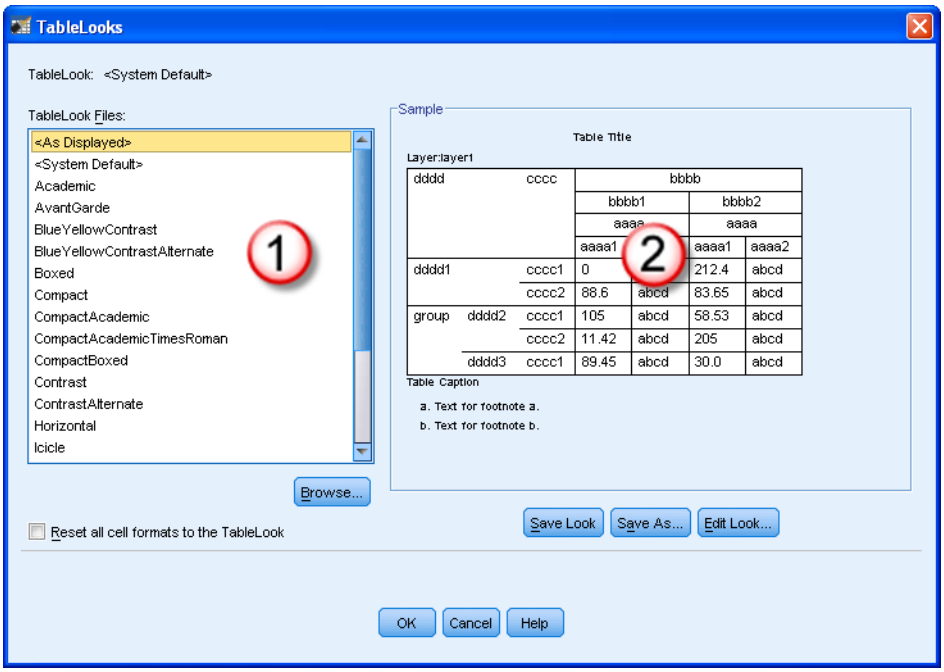
By default, PASW Statistics uses a system default TableLook. However, you can either choose from one of the several pre-defined TableLooks or create your own customized TableLook.

12.11 *Procedure: Create and Apply a Tablelook*

In the Pivot Table Editor, select **Format…TableLooks** to apply, create or save a TableLook. In the TableLooks dialog:

- There is a list of available TableLooks (1) that are available in a particular folder
- A preview of the selected TableLook is shown on the right (2).

Figure 12.9 TableLooks Dialog



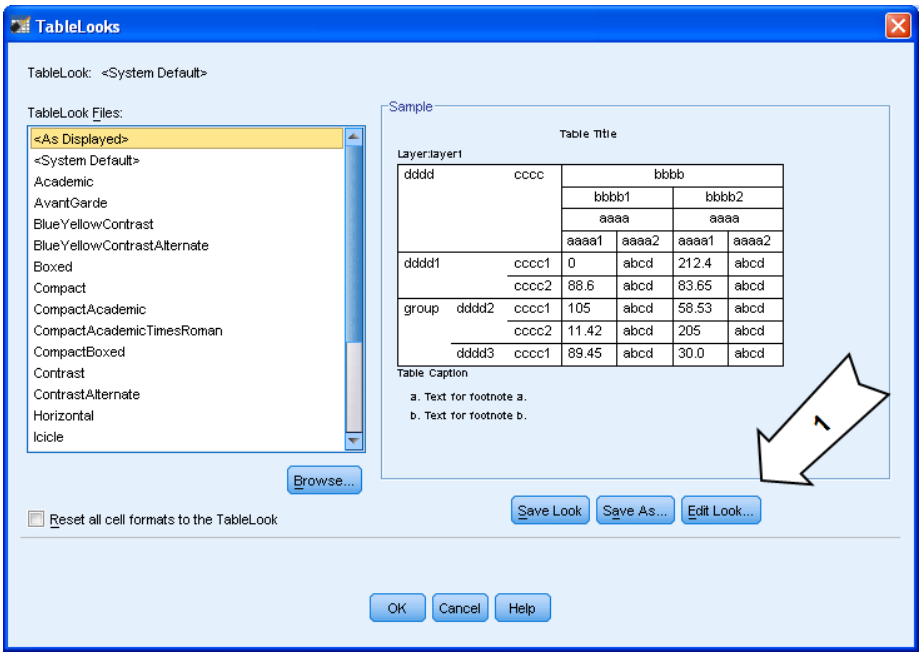
You can use a TableLook as it is, or you can edit a TableLook to better suit your needs. To apply an existing TableLook to the pivot table:

1) Select the TableLook from the list and select OK.

To edit a TableLook:

1) Select the Edit Look button

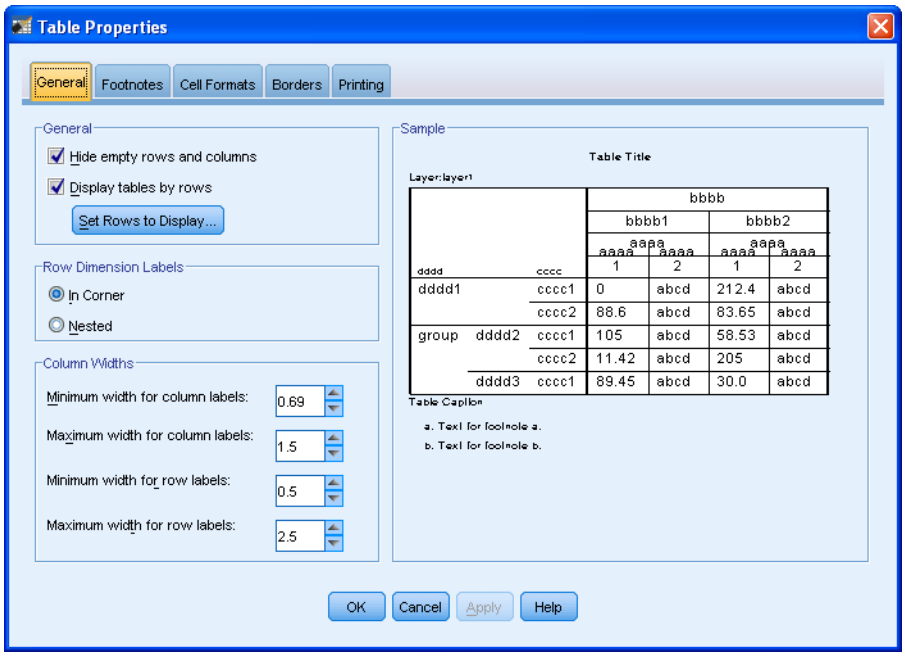
Figure 12.10 Request to Edit a TableLook



When you edit a TabeLook, the Table Properties dialog opens. Five tabs are available to customize the look of the table. Key features in each tab include:

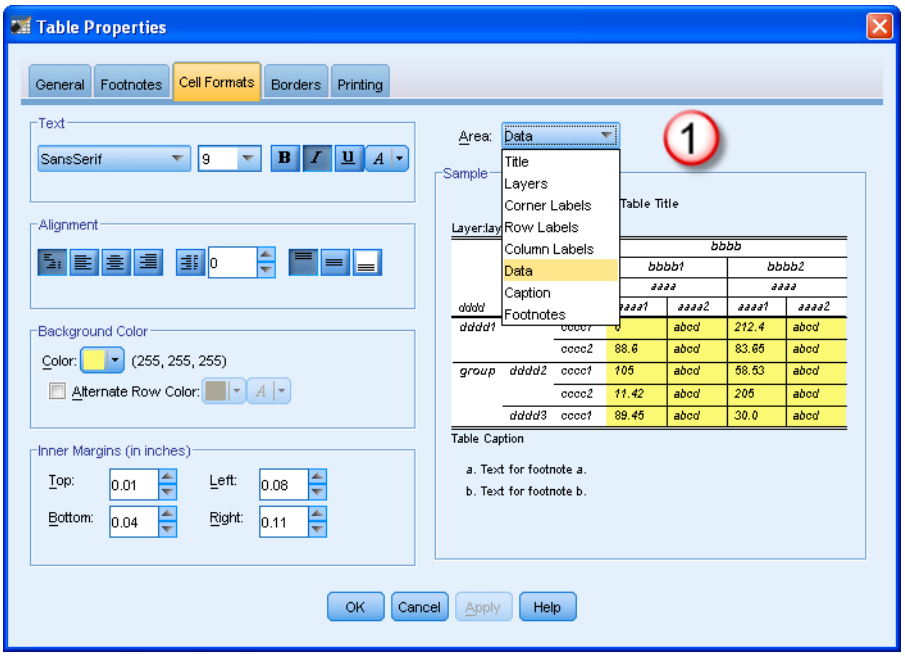
- **General:** Control maximum width for column and row labels
- **Footnotes:** Control display of footnotes
- **Cell Formats:** Control appearance of cells
- **Borders:** Control display and line width of borders
- **Printing:** Control how tables are printed, e.g., all layers

Figure 12.11 Table Properties Dialog



The Cell Formats tab deserves special attention; see the figure below. Different areas of a pivot table can be customized from here.

Figure 12.12 Cell Formats Tab in Table Properties Dialog



If you have edited and saved your TableLook, you can set it as the default TableLook for your pivot tables.

1) Select **Edit…Options**,
2) Select the **Pivot Tables** tab and browse to the TableLook you want to set as default.

