SPSS Step-by-Step

Tutorial: Part 2

For SPSS Version 11.5

1 *Transformations and recoding revisited* **5**

Introduction **5**

Value labels **5**

SPSS Tutorial and Help **6**

Using online help ***6***

Using the Syntax Guide ***7***

Using the statistics coach ***8***

Moving around the output window **10**

Sorting Revisited: Sorting by multiple variables **11**

Utilities: variable and file information **12**

Utilities > Variables ***12***

Utilities > File Info ***13***

Data Transformations **14**

Computing new variables ***14***

Performing calculations with a variable and a function 14

Creating expressions with more than one variable 16

Conditional expressions ***18***

Creating subsets ***20***

Deeper into crosstabs **23**

Crosstab Statistics ***23***

Crosstab cells ***24***

Adding layers to crosstabs ***26***

When to include zeros in a mean **27**

Gender, geography, and exercise: the universal variables **28**

Summary **28**

2 *Statistical procedures* ***31***

Introduction **31**

Measuring association **31**

Bivariate correlations ***32***

Partial correlation ***35***

Multiple correlation (multiple regression) ***36***

Crosstabs ***37***

Measuring differences **39**

T-Tests ***39***

ANOVA ***42***

One-Way ANOVA 43

Summary **47**

1 Transformations and recoding revisited

Introduction

In the first session, we'll explored the SPSS interface, some elimentary data management and recodes, and some basic charting. In this second session, we'll explore work with more complex data transformations like combining variables and subsetting populations and work with some of the primary statistical functions. We'll also look more closely at the online help and tutorial provided by SPSS. . But first, a small clean-up task from last week: displaying and hiding value labels.

Value labels

1. Open SPSS.
2. Open the data file by selecting File > Open > Data and finding the file Employee data.sav in the folder named SPSSTutorialData.
3. Make sure you're in the Data View of any data file.
4. From the menu, select View > Value Labels. If Value Labels is checked, the value labels will be displayed for variables for which you have defined value labels. If it is not checked, the actual values will be displayed.

5. Select View from the menu again and make sure that Value Labels is checked. If it isn't, click it once to select it. You can turn value labels on or off at any time during an SPSS session.

SPSS Tutorial and Help

SPSS provides extensive assistance through its online help, tutorial, syntax guide, and statistics coach.

Using online help

1. From the menu, select Help > Topics.
2. In the keyword field, type:

   crosstabs

   Notice that SPSS begins matching topics as soon as you begin typing.
3. From the list below the keyword field, double-click **assumptions.**
4. From the keyword list, double-click **formats.**

   Notice that the resulting help window informs you that you can "arrange rows in ascending or descending order of the values of the row variable." What's wrong with this statement? Hint: In the rest of the computer world, "format" applies to how you display text or numbers. Arranging in ascending or descending order is called **sorting.** Lesson: If a standard already exists, use it; don't confuse people by making it up as you go along.
5. From the **Related topics** list, click once on **-Related procedures.** The help window now displays information about modeling relationships between two or more categorical variables.
6. From the **Related topics** list, click once on **Model Selection Loglinear Analysis Data Considerations**. The help window now displays more information about this technique.
7. In the keyword field, type:

   chi-square
8. From the keyword list, double-click **Chi-square test.**
9. In the next window, click **Display.** Notice that the help window now displays general information about the Chi-square test. In addition to the list of related topics, the window also contains a **Show Me** link.

10. Click **Show Me.** SPSS now opens the tutorial to the chi-square topic in the form of an Internet page.
11. Click **Next.** In addition to an example of how to use a chi-square test, the window also identifies the sample data file you can use to follow the example for yourself.
12. Click **Next.**
13. Read the text on the right side of the screen. Here is where the tutorial explains each step. And yes, that is a typo in “you must first be weight the cases …” Ignore the “be.”
14. Click **Next** and read the steps.
15. Click **Next** and read the steps.
16. Click **Next.**
17. Click **Next.** SPSS now displays the sample output.
18. Close the tutorial window.
19. Close the Help window.

As with most help systems, you can use links to investigate topics related to the keyword you selected. In SPSS, however, you can also open the online tutorial to get more information about using a specific procedure.

Using the Syntax Guide

If you need information about SPSS syntax, you can open the online Syntax Guide. This guide explains each command and provides examples of its use. The Syntax Guide is in Adobe Acrobat format. In this format, you can search the guide for specific text or use the Bookmarks pane to find a specific command.

1. From the menu, select Help > Syntax Guide > Base.
2. Look at the bottom of the window where Acrobat lists the page count. Yes, that is 1490 pages. In other words, if you want, you can print the entire Syntax Guide.
3. In the bookmarks pane, click the “+” to the left of UNIVERSALS. The topic opens to display its subtopics.
4. Click the “+” to the left of Commands. (The Commands under Universals, that is, not the Commands further down the list.)
5. Click Syntax Diagrams. This topic provides information about the basic structure of SPSS syntax.

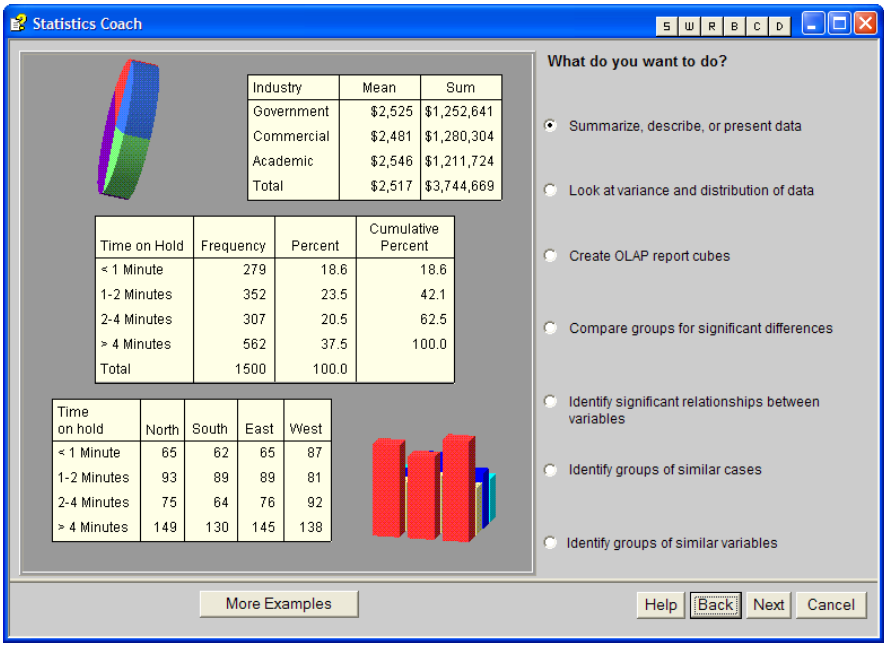
6. Now click the "+" to the left of the top-level COMMANDS topic. The window opens to display the subtopics for COMMANDS.
7. Scroll down until you can see the CROSSTABS entry.
8. Click the "+" to the left of CROSSTABS.
9. Click the word **CROSSTABS.** The Crosstabs page is now displayed and provides information about the complete format of the Crosstabs command.
10. Close the Syntax Guide window.