Tip

In the Pivot Tables tab in the Options dialog, you can control how wide pivot tables are copied into the clipboard for rich text format (rtf). You can select from Shrink width to fit or Wrap table. Both of these selections will retain the entire table. The default *Do not adjust width* can cause some wide tables to be truncated depending on the page size and margins settings of the application into which you are pasting the table.
From this tab, you can also change the default editing mode for pivot tables. For example, you might choose to edit all tables in a separate window by default.

12.12 *Demonstration: Applying a TableLook*

In this demonstration we will use the Academic TableLook and apply it to the Frequencies table in the Viewer.

Detailed Steps to Apply a TableLook

1) Open the Frequencies table for *hapmar* table in the Pivot Table Editor
2) Select **Format…TableLooks**..
3) Select **Academic** in the list of TableLooks.

Results from Applying a TableLook

The figure below depicts the results of applying the Academic TableLook.

Figure 12.13 Table with the Academic TableLook

HAPPINESS OF MARRIAGE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	VERY HAPPY	596	29.5	61.5	61.5
	PRETTY HAPPY	343	17.0	35.4	96.9
	NOT TOO HAPPY	30	1.5	3.1	100.0
	Total	969	47.9	100.0	
Missing	DON'T KNOW	1	.0		
	REFUSED TO ANSWER	7	.3		
	System	1046	51.7		
	Total	1054	52.1		
Total		2023	100.0		

12.13 *Pasting Results to Other Applications*

Your results can be used in many applications. Applications such as Microsoft's PowerPoint or Word can display your results as plain text, rich text, or as a metafile, which is a graphical representation of the output. Or, pivot tables are pasted or exported into Microsoft Excel with each cell of the pivot table in a separate Excel cell.

If you have a single table or a small number of tables, you can copy and paste these directly into a file opened in another application. Alternatively, PASW Statistics provides an export facility to export large numbers of tables and charts into a file in a variety of common formats: Excel, Portable Document Format (PDF), HTML, Text, Microsoft Word, PowerPoint, or Excel files

12.14 *Procedure: Pasting Results to Other Applications*

Moving output to other applications can be done in various ways. First, standard Copy & Paste actions can be used:

- **Copy & Paste**. You can copy the table in PASW Statistics and paste the table into Microsoft Word as a native Word table. Text formatting, such as font size and color are retained, and columns and rows are properly aligned. Because the table is in the appropriate format, the data can be edited after you paste it into your document.

- **Copy & Paste Special**. You can choose to paste the table as a Picture. The pasted table has the same look as the PASW Statistics pivot table and is the best choice if you have done all of your editing in the Pivot Tables Editor. To paste as a picture, use Paste Special instead of Paste and choose a Picture or graphic format from the Paste Special dialog box in Microsoft Word.

To illustrate this, we will copy the happiness of marriage frequency table to Word.

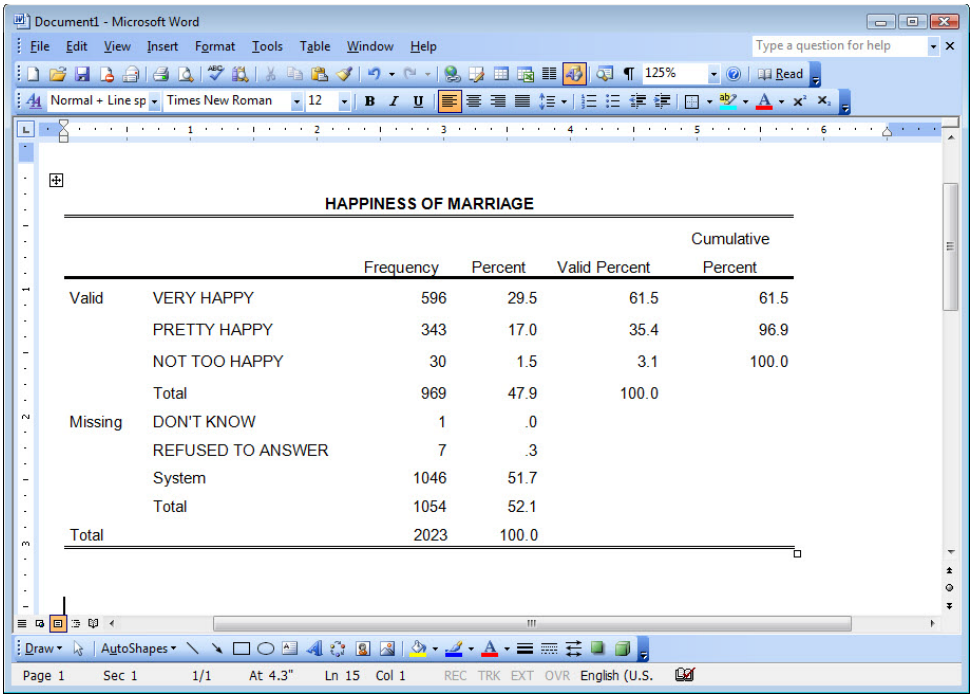
12.15 *Demonstration: Pasting a Table to a Word Processor*

Make sure that Microsoft Word or another word processor is open on your PC.

1) Select the **HAPPINESS OF MARRIAGE** frequency table
2) Select **Edit…Copy**
3) Switch to Word or another word processing application
4) Select **Edit…Paste**

The table is pasted and looks identical to that in the Viewer window. The table can now be edited.

Figure 12.14 Frequency Table Pasted into Microsoft Word

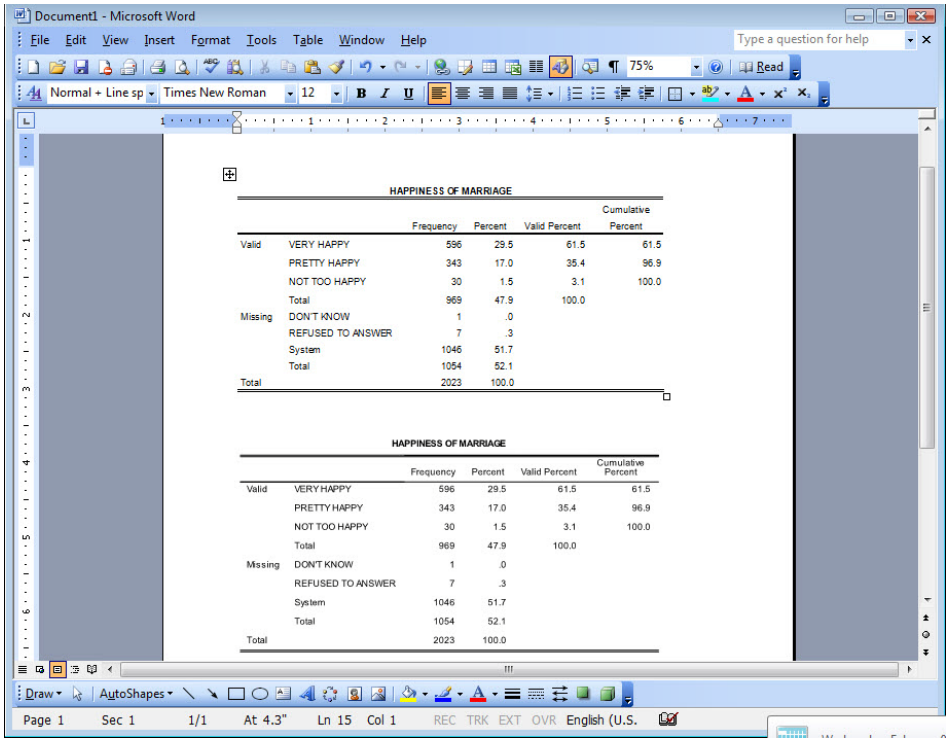


We can continue this demonstration by now pasting the table as a picture.

1) Move the cursor below the table in Word
2) Select **Edit…Paste Special**
3) Select **Picture (Enhanced Metafile)**

The table is pasted and looks similar, but not identical, to that in the Viewer, or the table we first pasted. If you try to edit it, you will discover that only limiting capability is available (chiefly moving table elements).

Figure 12.15 Frequency Table Pasted as a Picture



12.16 *Exporting Results to Other Applications*

Copying and pasting tables one-by-one might not be a good option if it involves many tables. You can export selected items, all visible items, or all items in the Viewer. If exporting to HTML, PDF, or Microsoft Word or PowerPoint, you can export charts.

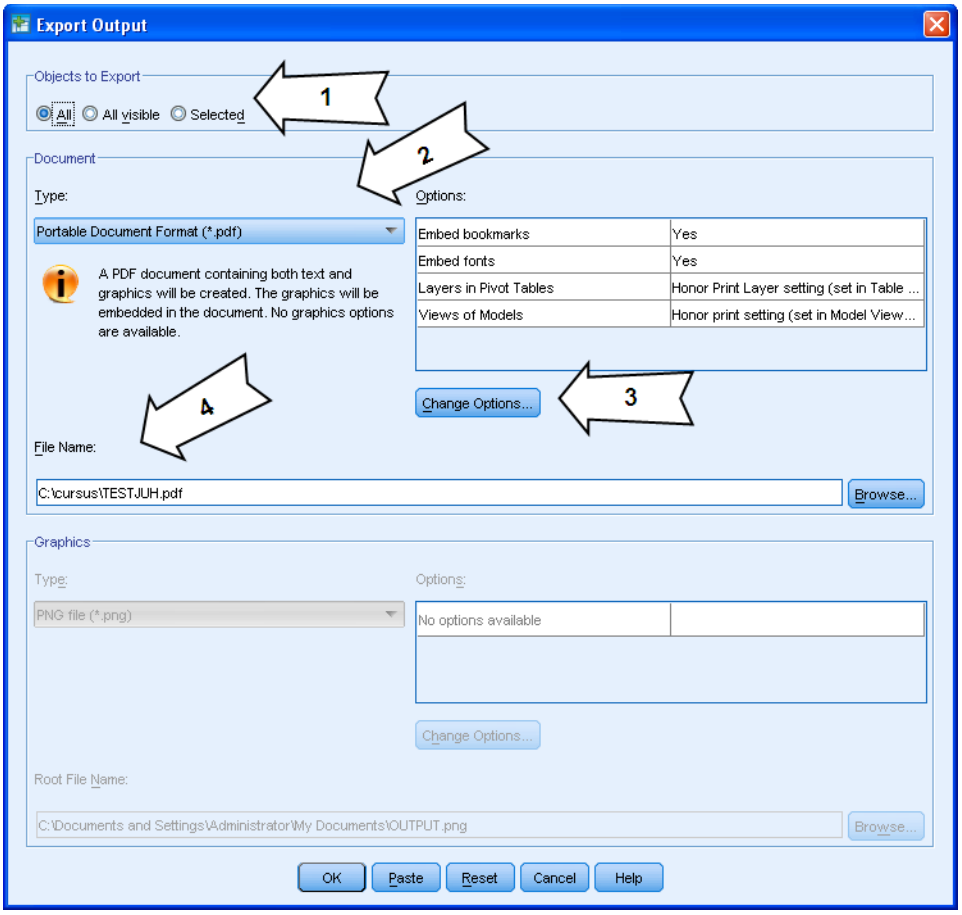
To export output:

1) Select the objects you want to export (or none if you want to export all objects)
2) Select File...Export (or right-click in the content pane of the Viewer and select Export from the pop-up menu).

Then, in the Export Output dialog:

1) Select objects to export.
2) Select file format.
3) Optionally, change options for the selected file format.
4) Specify file name for the export.

Figure 12.16 Export Output Dialog



In the Objects to Export area, you can specify whether you want to export All objects, All Visible Objects or Selected Objects. Instead of exporting all objects in the Viewer, you can choose to export only visible objects (open books in the outline pane) or those that you selected in the outline pane. If you did not select any items in the outline pane, you do not have the option to export selected objects.

There are several options for exporting the results. In the Type dropdown list, you specify the file type that you want to export to. The choices are:

- HTML file(*.htm)
- Text file (*.txt) – Plain text, UTF-8 or UTF-16 formats
- Excel file (*.xls)
- Word/RTF file (*.doc)
- PowerPoint file (*.ppt)
- Portable Document Format (*.pdf)
- None (Graphics only) – Several graphics formats are available

Each of these file types has default format options which you can change with the Change Options button. In the Options dialog for exporting to Word/RTF format, for example, you can control how layered tables are exported. You can control how wide pivot tables are exported:

- Wrap table to fit within page margins
- Shrink width to fit within page margins
- Do not adjust width.

You can also change the page setup (margins and page orientation) by clicking the Page Setup for Export button.

When you export results to Excel, you can choose to:

- Create a new workbook
- Create a new worksheet in an existing Excel file
- Modify an existing worksheet in an Excel file.

Further, you can name the worksheet and specify the location within the Excel worksheet for the exported tables.

12.17 *Demonstration: Exporting to Word*

In this first demonstration, we'll export the Frequencies and Crosstabs tables in the Viewer window to Word.

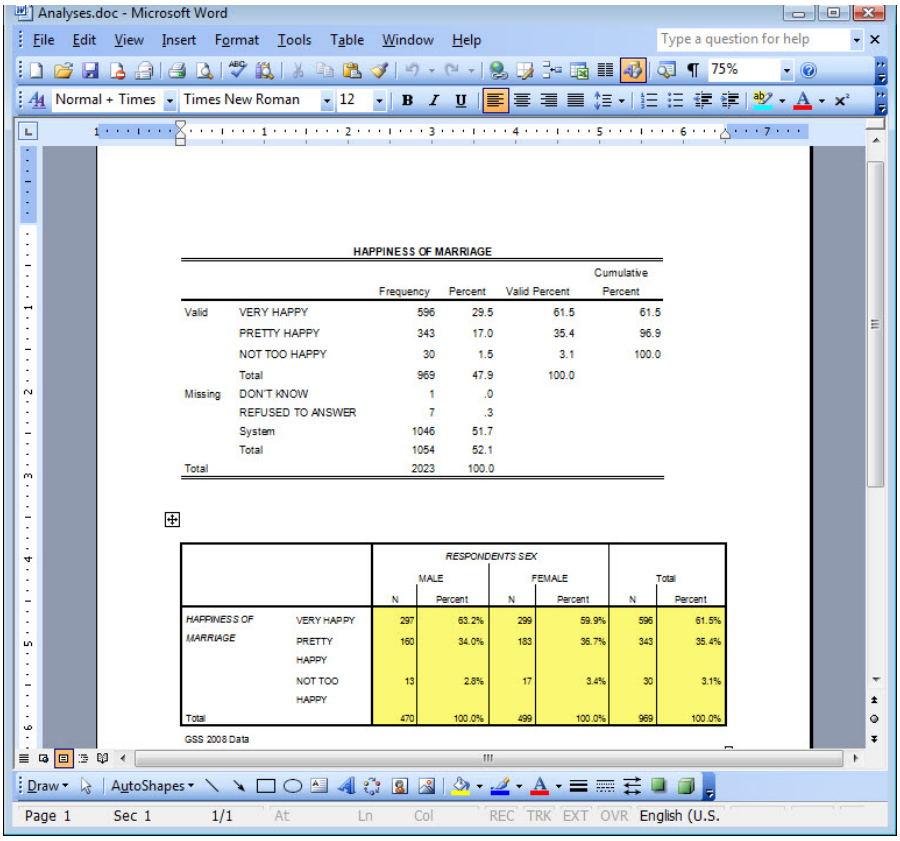
Detailed Steps to Export to Word

1) In the Viewer, select the frequency table and crosstabulation table.
2) Select **File…Export**.
3) Select **Selected** in the Object to Export: area
4) Select **Word/RTF (*.doc)** from the Type drop-down
5) Select **Browse**, and then move to the folder **c:\Train\StatisticsIntro**
6) Name the file **Analyses.doc.**
7) Select the **Change Options** button
8) Select **Shrink width to fit within page margins**.

Result of Exporting to Word

If we open the exported Word file, it will contain the two tables, and only that. The tables will fit the page width, thanks to the export option to shrink wide tables. Furthermore, the tables can be edited with the Word table editor, because the tables are exported as text, not as a picture.

Figure 12.17 Exported Output to Word



Apply Your Knowledge

1. To which of these formats/software can PASW Statistics export output? Name all that apply.
    a. Access
    b. Excel
    c. PDF
    d. HTML
    e. Word

12.18 *Lesson Summary*

We discussed the Viewer in this lesson, including managing output objects, editing pivot tables in the Pivot Table Editor, and exporting output to other applications.

Lesson Objectives Review

After completing this lesson students will be able to:

- Use the Output Viewer and its editing features

And, they should also be able to:

- Navigate through the Viewer
- Customize a pivot table
- Create and apply a template for a pivot table
- Export output to other applications

12.19 *Learning Activity*

The overall goal of this learning activity is to edit tables, apply a template for tables, and export output to other applications.

Supporting Materials

The data file *employee data.sav* contains information on employees of a major bank. Included is data on beginning and current salary position, time working, and demographic information.

1. Open the data file *employee data.sav*.

2. Use the Means procedure (Analyze…Compare Means…Means) to create a table for current salary (*salary*) by job category (*jobcat*). Hint: place *salary* in the Dependent List box).

3. Edit the table so that it looks like this:

Salaries by Job Category[a]

Variables=Current Salary

	Job Category			
	Clerical	Custodial	Manager	Total
Mean	27755.71	30975.00	63977.80	34450.27
N	357	26	84	467
Std. Deviation	7567.228	2147.987	18244.776	17188.710

a.Bank employee data

4. Apply several TableLooks to the table. Evaluate their strengths and weaknesses for the audience to whom you will be presenting your results

5. Hide the Total column. To do so, highlight the Total labels, right-click and choose Select…Data Cells and Label. Then select View…Hide from the menu. Then display it again with View…Show all Categories.

6. Export this table to a Microsoft Word file using options you prefer.

7. Export this table to an Excel file using options you prefer.

8. *For those with more time:* (Export this table to another format that you think you might use in your work.

Lesson 13: Syntax Basics

13.1 *Objectives*

After completing this lesson students will be able to:

- Use basic syntax to automate analyses

To support the achievement of this primary objective, students will also be able to:

- Use the Syntax Editor environment
- Create syntax
- Run syntax
- Edit Syntax using auto-completion of commands

13.2 *Introduction*

PASW Statistics syntax provides a method for you to run commands without navigating through dialog boxes, the Viewer, or the Data Editor. Instead, you control the application through a straightforward command language. Nearly every action you can achieve through the user interface can be achieved through syntax. Using PASW Statistics syntax, you save the exact specifications used during a session and then can run and/or modify them to repeat a series of operations. This is especially useful
if you have a series of PASW Statistics tasks that you repeat on a regular basis, such as reports that you must produce periodically.

You can paste commands from the menus, copy them from the Journal File or Log, or type them directly into the Syntax Editor window, to create syntax.