Using the statistics coach

One of the most useful functions in SPSS is the statistics coach, particularly when you're just starting to work with the program. The statistics coach provides prompts at which you can select what you want to do, the kind of data you're using, and the kind of output you want.

1. From the menu, select Help > Statistics Coach. SPSS opens the first Statistics Coach window (Figure 1).

FIGURE 1. Statistics Coach, opening window



Industry	Mean	Sum
Government	$2,525	$1,252,641
Commercial	$2,481	$1,280,304
Academic	$2,546	$1,211,724
Total	$2,517	$3,744,669

Time on Hold	Frequency	Percent	Cumulative Percent
< 1 Minute	279	18.6	18.6
1-2 Minutes	352	23.5	42.1
2-4 Minutes	307	20.5	62.5
> 4 Minutes	562	37.5	100.0
Total	1500	100.0	

Time on hold	North	South	East	West
< 1 Minute	65	62	65	87
1-2 Minutes	93	89	89	81
2-4 Minutes	75	64	76	92
> 4 Minutes	149	130	145	138

What do you want to do?

● Summarize, describe, or present data

○ Look at variance and distribution of data

○ Create OLAP report cubes

○ Compare groups for significant differences

○ Identify significant relationships between variables

○ Identify groups of similar cases

○ Identify groups of similar variables

2. Take a moment to review the choices offered in this window.
3. Click **More Examples** a few times and notice that different types output available to you.
4. For the moment, we'll use the default task, **Summarize, describe, or present data**, so click **Next.**
5. In this case, we want to create a summary of gender by job category. Both variables are categorical, so click **Next.**
6. This time we'll change the output from the default, so select **Charts and graphs** by clicking its radio button.
7. We want a simple two-dimensional chart, so click **Next.**
8. Click **Next.**
9. We want a bar chart, so click **Finish.**
10. SPSS now opens the correct window for creating a bar chart with the type of data we have selected.
11. Drag Employment Category to the horizontal axis.
12. Drag Gender to the Legend Variables > Color field.
13. Close the remaining help window.
14. Click **OK.** The chart appears in the output window. Next, we'll use the statistics coach for a more complicated task.
15. From the menu, select Window > Employee Data.sav - SPSS Editor to get back to the Data window.
16. From the menu select Help > Statistics Coach.
17. Select **Compare groups for significant differences.**
18. Click **Next.**
19. We're using categorical data, so click **Finish.** The How To window appears to guide you through the steps along with the Crosstabs window. (You might have to move them around a bit so you can see both windows.
20. Select Gender and move it to the Rows pane.
21. Select Employment Category and move it to the Columns pane.
22. Click **Statistics.**
23. In the Crosstabs: Statistics window, select Chi-square.
24. Click **Continue.**
25. In the How To window, click **Tell Me More.** SPSS now displays the Data Requirements window that applies to using Chi-square.
26. Click **Next.** Surprise! There is no next topic. Ha, ha, SPSS. Very funny.

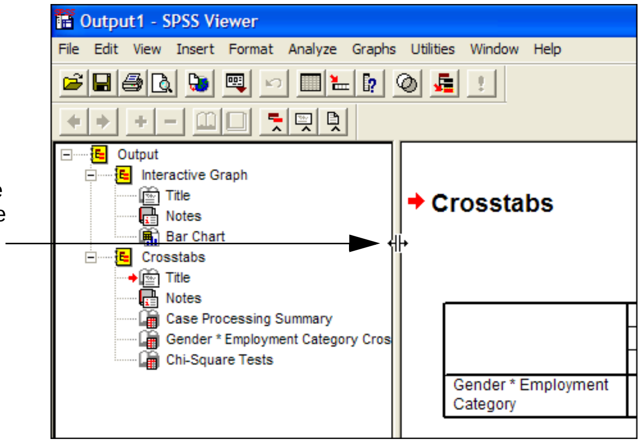
27. Click **OK.**
28. Click **Back.** Oh look! There is no **Back.**
29. Close the Data Requirements window.
30. In the Crosstabs window, click **OK.** The output is now displayed in the Output window.
31. Close any remaining Help windows.

Moving around the output window

Now that you have created some charts, crosstabs, and statistics results, it's a good time to take a closer look at moving around the output window.

1. From the menu, select Window > Output1 - SPSS Viewer.
2. Move your cursor over the border between the panels, hold down the mouse button, and move the border to the right until you can read all the titles in the icons in the contents pane as in Figure 2.

FIGURE 2. Changing pane width in the output window



hold and drag here to change the pane size in the output window

3. Notice the icons on the left arranged in outline format.
4. Click the icon named Interactive Graph. The output displays moves to the first graph you created in this session.

5. Notice the "-" to the left of the Output icon. The "-" indicates that a topic is fully expanded.
6. Click the "-" next to Output. The "-" changes to a "+" and all the output is now hidden. If you ever "lose" output on the window, check to see if the output is hidden.
7. Click "+" next to Output to expand the items again.
8. Click the icon named "Crosstabs."
9. Holding down the left mouse button, drag "Crosstabs" up above "Interactive Chart." You can use the drag function to arrange your output in any order you like.
10. Below Crosstabs, click the icon for Title.
11. Click the icon for Chi-square tests. Notice that the red arrow next to the icon corresponds to the red arrow in the actual output window.
12. Under Interactive Graph, click Bar Chart.
13. Above the actual chart, double-click the title ("Interactive Graph").
14. Select all the text and type:

    Distribution of job category by gender

15. Click anywhere outside the title to apply the change.
16. In the navigation pane, click Title under Interactive Graph.
17. Wait about one second, then click Title again to activate the text.
18. With the text selected, type:

    Distribution of job category by gender

19. Press **Enter** or click anywhere outside the title to apply the change.

    Use the title function in the navigation pane to indicate what each piece of output it contains. "Distribution of job category by gender" is a lot more informative than "Title."

Sorting Revisited: Sorting by multiple variables

SPSS provides sophisticated sorting functions. You can sort by multiple variables, and you can set the sort order for each variable. For example, you could sort in order of increasing income, decreasing birthdate, and increasing expenditures. In the following task, you'll sort the employee data set by gender (increasing) and current salary (decreasing).

1. Switch back to the data view by selecting Window > Employee data.sav - SPSS Data Editor.
2. From the menu, select Data > Sort Cases.
3. Clear any criteria that might already be in the **Sort by** pane by double-clicking them.
4. Double-click gender and current salary to move them to the Sort by pane.
5. Click once on **gender**.
6. If it is not already selected, select **Ascending**.
7. Click once on **current salary**.
8. Click **Descending**.
9. Click **OK.** Notice that all female employees are now listed first, in descending salary order.

Utilities: variable and file information

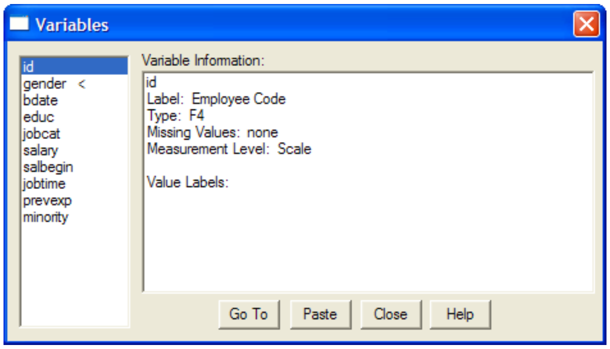
Tucked quietly under the Utilities menu are two especially useful functions: Variables and File Info. You can use these functions to get a snapshot of each variable in the file (Variables) and all variables together (File Info).

Utilities > Variables

The Variables function provides all the information about each variable in your data file, including any categorical codes and their value labels.

1. From the menu, select Utilities > Variables. (Figure 3)

FIGURE 3. Variables window



```
id
gender <
bdate
educ
jobcat
salary
salbegin
jobtime
prevexp
minority
```

```
id
Label: Employee Code
Type: F4
Missing Values: none
Measurement Level: Scale

Value Labels:
```

2. In the variable list, click **jobcat.** Notice that the Variable Information pane displays the variable name, label, defined missing value, measurement level, values, and value labels.
3. Click **salary.** This value is a scale variable (continuous) and so has no value labels.
4. The Go To button takes you to the specific variable within a selected case or to the variable in the first case if no case is selected.
5. Click **Close.**

Utilities > File Info

The File Info window provides information about all variables in the file. This is extremely useful information. We recommend that you print out the file definition regularly and keep it close at hand.

1. From the menu, select Utilities > File Info.
2. Scroll through the output to see how each variable is described. Because this information goes to the output window, be sure that you print only the File Info output.
3. In the navigation pane, click once on the File Information icon.
4. Press **Ctrl-P** to open a print dialog window. Notice that under Print Range you can select *All Visible Output* or *Selection.* When you print the File Info, be sure to select Selection.
5. Click **Cancel.**

Data Transformations

SPSS provides a number of funtions you can use in computing new variables, including:

- Tarithmetic funtions
- statistical functions
- string functions
- date and time functions
- distribution functions
- random variable functions
- missing value functions

In this session, we'll be looking at only the arithmetic functions.

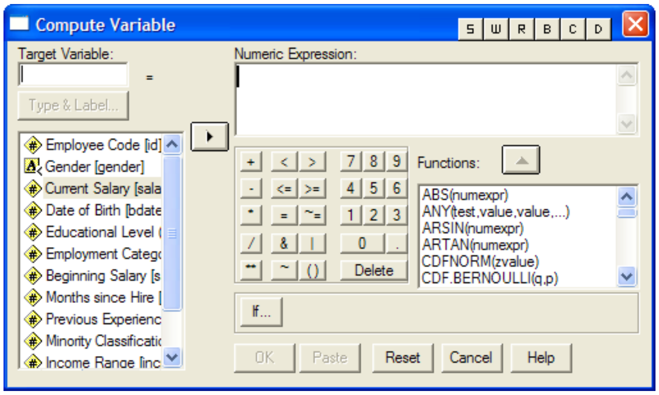
Computing new variables

Performing calculations with a variable and a function

In some cases, you might want to calculate new variables based on values in existing variables and some arithmetic function like multiplying or dividing. For example, if you have a variable that contains an annual salary, you might want to calcuate a monthly salary. To create the new variable, you use the **Compute** function.

1. In the Data window, select from the menu Transform > Compute. (Figure 4)

FIGURE 4. Compute window

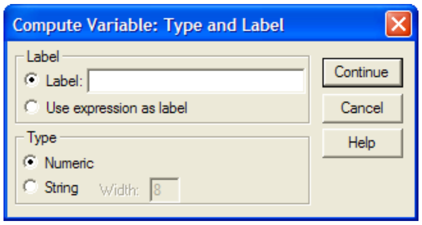


2. In the Target Variable field, type:

salmonth

3. Click **Type & Label.** (Figure 5)

FIGURE 5. Compute: Type and label window



4. In the Label field, type:

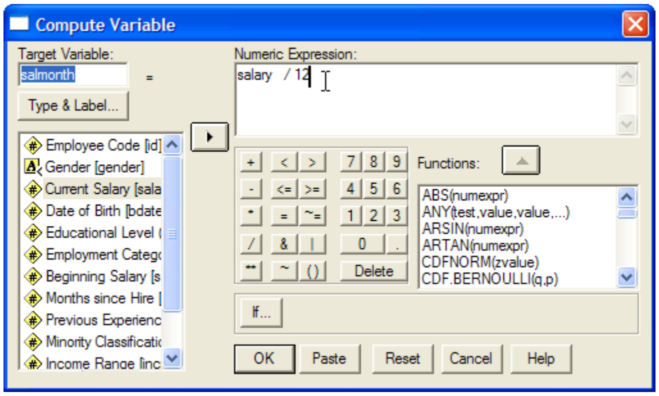
Average monthly salary

5. Click **Continue.**
6. In the Compute Variable window, select Current Salary and move it to the Numeric Expression pane by clicking the right arrow.
7. In the Numeric Expression pane, click the cursor after **salary** and type:

/12

Your window should now look likeFigure 6

FIGURE 6. Entering a compute formula



Target Variable: salmonth

Numeric Expression: salary / 12

8. Click **OK.** The Compute Variable window closes and the new variable is displayed in the Data window. You can now use the new variable in procedures such as crosstabs or in further calculations. For example, you could create a new variable for monthly withholding that calculates withholding as a percentage of monthly salary. You could then subtrack the new withholding variable from the monthly salary to create still another variable for monthly net.

Creating expressions with more than one variable

Let's use the previous example of calculating withholding and net to compute variables based on more than one variable. First you'll compute the withholding variable, then you'll compute the net variable.

1. From the menu, select Transform > Compute.
2. In the Target Variable field, type:

withhold

3. Click **Type & Label.**
4. In the Label field, type:

Monthly withholding

5. Click **Continue.**
6. Select all the text in the Numeric Expression field and delete it.

7. Move the new variable, Average Monthly Salary, to the Numeric Expression field.
8. Click after **salmonth** and type:

*** .05**

Note: If you haven't worked with computer programs before to make calculations, the asterisk denotes multiplication. A double asterisk (**) denotes exponentiation. In SPSS, a vertical bar (|) denotes "OR", and the ampersand (&) denotes "AND".