Business Context

The topics in this lesson can help in answering questions such as:

- If I have to run the same analyses periodically, what is the best way to proceed?
- If I have run analyses for a selection of cases, but I want to repeat these analyses for another selection of cases, how can I repeat this analysis efficiently?
- How can I keep track of exactly what I've done in a PASW Statistics session?

Supporting Materials

The file *census_small.sav*, a PASW Statistics data file from a survey done on the general adult population. Questions were included about various attitudes and demographic characteristics.

Further Info

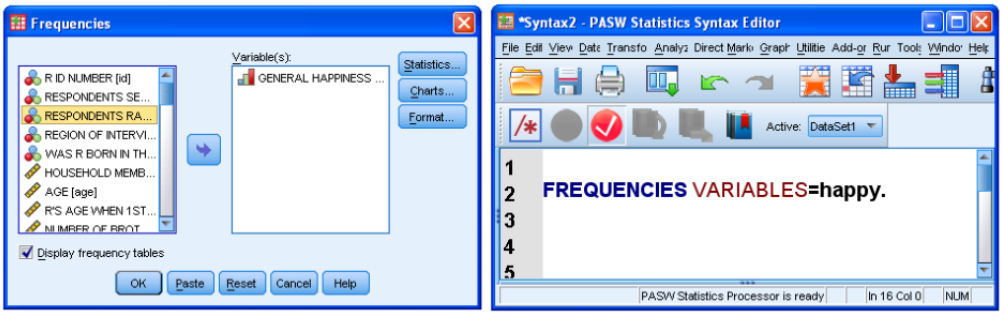
This lesson discusses the basics of working with syntax. The *Syntax I* and *Syntax II* courses provide complete information on working with syntax.

13.3 *Syntax Compared to Commands from Menus*

With syntax you control PASW Statistics through a command language. This command language is actually generated when you use the menus, and is visible in the Log in the Viewer window.

To compare menus to syntax, the figure below shows the **Frequencies** dialog box on the left, and on the right the Frequencies command in syntax that is the equivalent.

Figure 13.1 Dialogs and Syntax Compared



When you click OK in the **Frequencies** dialog box, the Frequencies command is run automatically behind the scenes. Working with syntax, the user constructs the Frequencies command in the Syntax Editor window and then runs the command from there. But in both cases, the same Frequencies command is run.

Most commands are available from the menus. However, some commands and options are only available by using the command language, which is another motivation to learn syntax.

13.4 *Pasting From Dialogs to Generate Syntax*

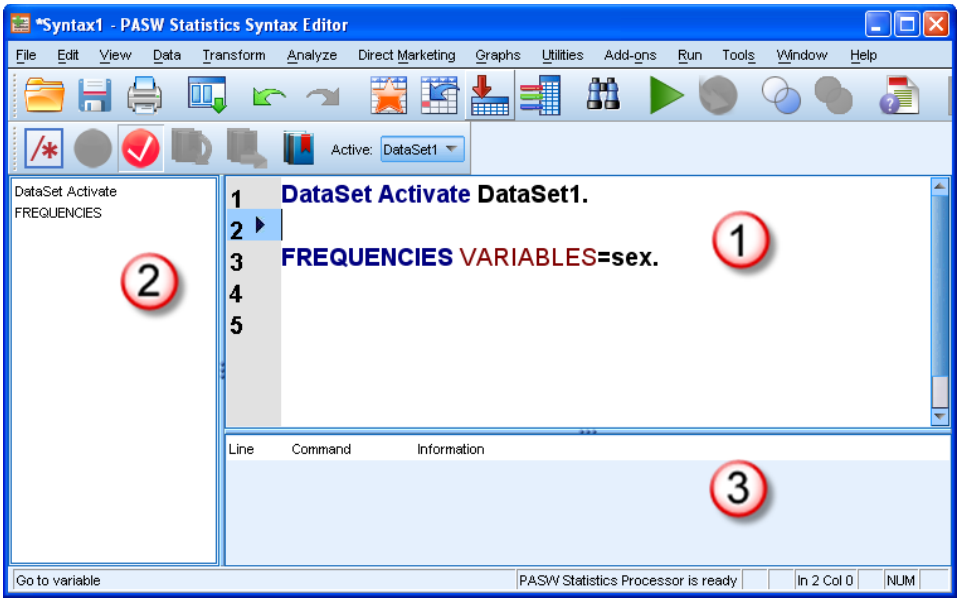
There are several methods to create syntax. In this lesson, we will discuss one of these, pasting syntax from a dialog window. From this perspective, dialogs are essential to create syntax: syntax is built from the dialog boxes by selecting the Paste button. So, it´s not an either-or choice between working with dialogs and syntax, as one method complements the other. As a new user of PASW Statistics, the dialog boxes are your best friend to learn and use syntax.

13.5 *The Syntax Editor*

The Syntax Editor window provides an environment in which to write, modify, and run syntax commands.

- The syntax commands are edited in the editor pane (1).
- The navigation pane (2) displays a list of all of the commands and allows you to easily navigate to any command.
- The error pane (3) at the bottom displays any errors encountered in the commands when they are run.

Figure 13.2 The Syntax Editor



A very helpful feature is that commands are color coded.

- Commands are colored blue and in bold text.
- If, however, there is a recognized syntactical error within the command, such as a missing parenthesis or unmatched quotes, the command name is colored red and in bold text.
- A subcommand is colored green in the Syntax Editor and keywords are colored maroon.
- User-specified values, such as file names, variable names and numbers are not color coded.

Pasting Syntax

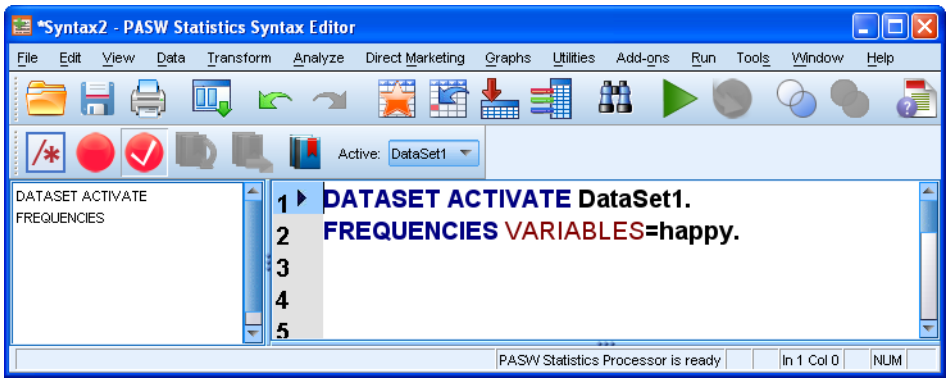
Creating and using syntax from the dialog boxes involves the following steps:

1) Select the dialog for the procedure you want to run.
2) Complete the specifications for the dialog.
3) Select the Paste button in the dialog.
4) Switch to the Syntax Editor.
5) Run the syntax from the Syntax Editor.
6) Inspect the output.

Pasted Syntax

Pasting from a dialog will put the commands in the Syntax Editor window. Occasionally, a Paste will not result in one command only, but two or more. For instance, if we paste from the **Frequencies** dialog, the pasted syntax might look as depicted below. Here, two commands are generated, DATASET ACTIVATE and FREQUENCIES. The navigation pane on the left outlines the commands.

Figure 13.3 Pasted Syntax --- More Than One Command Pasted from a Dialog

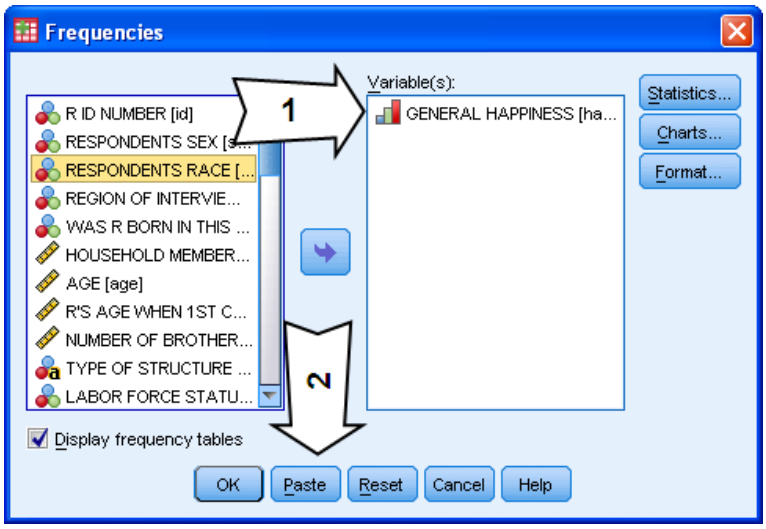


13.6 *Procedure: Pasting and Running Syntax*

Creating and running syntax is accomplished with these steps:

1) Open the dialog and complete specifications.
2) Select Paste in the dialog.

Figure 13.4 Pasting from a Dialog



When the user hits Paste and there is no Syntax Editor window open, PASW Statistics will create a new Syntax Editor window. This window contains the pasted syntax and the focus is automatically on this window. However, if there is already a Syntax Editor window open, and the user then selects the Paste button in a dialog box, syntax will be pasted to the open window. However, focus is not automatically on that window, and the user has to switch to that window in order to get to the syntax.