9. Click **OK.** The new variable appears in the data view. In the next step, you'll use two variables to calculate a third.
10. From the menu, select Transform > Compute.
11. In the Target Variable field, type:

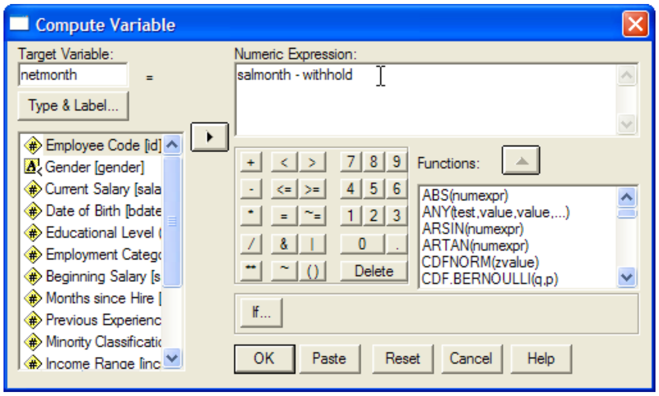
netmonth

12. Click **Type & Label.**
13. In the Label field, type:

Monthly net

14. Click **Continue.**
15. Select all the text in the Numeric Expression field and delete it.
16. From the list of variables, select Average Monthly Salary and move it to the Numeric Expression field.
17. Click after salmonth in the Numeric Expression field.
18. Using the keypad in the Compute Variable window, click **"-".**
19. From the list of variables, select the new variable Monthly Withholding.
20. Click the right arrow to move it to the Numeric Expression pane. Your Compute Variable window should now look like Figure 7.

FIGURE 7. Complete Compute Variable window for monthly net



Target Variable: netmonth =

Numeric Expression: salmonth - withhold

21. Click **OK.** The new variable appears in the data view.

Conditional expressions

In some cases, you might want to look at only a specific subset of your data. Say you want to send a monthly newsletter to only female clerical staff. To identify these staff, you'll calculate a new binary variable (one that has only two values) using the IF statement to set the condition.

1. From the menu, select Transform > Compute.
2. In the Target Variable field type:

femclerk

3. Click **Type & Label.**
4. In the Label field type:

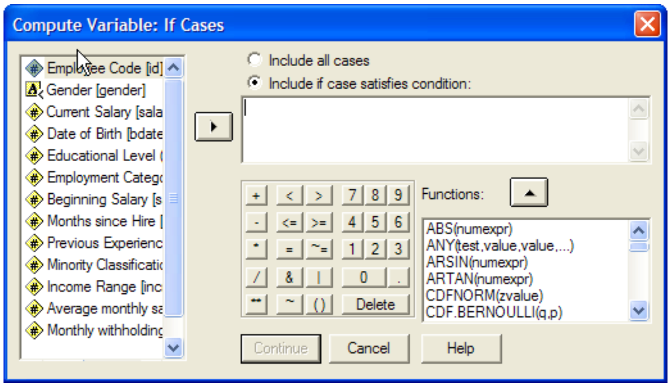
Female Clerical

5. Click **Continue.**
6. Select all the text in the Numeric Expression field and delete it.
7. In the Numeric Expression field type:

1

8. Click **If** to open the Compute Variable: If Cases window (Figure 8).

FIGURE 8. Compute Variable: If Cases window



Compute Variable: If Cases

○ Include all cases

● Include if case satisfies condition:

9. Select **Include if case satistifes condition.**
10. Double-click **Gender** to move it to conditions field.
11. Click after Gender in the conditions field and type:

= "f"

Note: Whenever you create a condition, you must use the actual values in the variable, not their labels. Thus, setting a condition to **gender = "Female"** would not select any cases.

12. Click after "f" and type a space.
13. Using the keypad in the Compute Variable window, click **&.** You use the ampersand to add a second condition.
14. From the field list, double-click **Employment category** to move it to the calculation pane.
15. In the calculation pane, type:

= 1

Note that you don't use quotation marks this time because is a numeric variable.

16. Click **Continue.**

17. Click **OK.** The new variable appears in the Data window. Scroll through the records to see how the values in the new variable. Notice that cases where gender is not female and job category is not manager have only a period, indicating a missing value. Only those cases where gender is female **and** jobcat is manager contain a 1 in the new variable.

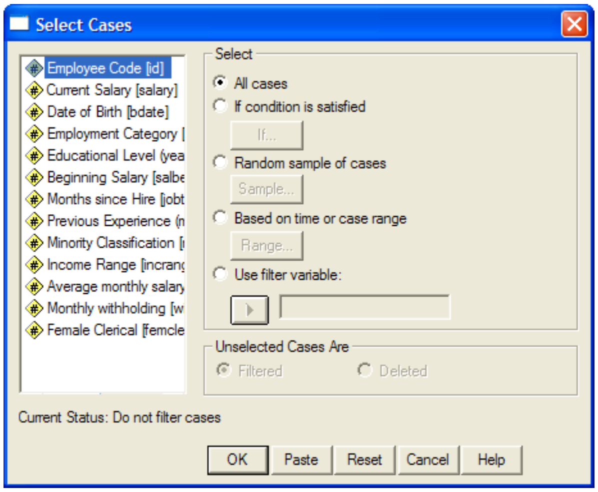
Creating subsets

In some instances, you might want to use only part of the file in an analysis. For example, you might want to look at changes in income among single working mothers. Or you might want to consider only staff born before a specific date.

To select a *subset* of the cases in your file,

1. From the menu, select Data > Select Cases (Figure 9).

FIGURE 9. Select cases window



Current Status: Do not filter cases

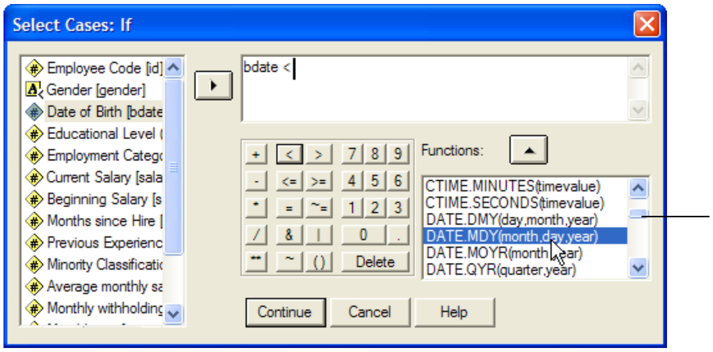
2. Select **If condition is satisfied** by clicking its radio button.
3. Click **If....** Notice that the **Select Cases: If** window looks exactly like the If window you used in the earlier compute procedures.
4. From the variable list, double-click Date of Birth.
5. Click the cursor anywhere after **bdate** in the calculation pane.

6. Type (or select from the keypad):

<

7. Scroll through the Function menu and double-click DATE.MDY(month,day,year). (Figure 10)

FIGURE 10. Selecting a function



bdate <

scroll here to select a function

In the next step, you'll set the date criterion. SPSS adds the function to the calculation pane, substituting question marks to indicate that you need to specify the values.