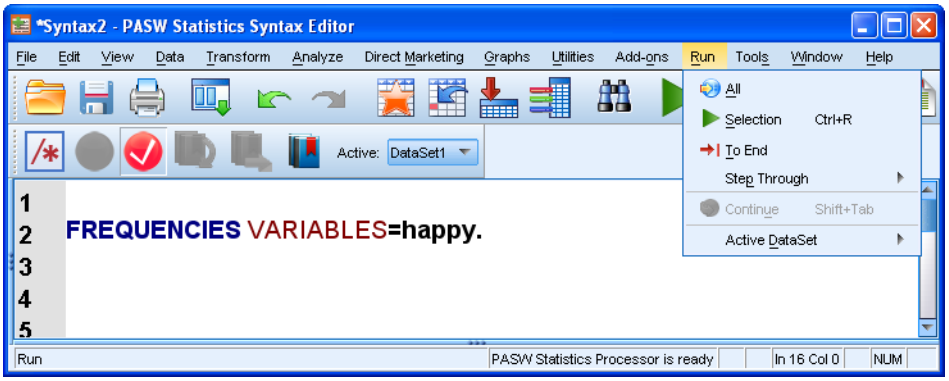
To run (execute) the syntax, the user has several options:

- **Run…All**: runs all commands in the Syntax Editor window
- **Run…Selection**: runs commands selected (highlighted with the mouse) in the Syntax Editor. If no commands are selected, it will run the command where the cursor has focus.
  Run…To End: runs from cursor to the last command
- **Run…Step Through**: runs one command at a time from the first command or the current command

- **Run…Continue:** to further control the execution of the Syntax commands, you can set breakpoints which allow you to stop execution of command syntax at specified points and continue execution when ready.

The Toolbar button  provides a quick alternative to run a selection or the current command that has focus.

Figure 13.5 Running Syntax



Note

The Active DataSet window can be selected from the Run menu. If there are multiple datasets, the user has control over the dataset on which the command should be run.

Tip

If you build syntax and want to test the commands, the best way is to run the commands one-by-one and see if they run successfully. If you are sure that a series of commands does the requested job, then you can run all commands at once.

13.7 *Demonstration: Pasting and Running Syntax*

We will work with the *census_small.sav* data file in this lesson. We will build syntax to run the **Frequencies** command on the variable *HAPPINESS OF MARRIAGE [hapmar]*. The starting point is that no Syntax Editor window is open yet.

Detailed Steps for Pasting And Running Syntax

1) Select **Analyze…Descriptives Statistics…Frequencies**.
2) Place ***HAPPINESS OF MARRIAGE [hapmar]*** in the Variables box
3) Select the **Paste** button.

Results for Pasting and Running Syntax

The **Frequencies** command is pasted to a new Syntax Editor window and focus is set to this window. Actually, two commands are pasted: **Dataset Activate** and **Frequencies**. The **Dataset Activate** command assures that the correct dataset is chosen on which the **Frequencies** command will be run.

If we paste syntax from additional dialog boxes, the commands will be pasted into this open syntax window.

Now we run the **Frequencies** command.

1) Select **Run…Selection**

The output from Frequencies is created, just as if we'd used the dialog box only.

Figure 13.6 Frequency Table for Happiness of Marriage

HAPPINESS OF MARRIAGE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	VERY HAPPY	596	29.5	61.5	61.5
	PRETTY HAPPY	343	17.0	35.4	96.9
	NOT TOO HAPPY	30	1.5	3.1	100.0
	Total	969	47.9	100.0	
Missing	DON'T KNOW	1	.0		
	REFUSED TO ANSWER	7	.3		
	System	1046	51.7		
	Total	1054	52.1		
Total		2023	100.0		

Apply Your Knowledge

1. True or false? PASW Statistics generates syntax even when you select the OK button in a dialog box?

13.8 *Editing Syntax and Running Edited Syntax*

When you are in the Syntax Editor window you can create additional commands yourself. You can:

- Write a command from scratch
- Type in a command or part of a command and use the feature of auto-completion to complete the command

Of course, there is no guarantee that syntax created in this way will run without errors. In this section, we will demonstrate auto-completion and focus on errors arising from running syntax.

Steps in Editing and Running Syntax

The steps to edit syntax and run the edited commands are:

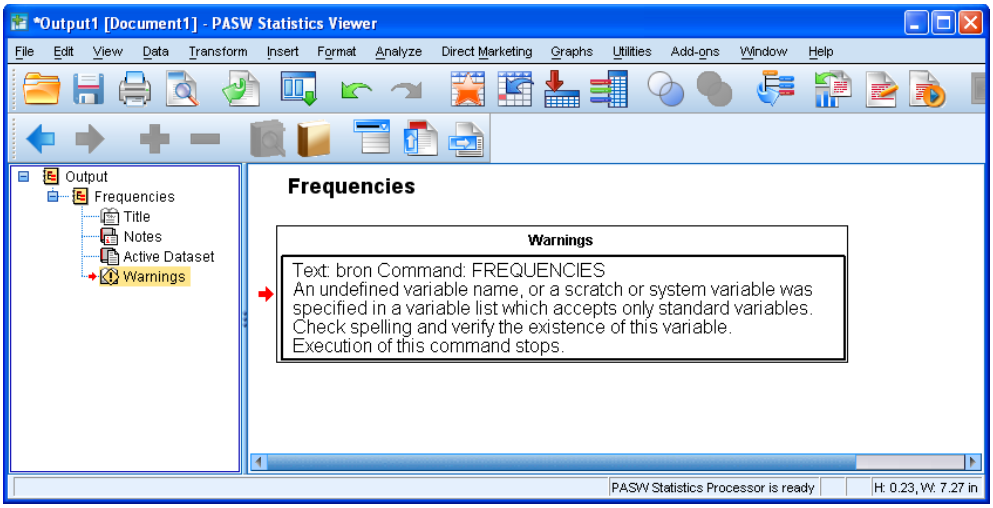
1) Switch to the Syntax Editor.
2) Edit the syntax.
3) Run the syntax from the Syntax Editor.
4) Inspect the output for errors.

Results of Running Edited Syntax

In this context, the term "output" refers not only to the Viewer document, but to messages in the Syntax Editor itself. When the command runs okay, then the result will be displayed in the Viewer, just like any analysis run from the dialogs. But if a command results in a warning or error, a message

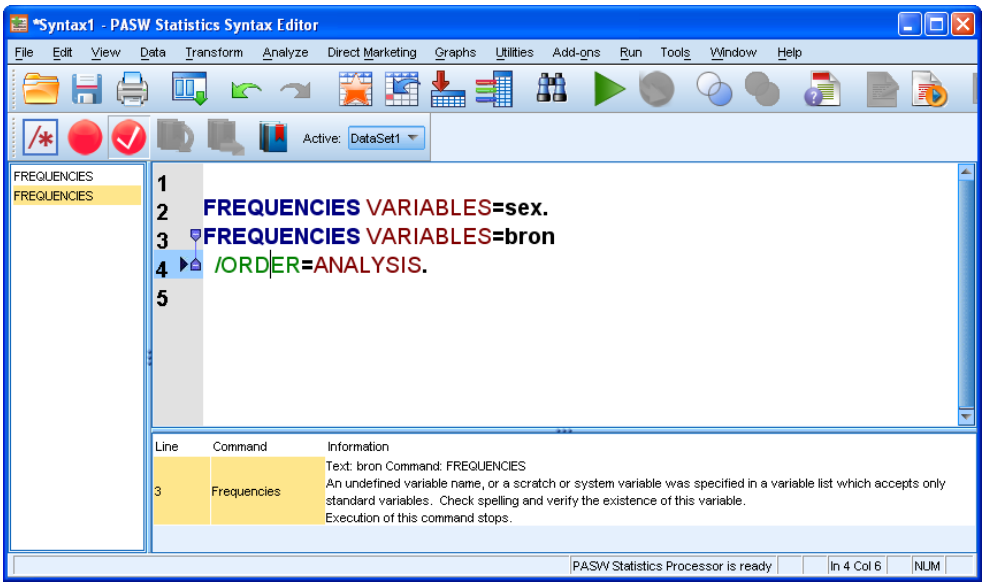
will be displayed both in the Viewer and in the Syntax Editor. The Viewer will indicate what command generated the error, but does not report all pertinent details.

Figure 13.7 Error Message in the Viewer



In the Error pane, below the Editor pane, the line number of the command that caused the warning/error is listed. Therefore, the Error pane in the Syntax Editor complements the information in the Viewer.

Figure 13.8 Error Message in the Syntax Editor

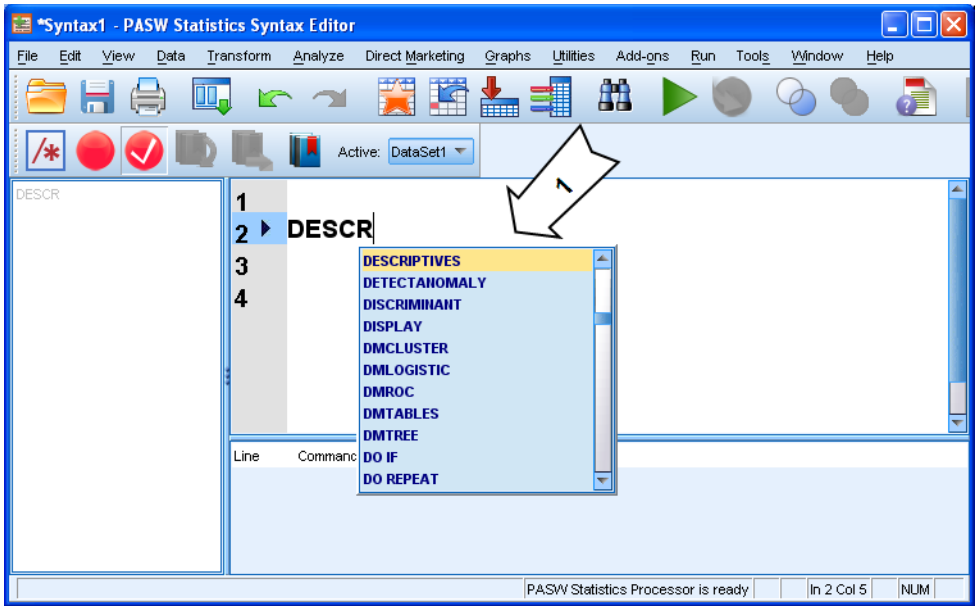


13.9 *Editing Syntax*

To edit syntax you use standard methods like copy, cut and paste. The Syntax Editor also provides assistance in the form of auto-completion of commands, subcommands and keywords.