8. Select the first question mark and type:

1

9. Select the second question mark and type:

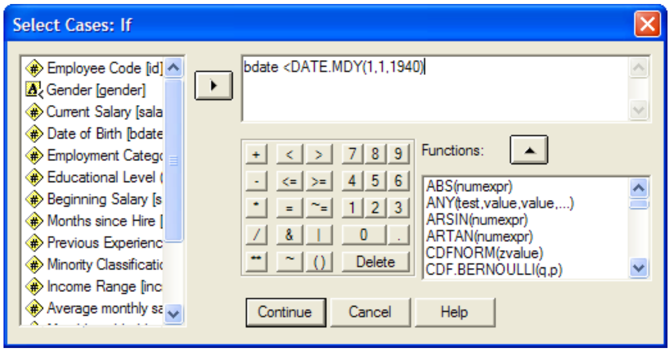
1

10. Select the third question mark and type:

1940

Your completed window should look like Figure 11.

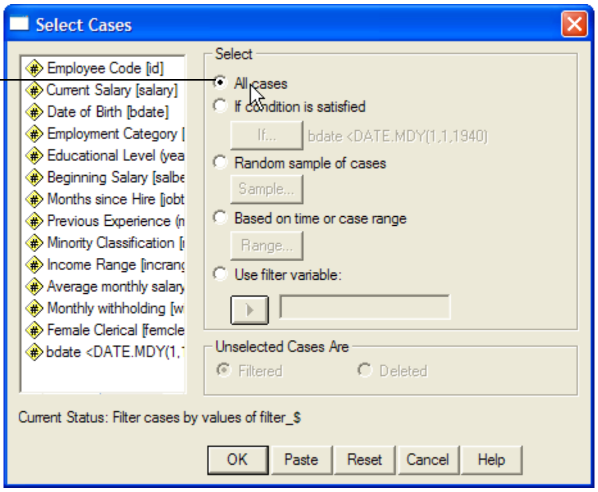
FIGURE 11. Completed Select Cases: If window



bdate <DATE.MDY(1,1,1940)

11. Click **Continue.**
12. Click **OK.** Notice that many of the records are marked with a diagonal line through the record number. These cases are excluded from any further calculations until you specifically include them again.
13. To see the effect of the subset selection, right click the heading for **bdate.**
14. From the pop-up menu, select Sort Ascending. Notice that all employees born before 1940 are selected, except for the person with the missing date of birth. In the next step, you'll instruct SPSS to include all cases until otherwise instructed.
15. From the menu select Data > Select Cases.
16. Select All Cases by clicking its radio button (Figure 12).

FIGURE 12. Selecting all cases



Current Status: Filter cases by values of filter_$

click here to include all cases

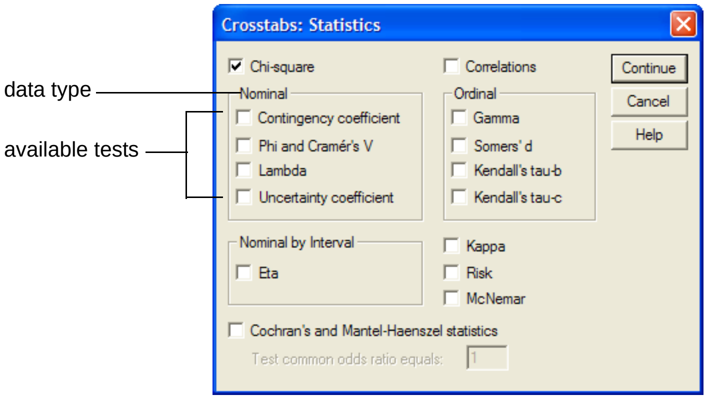
17. Click **OK.** The diagonal lines appear to be gone, but to be sure, right-click the heading of **bdate** again and select Sort Descending, so that the youngest employees are listed first. Notice that they are no longer excluded.

Deeper into crosstabs

Crosstab Statistics

When you the various statistical techniques, SPSS will frequently tell you what kind of statistical tests are available for that procedure. For example, if you ask for crosstabs, SPSS offers a number of statistics based on the type of data you're using. (Figure 13)

FIGURE 13. Statistics indicated by data type in Crosstabs: Statistics window



Crosstabs: Statistics

☑ Chi-square

☐ Correlations

Nominal
☐ Contingency coefficient
☐ Phi and Cramér's V
☐ Lambda
☐ Uncertainty coefficient

Ordinal
☐ Gamma
☐ Somers' d
☐ Kendall's tau-b
☐ Kendall's tau-c

Nominal by Interval
☐ Eta

☐ Kappa
☐ Risk
☐ McNemar

☐ Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1

Continue
Cancel
Help

For example, when you select a statistic like Chi-square, SPSS indiates the particular Chi-square *technique* that should be used based on the type of data. If you are using nominal or ordinal data, SPSS provides a number of methods you can include in your output. To see how SPSS makes the selection of techniques and to see the description of the data types and corresponding statistics, try the following.

1. From the menu, select Help > Topics.
2. Click the Index tab and enter:

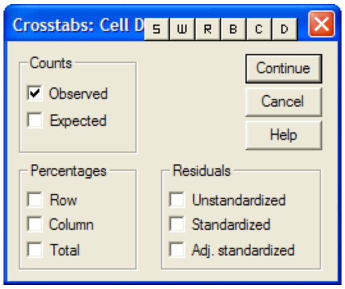
crosstabs

3. From the list below your entry, select **Statistics.** SPSS Help displays a description of the types of statistics to select based on the type of data you're using.
4. Close the Help window.

Crosstab cells

Using the Crosstab cell display window (Figure 14) you can determine what data will be displayed in each cell of the crosstab.

FIGURE 14. Crosstab cell display



Counts
☑ Observed
□ Expected

Percentages
□ Row
□ Column
□ Total

Residuals
□ Unstandardized
□ Standardized
□ Adj. standardized

Try it:

1. From the menu, select Analyze > Descriptive Statistics > Crosstabs.
2. Move **Gender** to the Rows list.
3. Move **Employment Category** to the columns list.
4. Click **Cells.**
5. Under **Percentages**, select Row, Column, and Total.
6. Click **Continue.**
7. Click **OK.** SPSS displays the completed crosstab in the output window.

Gender * Employment Category Crosstabulation

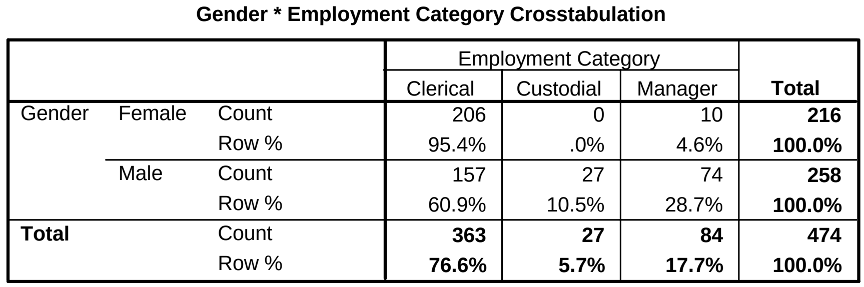
			Employment Category			
			Clerical	Custodial	Manager	**Total**
Gender	Female	Count	206	0	10	**216**
		Row %	95.4%	.0%	4.6%	**100.0%**
		Column %	56.7%	.0%	11.9%	**45.6%**
		Total %	**43.5%**	**.0%**	**2.1%**	**45.6%**
	Male	Count	157	27	74	**258**
		Row %	60.9%	10.5%	28.7%	**100.0%**
		Column %	43.3%	100.0%	88.1%	**54.4%**
		Total %	**33.1%**	**5.7%**	**15.6%**	**54.4%**
Total		Count	**363**	**27**	**84**	**474**
		Row %	**76.6%**	**5.7%**	**17.7%**	**100.0%**
		Column %	**100.0%**	**100.0%**	**100.0%**	**100.0%**
		Total %	**76.6%**	**5.7%**	**17.7%**	**100.0%**

Notice that each cell contains the count, the percent of the row, the percent of the column, and the percent of the total. Crosstabs like this are useful for both a general overview and closer study of your data. For publication, however, you may want to simplify the output by including only a column or row percentage, depending on the issue you’re addressing.

Suppose you only want to know the distribution of job categories within gender. In the next task, you’ll create a cross-tab that includes only *row* percentages.

8. From the menu, select Analyze > Descriptive Statistics > Crosstabs.
9. Move Gender to the Rows pane.
10. Move Employment to the Columns pane.
11. Click **Cells.**
12. Under percentages, make sure only Row is selected. Clear any others that are already selected.
13. Click **Continue.**
14. Click **OK.** Your new output will look like Figure 15. If you wanted to know percentages of each job category across gender, you would select only column percentages.

FIGURE 15. Crosstab showing only row percentages



Gender * Employment Category Crosstabulation

			Employment Category			**Total**
			Clerical	Custodial	Manager	
Gender	Female	Count	206	0	10	**216**
		Row %	95.4%	.0%	4.6%	**100.0%**
	Male	Count	157	27	74	**258**
		Row %	60.9%	10.5%	28.7%	**100.0%**
Total		Count	**363**	**27**	**84**	**474**
		Row %	**76.6%**	**5.7%**	**17.7%**	**100.0%**

Adding layers to crosstabs

So far, we have worked with just a single variable for rows. You can make your crosstabs much more specific, however, by adding multiple row variables or by adding layers. In this exercise, you’ll create layered crosstabs that provide a more detailed breakdown of the data.

1. From the menu, select Analyze > Descriptive Statistics > Crosstabs. Notice that the variables from the previous crosstab are still selected.

2. From the variable list, move Minority Classification to the Layer pane.
3. Click **OK.** Notice that you're still getting row percentages only because we did not reset the cells information. Notice also that the Layers variable becomes the uppermost level, followed by gender. In the next task, you'll reverse that order.
4. From the menu, select Analyze > Descriptive Statistics > Crosstabs. Notice that the variables from the previous crosstab are still selected.
5. In the Rows pane, double-click Gender to move it back to the variable list.
6. In the Layers pane, double-click Minority Classification to move it back to the variable list.
7. Move Minority Classification to the Rows pane.
8. Move Gender to the Layers pane.
9. Click Cells.
10. Under percentages, select Row, Column, and Total.
11. Click **Continue.**
12. Click **OK.** In the new output, the crosstab displays Minority Classification within Gender.

You can continue to add layers and multiple rows to your crosstabs. Remember Robin's rule, however: ALWAYS GET THE UNIVERSE FIRST. That is, print out the most general crosstabs before getting into the detail.

When to include zeros in a mean

We were once working on a project looking into some of the economic issues of child care, and one of the researchers asked if we should include in the calculation of the average price the number of people who didn't use child care. The answer is: it depends. If you want to look at, say, an average price *paid*, then, no, you would not include people who didn't buy the product. If, on the other hand, you want to know the *per capita* cost, then, yes, you would include everyone.

The question involves the issue of whether to include zero values in calculating statistics such as means or standard deviations or conducting statistical tests such as t-tests and ANOVAs. And to some degree the answer lies in the question itself. If you say, "How much did people pay for child care?" (or gasoline or televisions or clothes) then you want to look at the actual purchase price among those who actually purchased the product. If, for example, I tell you that the average *per capita* cost of gasoline is thirty cents a gallon, that's not going to tell you what to expect the next time you drive up to the pump. What you really want to know is, what's the average price today in this particular area. If, on the other hand, you're working for the Council of Economic Advisors and you want to know how the cost of gasoline factors into the overall expenditures of an average family, you do want to use a *per capita* ***cost.***

The other question to consider is whether zero is a valid value in your data set. In medical research, for example, particularly in dose-response research or lab values, zero is obviously a valid value. In other cases, such as a five-point Likert scale beginning with 1, zero is not a valid value and should be treated as missing on an error. In other words, the decision about whether to include zero in a particular test depends on whether it's a valid value and on the particular question you're asking.

Gender, geography, and exercise: the universal variables

There are certain variables that will affect nearly any statistic or test you use. Our particular favorites are gender, geography, and exercise. These are variables whose effect is so pervasive that failing to take them into account can seriously affect the validity of your research. There are others, of course, that will affect whatever data you work with to varying degrees. Age, of course, is certainly one, along with diet and ethnicity. If you're conducting medical research, for example, you must always take into account age, gender, and ethnicity. More and more, however, the level of exercise is being included as a concomitant variable'. We once had a research psychologist tell us that "Exercise is implicated in every variable we look at." In social science research, age, gender, ethnicity, and education are critical factors. The moral: when you are designing a research project, make sure you have accounted for *all* the variables that might affect the outcomes, not just the ones of immediate interest.

Summary

If you think we have spent an awful lot of time dealing with data management and crosstabs, you're right. You'll spend about eighty percent (a very rough guess) of your time on these two tasks. Then, finally, when you have the data exactly the way you need it, you run a couple of quick statistics and then . . . you start all over again. Remember that data analysis is iterative, so get ready now to do the same types of tasks over and over and over.

And over.

2 Statistical procedures

Introduction

This tutorial is not a replacement for a course in basic statistics or for a textbook on that subject. The primary emphasis here is on the use of SPSS to explore data and to answer some of the statistical questions you might ask about your data.

In this chapter we will review some of the SPSS procedures for evaluating the association among two or more variables, and procedures for measuring differences among groups.

Measuring association

Typically the association between two variables is evaluated by using a bivariate correlation procedure. If the two variables are continuous and you want to predict one variable using the value of the other, a simple linear regression or some method of curve estimation can be used.

If there are more than two variables and they are continuous, use a partial or a multiple correlation procedure. If you want to predict one of the variables using the values of the other variables, a multiple regression can be used.

If you have frequency distributions based upon one or more categorical variables, you should consider crosstabulation or Chi-square. A categorical variable could be one that naturally exists, such as eye color, or could be derived by “categorizing” a continuous variable.