1) Type in the first letters of a command, subcommand or keyword and select the relevant option from the pop-up menu.

Figure 13.9 Auto-completion of Syntax



Tip

You can display the auto-completion list on demand by pressing Ctrl+Spacebar, and you can close the list by pressing the Esc key.

13.10 *Demonstration: Editing and Running Syntax*

Suppose we want to customize the **Frequencies** command in order to request a bar chart with percentages, for the variable *happy* (rather than *hapmar*).

Detailed Steps for Editing and Running Syntax

1) Switch to the Syntax Editor
2) Replace ***hapmar*** with ***happy***
3) Put the cursor behind the variable name ***happy*** and add a **space**.
4) Type a slash ***/***.
5) Select BARCHART from the pop-up menu.
6) Add a **space**.
7) Hit **Ctrl+Spacebar** and select **PERCENT** from the pop-up menu.
8) Run the command with the **Run** button

Results Running Edited Syntax

Running the **Frequencies** command will produce the requested graph in the Viewer along with the frequencies table.

Figure 13.10 Bar Chart of General Happiness

13.11 *Demonstration: Syntax Errors*

In this demonstration we rerun the **Frequencies** command, but now will misspecify a variable name to illustrate what happens when PASW Statistics encounters an error.

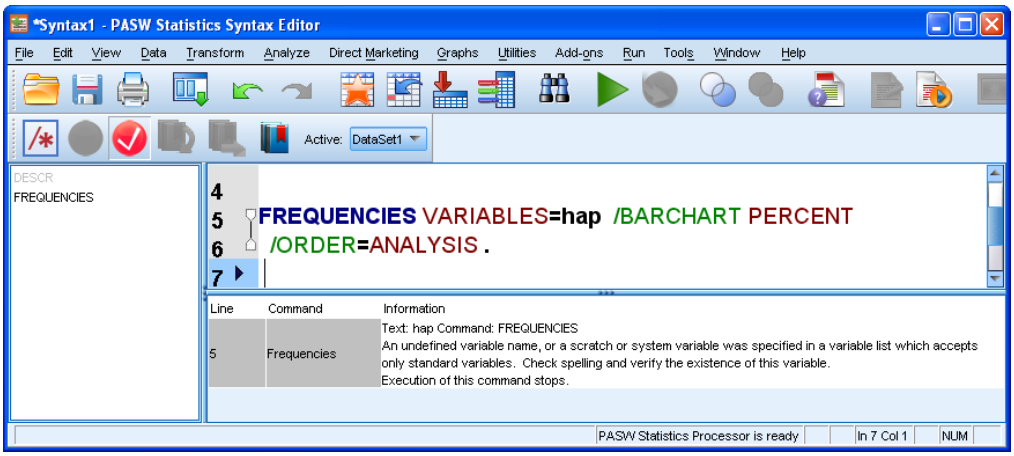
Detailed Steps for Syntax Errors

1) Switch to the Syntax Editor.
2) Replace ***happy*** with ***hap***.
3) Run the **Frequencies** command.

Results from Syntax Errors

Both the Viewer and the Syntax Editor display a warning. The message in the Syntax Editor is more informative, as it details the line number of the probable cause.

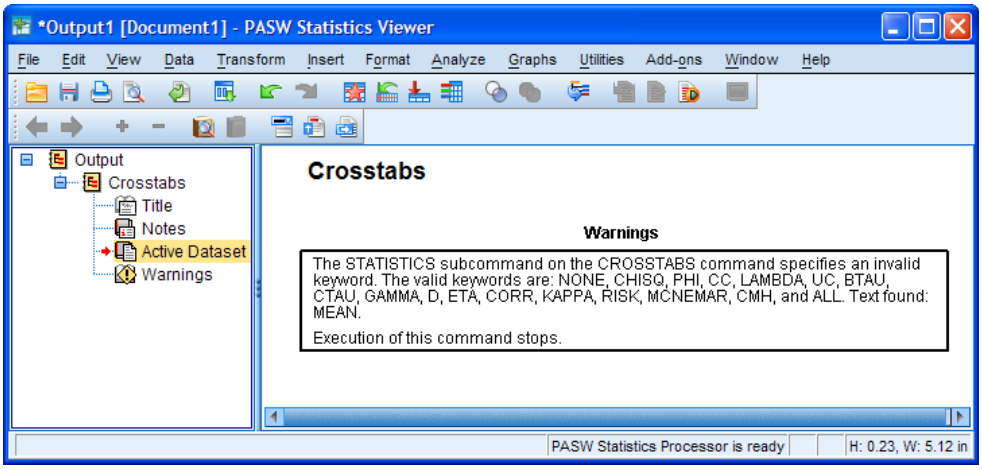
Figure 13.11 Error Message in Syntax Editor



The message in the Error pane of the Syntax Editor window clearly refers to line 5 where the error occurred and suggests that the variable name is inappropriate.

Apply Your Knowledge

1. Suppose you run this command: **CROSSTABS hapmar BY sex / STATISTICS = MEAN.** Why would you receive the warning below in the Viewer?
   a. The command name is incorrect
   b. There is no STATISTICS subcommand
   c. The wrong statistic is specified on the STATISTICS subcommand
   d. The variable name is incorrect



13.12 *Lesson Summary*

After completing this lesson students will be able to:

- Use basic syntax to automate analyses

And, they should also be able to:

- Use the Syntax Editor environment
- Create syntax
- Run syntax
- Edit Syntax using auto-completion of commands

13.13 ***Learning Activity***

In this exercise you will use the data file *employee data.sav* to create and edit syntax.

Supporting Materials

The data file *employee data.sav* contains information on employees of a major bank. Included is data on beginning and current salary position, time working, and demographic information.

1. Create syntax to open the data file *employee data.sav*. You can use the dialog box under File…Open…Data.
2. Create syntax to run frequencies for *gender* and *jobcat*. Add a barchart and ask for percents.
3. Create syntax to run a crosstabulation of *jobcat* by *gender*, with column percentages.
4. Run these commands one-by-one. Then run all the commands together.
5. Edit the syntax to request a chi-square test for the crosstabulation. Hint: Use the Statistics subcommand.
6. Edit the syntax to leave out the word “by” in the CROSSTAB command. What error do you receive when you run the command?
7. Fix the CROSSTAB command. Then save the syntax file to *my syntax.sps*.
8. Exit PASW Statistics, and then open the software again. Open the file *my syntax.sps* and run all the commands again.

Lesson 14: Course Summary

14.1 *Course Objectives Review*

Now that you have completed the course, you should be able to:

- Perform useful analyses on their data using PASW Statistics

And, you should also be able to:

- Explain the use of PASW Statistics for basic data analysis
- Import data from different types of file formats
- Define, save and view variable properties
- Use the Data Editor to enter data values, work with Data Editor features and multiple Data Editor windows
- Summarize individual variables using tables and graphs
- Group values of variables using various methods
- Create new variables using the **Compute Variable** dialog
- Analyze relationships between categorical variables and between categorical and scale variables
- Select cases in a data file using various methods
- Present results with charts
- Use the Output Viewer and its editing features
- Use basic syntax to automate analyses
- Use the Toolbars
- Use all types of help available in PASW Statistics

14.2 *Course Review: Discussion Questions*

1. Which method would you use to add or edit variable properties? Why?
2. How do or would you use charts to present your results? Would you use PASW Statistics or a different application to create the charts? What advantages does each have?
3. What order of display of variables in dialog boxes do you prefer? Why? Do you prefer variable names or labels?
4. What other statistical methods beyond those covered in this course to analyze your data? Why?
5. How might you use Visual Binning facility on your data?
6. In what format do you present your results? How might you use the editing and exporting facilities in PASW Statistics to achieve the presentation that you need?

14.3 *Next Steps*

Thought Starters

How can you immediately apply what you have learned about using PASW Statistics to better understand your data and provide results that will aid in decision making?

List your top three:

1._____

2._____

3._____

Next Courses

This course introduced you to PASW Statistics. In this section we provide direction for what courses you can attend to broaden your knowledge in specific areas.

If you want to learn more about this:	Take this course:
Data management and manipulation	Data Management and Manipulation with PASW Statistics
Basic syntax and its application	PASW Statistics Syntax I
More advanced use of syntax	PASW Statistics Syntax II
Basic statistics: Crosstabs, T Tests, Correlations, etc.	Introduction to Statistical Analysis Using PASW Statistics
Custom Tables	Presenting Data with PASW Statistics Custom Tables

Appendix A: Menus and the Help System

A.1 *Objectives*

After completing this lesson students will be able to:

- Use the Toolbars
- Use all types of help available in PASW Statistics

To support the achievement of this primary objective, students will also be able to:

- Describe the operation of the toolbar buttons
- Describe the options for help in PASW Statistics
- Access the various types of help

A.2 *Introduction*

The toolbars in PASW Statistics have many features in common with other software applications, but also some unique characteristics. In this appendix we review the actions that can be taken with the choices in each menu item in the various PASW Statistics windows.