Bivariate correlations

Bivariate correlations measure the degree of association between two variables. If the two variables are continuous, the Pearson product moment correlation is an appropriate measure. If they are not continuous (that is, if they are discrete or categorical), it would be more appropriate to use Spearman’s rho or Kendall’s tau-*b*.

The correlation coefficient, which ranges from -1 to +1 is both a measure of the strength of the relationship and the direction of the relationship. A correlation coefficient of 1 describes a perfect relationship in which every change of +1 in one variable is associated with a change of +1 in the other variable. A correlation of -1 describes a perfect relationship in which every change of +1 in one variable is associated with a change of -1 in the other variable. A correlation of 0 describes a situation in which a change in one variable is not associated with any particular change in the other variable. In other words, knowing the value of one of the variables gives you no information about the value of the other.

The correlation squared is another measure of the strength of the relationship. In fact, *the correlation squared is the percent of the variance in the dependent variable that is accounted for or predicted by the independent variable.*

You can also determine the statistical significance of the correlation coefficient. If the direction of the association is hypothesized in advance, you can use a one-tailed test to determine whether the correlation is statistically significantly different from zero, otherwise use a two-tailed test.

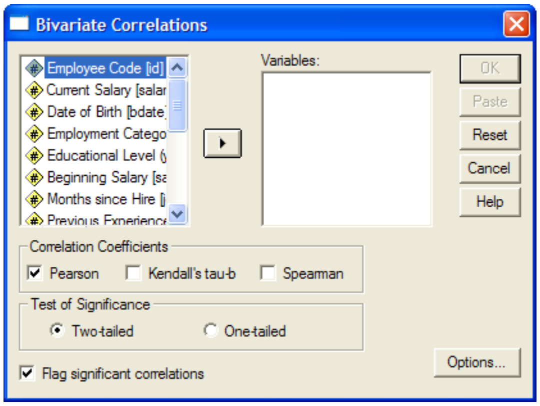
Correlation is not causation. If we looked at the correlation of the time the paper boy delivers the morning paper and the time of the sunrise, we would find a very strong positive correlation. And yet, we would be reluctant to claim that the newspaper boy causes the sun to rise.

How to: In this exercise, you’ll calculate a Pearson correlation between current salary and months since employment,

1. In the data window, open the file named Employee data.sav in the folder named SPSSTutorialData.

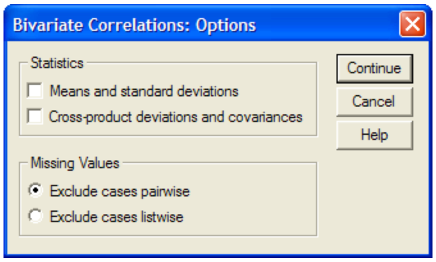
2. In the data window, from the menu select Analyze > Correlate > Bivariate (Figure 16).

FIGURE 16. Bivariate correlations window



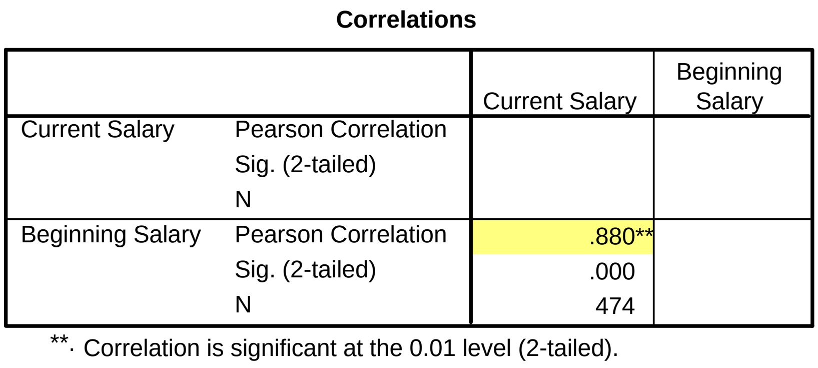
3. Double-click **Current Salary** to move it to the Variables list.
4. Double-click **Beginning Salary** to move it to the Variables list.
5. Click **Options.** (Figure 17)

FIGURE 17. Bivariate correlations: options window


Bivariate Correlations: Options
Statistics
□ Means and standard deviations
□ Cross-product deviations and covariances
Missing Values
● Exclude cases pairwise
○ Exclude cases listwise
Continue
Cancel
Help

6. In the Statistics pane, select **Means and standard deviations** by clicking its check box.
7. Click **Continue.**
8. Click **OK.** The output is displayed in the Output window. (Figure 18)

FIGURE 18. Pearson correlation for current salary x beginning salary



Correlations

		Current Salary	Beginning Salary
Current Salary	Pearson Correlation		
	Sig. (2-tailed)		
	N		
Beginning Salary	Pearson Correlation	.880**	
	Sig. (2-tailed)	.000	
	N	474	

**. Correlation is significant at the 0.01 level (2-tailed).

9. Notice that the correlation is particularly high (.880). The footnote to the table indiates that correlation is significant at the .01 level. (And, no, we can't find any documentation on why SPSS has highlighted the particularly high correlation. Just another one of those moments of cryptic helpfulness.)

10. On the contents pane of the output window, click the Correlations icon indicated by the red arrow (you may have to scroll down a bit to see it), then click it again to allow you to change the name.

11. Type:

Pearson: Curr Sal X Beg Sal

12. Press **Enter** or click anywhere away from the title to apply the change. Try to get in the practice of naming each portion of output so that you can quickly find what you're looking for.

Homework: The SPSS tutorial has an excellent section on using correlations. To find it, open the Help menu and enter **correlations, bivariate** as the search term. In the Help window, click **Show Me.**

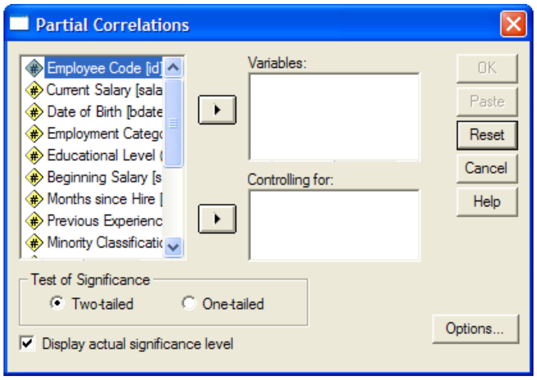
Partial correlation

Partial correlation is used to measure the association of two continuous variables after controlling for the association of other variables. Conceptually, what is being done is to first calculate the variance in the dependent variable that can be explained or accounted for by all of the *control* variables. The variance accounted for by the control variables is then removed from the dependent variable. Finally, the degree of association is measured between the variance remaining in the dependent variable and the non-controlled variable.

How to: In this exercise, you’ll compare current salary to beginning salary, after controlling for previous experience.