PASW Statistics includes several types of Help, including tutorials, case studies, reference documents, and standard help as in other software applications. We review the various types of Help so you are aware of and comfortable accessing them to assist your use of the software.

A.3 *PASW Statistics Main Menu and Toolbar*

Each type of window in PASW Statistics has its own specific menu system and toolbar. Several of the menu selections can be accessed from more than one menu. For example, all procedures can be accessed from both the Output Viewer menu and the Data Editor menu. The selections available on the main menu in the Data Editor window are:

File Use the File menu to read data from existing PASW Statistics data files, and spreadsheets, text, or database files created by other software. You also use this menu to save the data to a PASW Statistics data file or to export to other file formats or databases, to print the contents of the Data Editor, to display data file information, to connect to the PASW Statistics Server or PASW® Collaboration and Deployment Services systems (if installed), to exit PASW Statistics, and to perform other functions common to the File menu item on most applications.

Edit Use the Edit menu to perform standard Windows functions to cut, copy and paste selections and to find and replace data values. You also use this menu item to insert variables and cases and to modify PASW Statistics environment settings.

View Use the View menu to display gridlines, labels, the status bar and toolbars, to change the display font and to edit the menus. From the Variable View tab, you can select variable attributes to display and reorder their display in the Variable View.

Data Use the Data menu to access the PASW Statistics facilities that make global changes to PASW Statistics data files, such as sorting files by cases or variables, restructuring files, merging files and creating subsets of cases for analysis. These changes are made only on the temporary file displayed in the Data Editor and do not change the permanent file unless you explicitly save the file with these changes. You can define a custom attribute from this menu as well.

Transform Use the Transform menu to access PASW Statistics facilities that modify or create new variables in the data file. You can compute new variables, recode variables, bin values of scale variables, manipulate date/time variables, and create variables with lagged values from this menu.

Analyze Use the Analyze menu to select the PASW Statistics statistical and reporting procedures you have installed with PASW Statistics. This menu contains all of the PASW Statistics procedures included in the PASW Statistics Base system, such as Frequencies, Crosstabs (covered in this course) as well as other descriptive procedures, regression, analysis of variance and many more. As well, this menu will contain procedures from any of the PASW Statistics add-on options that you have installed such as Tables (Custom Tables Option), Mixed Models (from Advanced Statistics Option), etc.

Graphs Use the Graphs menu to create charts using the Chart Builder or the Legacy Dialogs. As of PASW Statistics 17, the new Graphboard Template Chooser allows you to produce chart visualizations using a visualization template. Some statistical procedures also optionally generate charts.

Utilities Use the Utilities menu to display variable information and to define and use variable sets to control the variables that appear in the Data Editor and in the variable lists of dialog boxes. You can also add comments to your data file and access the OMS (Output Management System). You can create, edit and run Production Jobs from this menu and check spelling of variable and value labels in the Variable View and check spelling in string (alphanumeric) data values. Also, you access the feature to build, save, and install Custom Dialogs from this menu.

Add-ons Use the Add-ons menu to display information about other PASW Statistics add-on options and SPSS, an IBM Company software that you haven't installed as well as information about Consulting and Training Services, PASW Statistics Programmability Extension feature, and Statistics Guides.

Window Use the Window menu to switch between PASW Statistics windows and manipulate how they appear on the screen.

Help Use the Help menu item to access the many Help features of PASW Statistics. Context sensitive help is also available using the Help button in the dialog boxes and through activated pivot tables and charts.

You can also access many of the most commonly used menu features using the toolbar in each PASW Statistics window. A toolbar is associated with each type of window (Data Editor, Output Viewer, Chart Editor, Pivot Table Editor or Syntax Editor) and contains tool buttons specific to that particular window. You can move toolbars, attach them to a new location or create a separate toolbar palette using standard click and drag Windows operations.

Figure A.1 PASW Statistics Data Editor Toolbar

The Data Editor toolbar contains common edit operations and operations associated with the Main Menu. You can place the cursor over any tool button and an explanation label will appear. The most commonly used tools on this toolbar are listed below. The first three tool buttons are common to most Windows applications and are common to all types of PASW Statistics windows.

Open document – Displays the *Open* dialog box for the type of document that is in the currently active window. For example, we used it to open a data file in the Data Editor window.

Save this document – Saves the contents of the current window. If the file has not been named the *Save As* dialog box will open. Otherwise, you will overwrite the current file with contents of the current window.

Print – Opens the *Print* dialog box appropriate for the current window.

Recall recently used dialogs – Recalls a list of the twelve most recently used dialog boxes from which you can select. This is a useful shortcut for accessing frequently used dialogs.

The items below are all specific to a Data Editor window. So long as you are in the same Data Editor window in the same PASW Statistics session, the settings made previously are retained.

Undo/Redo – Undoes or redoes the most recent user action made in the Data Editor window. It does not undo/redo actions made from dialog boxes.

Go to Case – Opens the *Go to* dialog *Cases* tab where you can enter a case's row number to go directly to any case in the Data Editor.

Go to Variable – Opens the *Go to* dialog *Variable* tab where you can select a variable name to move directly to that variable in either the Data View or the Variable View.

Variables – Activates the *Variables* dialog which lists the variable information for each variable in the current dataset.

Find – Finds and optionally replaces specified data values in the current variable. Can be used in the Variable View tab, to find and replace content in variable name and label, value labels, and missing values attributes.

Insert Cases – Inserts a case above the currently selected case.

Insert Variables – Inserts a variable to the left of the currently selected variable.

Split File – Splits the working data file into subgroups for analysis. Also available from the **Data…Split File** menu (discussed in the lesson *Selecting Cases*).

Select Cases – Selects a subset of cases for analysis. Also available from the **Data…Select Cases** menu (discussed in the lesson *Selecting Cases*).

Value Labels – Toggles between displaying data values and value labels in the Data View.

Spell Check – In the Variable View, spell check variable and value labels. Also, you can spell check the contents of string (text) variables in the Data View.

Use Variable Sets – Opens the *Variable Sets* dialog to select which variable sets to use in dialog boxes and display in the Data Editor. Also available from the **Utilities…Use Variable Sets** menu.

A.4 *Using the PASW Statistics Help System*

There are several types of help available in PASW Statistics. Every window has a Help menu on the menu bar, and these types of help are accessed from there:

- **Topics.** Provides access to the standard Help system, where you can use the Contents, Index, and Search tabs to find topics.
- **Tutorial.** Provides access to the introductory tutorial. Other items provide more specialized information such as the statistical algorithms used in PASW Statistics.
- **Statistics Coach.** Provides a wizard-like method for suggesting a statistical or charting procedure for the type of analysis you want to do.
- **Case Studies.** Provides hands-on examples of how to set up various types of statistical analyses and interpret the results. The sample data files used in the examples are also provided so that you can work through the examples to see exactly how the results were produced.

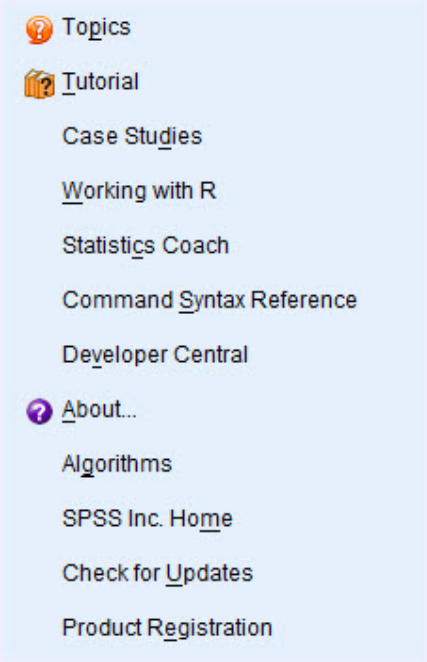
Help is available in other areas of PASW Statistics. This includes:

- **Dialog box Help buttons.** Most dialog boxes have a Help button that takes you directly to a Help topic for that dialog box. The Help topic provides general information and links to related topics.
- **Context Help in Pivot Tables.** Context menus are available using the right mouse button. Right-click on terms in the Pivot Table Editor in the Viewer and select “What's This?” from the context menu to display definitions of the terms.

Standard Help

The Help menu provides a wide range of help about PASW Statistics including a tutorial, complete documentation on the statistical algorithms used in all PASW Statistics procedures, and the Command Syntax Reference that documents the PASW Statistics command language.

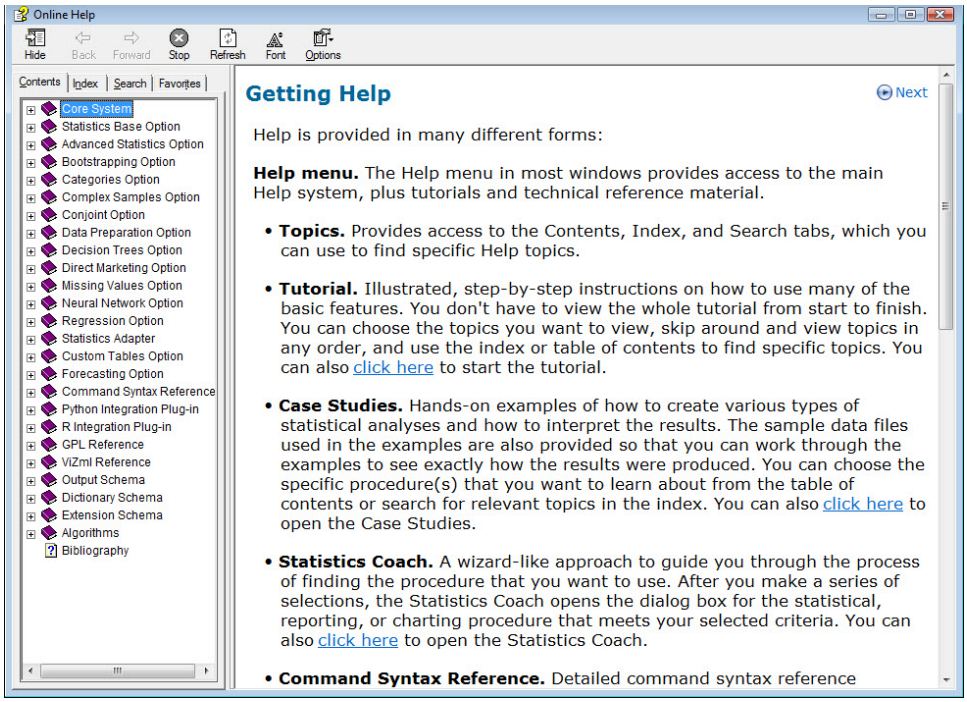
Figure A.2 PASW Statistics Help Menu



The Topics item on the Help menu opens a Help window.

The four standard Help tabs, *Contents*, *Index*, *Search*, and *Favorites* are available. The *Contents* tab lists the PASW Statistics Base and all of the add-on Options. It is most useful if you're looking for general information or are unsure of what index term to use to find what you're looking for. The *Index* tab provides a searchable index that makes it easy to find specific topics. It uses incremental search to find the closest match to what you're looking for. The Index topics are organized in alphabetical order, just like a book index. You can use the *Search* tab for a broader search than the Index provides and use the *Favorites* tab to "tag" frequently used Help topics.

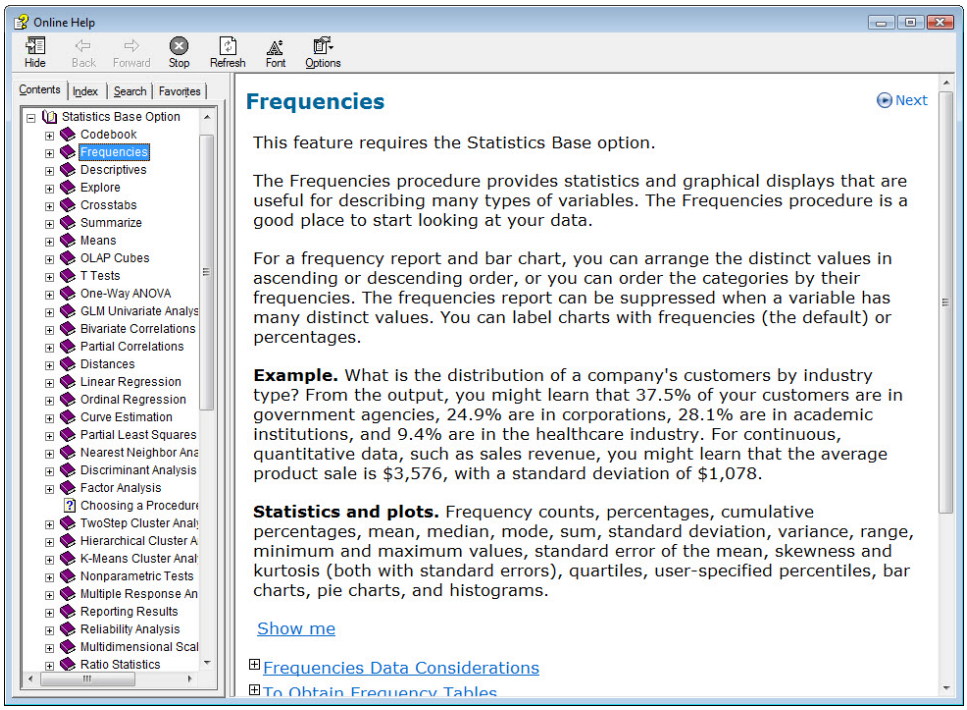
Figure A.3 Help Contents Tab



Dialog Box Help

Most dialog boxes have a Help button that displays information about what the dialog box does and how to do it. It also provides additional links related to the procedure or features. The figure below shows the Help available from the main Frequencies dialog.

Figure A.4 Frequencies Dialog Box Help

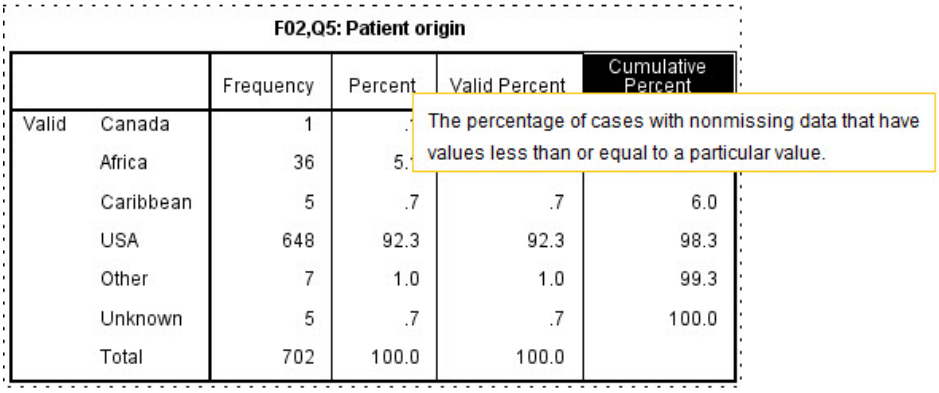


Context Help in Pivot Tables

Context-sensitive help is available for pivot table results. A definition of the term or additional information about how to use the option is displayed. Use the right-mouse button to access this help from an activated pivot table and select “What's This?” from the context menu to display definitions of the terms.

For example, you can use context-sensitive help to get definition of statistical term in a Frequencies table, as shown in the next figure, where a definition of cumulative percent is provided.

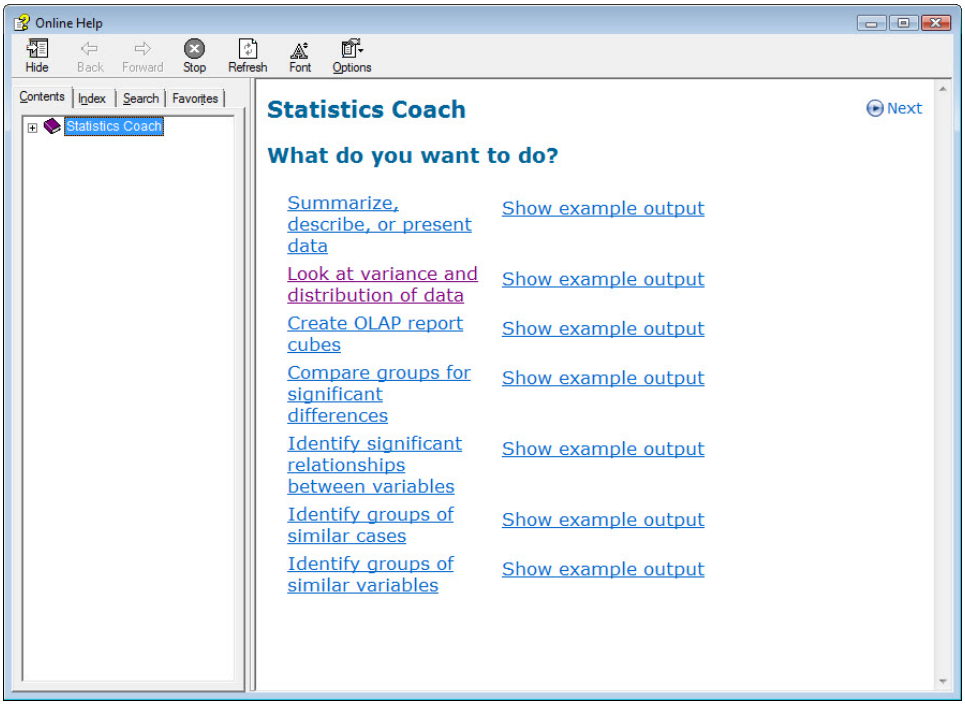
Figure A.5 Context What’s This? Help in Pivot Tables



Statistics Coach

The Statistics Coach can help to guide you through the process of finding the most appropriate procedure for the type of analysis you want to do with your data. The Statistics Coach provides a wizard-like method for suggesting a statistical or charting procedure based on your answers to a series of questions.

Figure A.6 Statistics Coach: First Question on Opening Dialog



Case Studies

Case studies provide comprehensive examples of each procedure. Data files used in the examples are installed with PASW Statistics in the Samples sub-folder of the PASW Statistics program folder, so you can follow along, performing the same analysis--from opening the data source and selecting variables for analysis to interpreting the results.

Note

As of PASW Statistics 17, the sample data files are available in several languages and are stored in sub-folders such as English, German, etc. of the Samples folder.

Figure A.7 Case Studies Table of Contents

A.5 *Lesson Summary*

Students who have completed this lesson should be able to:

- Use the Toolbars
- Use all types of help available in PASW Statistics

And,they should also be able to:

- Describe the operation of the toolbar buttons
- Describe the options for help in PASW Statistics
- Access the various types of help