1. From the menu, select Analyze > Correlate > Partial. (Figure 19).

FIGURE 19. Partial correlations window



Test of Significance
● Two-tailed ○ One-tailed

☑ Display actual significance level

2. Notice that this time, in addition to selecting the variables to be compared, you can also select **Controlling for.**
3. Select **Current Salary** and **Beginning Salary** and move them to the **Variables** pane.
4. Select **Previous Experience** and move it to the **Controlling For** pane.
5. Click **Options** to select the statistics you want displayed.
6. Select **Means and standard deviations** by clicking its check box.
7. Click **Continue.**
8. Click **OK.** The new output is displayed in the output window.

9. In the contents pane of the output window, click Partial Corr, then click it again to activate it. At the end of the text, type:

:Curr Sal X Beg Sal CF Prev Exp.

10. Click anywhere away from the title to apply the change.

Note: During your analysis, you're likely to generate a great number of charts, tables, and other output. Try to come up with a consistent abbreviation scheme that will help you figure out at a glance what you're looking at. In the example above, the X stands for "by" and "CF" stands for "Controlling For."

Multiple correlation (multiple regression)

Multiple correlation looks at the association between one continuous variable (often called the dependent variable) with a group of two or more continuous variables (usually called predictors).

One use for a multiple correlation is to find out if there is a relationship between an independent variable and a dependent variable *after controlling* for a subset of all other variables. In this sense the multiple correlation or multiple regression is used as a more sophisticated method of exploring partial correlations.

When you run a step-wise multiple regression, SPSS will find the one variable in the group of predictors which has the highest correlation with the dependent variable. It will then statistically remove that variance from the dependent variable that the predictor variable accounts for. The procedure will then go to the list of remaining predictors and select the variable which has the highest correlation with the remaining variance in the dependent variable, remove *that* variance, then select the next predictor and so on until some criterion is met. Typical criteria that you can specify are the amount of additional variance accounted, the level of statistical significance for the change in variance accounted for, and the maximum number of predictors that can be selected.

Example: In a study of the effectiveness of entitlement programs, you want to find out which set of variables can best predict client's income once they are no longer receiving benefits. All entitlement data are quantitative, including time receiving benefits, individual benefit values, length of job training, and family size. A single categorical variable — minority/non-minority — is included in the calculations as a binary variable. In this example, post-eligibility income is the dependent variable, while the independent (or predictor) variables include value of benefits, length of eligibility, and the like.

Crosstabs

To quote the famous purple book, "A crosstabulation is a joint frequency distribution of cases according to two or more classificatory variables. The display of the distribution of cases by their position on two or more variables is the chief component of contingency table analysis *and is indeed the most commonly used analytic method in the social sciences.*"[1] [Emphasis added.]

The Chi-square test can be used to determine whether the frequency distributions of one or more categorical variables are statistically independent. The crosstab can be used to provide measures of the associations of categorical variables. Some of the measures of association are the contingency coefficient, phi, tau, gamma, etc.. These measures describe the degree to which the values of one variable predict or vary with those of another.

Data requirements: Crosstabs require categorical data or continuous data recoded into categories, such as income or age ranges. The frequencies for each variable in the population should be approximately normal.

Suppose we had a group of 300 people for whom we knew hair color and eye color. We could create a contingency table similar to the displayed in Table 1. In the table, it appears that non-blue eyes and brown hair tend to go together, but because the total number of people is different in each cell, it's hard to arrive at a valid conclusion.[2]

TABLE 1. Sample crosstab

	Hair color:		
Eye color	**Blond**	**Brown**	**Total**
Blue	75	40	115
Non-blue	25	160	185
Total	100	200	300

[1] *SPSS: Statistical Package for the Social Sciences,* Second Edition, by Norman H. Nie, et al., McGraw Hill 1975, p. 218.

[2] This example is based on the example given in the Purple book, ibid., p. 219.

To help us determine if the counts of people in these four cells are random, we can calculate the number of people we would expect in each cell if the distribution of people with a particular eye color were independent of the distribution of people with a particular hair color.

The expected frequency for any cell in a contingency table is the product of the row frequency times the column frequency for that cell, divided by the total for the table. Thus, the expected frequency for blond hair/blue eyes would be 115 X 100/ 300 or 38.3 people. In fact, the observed frequency for this group is 75, indicating that the distributions are not independent. Something is going on here.

TABLE 2. Observed/Expected Crosstab values

	Hair color:		
Eye color	**Blond**	**Brown**	**Total**
Blue			
Observed	75	40	115
Expected	38.8	76.7	
Non-blue			
Observed	25	160	185
Expected	61.7	123.3	
Total	100	200	300

The degrees of freedom for a contingency table is the number of columns -1 times the number of rows -1. Conceptually, the degrees of freedom is a count of how many cells in which you are free to enter any number you want given that you know the margins. In general it indicates the number of data points you can specify before all remaining data points are determined.

To test whether the observed frequencies are statistically different from the expected frequency, we calculate a Chi-square: For each cell subtract the expected frequency (count) from the observed frequency, square the difference, and divide by the expected value for that cell, then sum these values for all the cells.

The greater the discrepancy between the expected and actual frequencies, the larger the Chi-square becomes. Whether there will be a statistically significant difference depends on the size of the difference and the degrees of freedom. *As the degrees of freedom increase, the larger the value of Chi-square needed to be statistically significant.*

How to: In this exercise, you’ll use a data file for hair color and eye color that will generate the Chi-square table shown above.

1. From the menu, select File > Open > Data.
2. Navigate to the file named ChiSquare.sav and open it.
3. When prompted to save the current data, click **Yes.**

4. From the menu, select Analyze > Descriptive Statistics > Crosstabs.
5. Select Hair Color and move it to the Columns pane.
6. Select Eye Color and move it to the Rows pane.
7. Click **Statistics.**
8. Select Chi-square by clicking its check box.
9. Click **Continue.**
10. Click **Cells.**
11. Select **Observed** and **Expected.**
12. Click **Continue.**
13. Click **OK.** The Chi-square results appear in the Output window. Notice that all the significance levels are less than .001. Something is definitely going on here.

Measuring differences

Typically, differences in one or more continuous dependent variables based on differences in one of more categorical variables are evaluated using a t-test or an analysis of variance. If we have only one continuous dependent variable and only one categorical independent variable with no more than two values, the t-test can be used to look for differences. If we have more that one dependent continuous variable or more than two values across the categorical independent variables or we have both categorical and continuous independent variables, we need to use an analysis of variance (ANOVA).

T-Tests

There are three types of t-tests:

An Independent Samples t-test is used when cases are randomly assigned to one of two groups. After a differential treatment has been applied to the two groups, a measurement is taken which is related to the effect of the treatment. The t-test is calculated to determine if any difference between the two groups is statistically significant.

A Paired Samples t-test can be used to evaluate differences between two groups who have been matched on one or more characteristics or evaluate differences in before/after measures on same person. If you want to use pre/post measures, make sure the post-test is the same as the pre-test. This is one of the most common errors in research.

A One Sample t-test is used to evaluate whether the mean of a continuous dependent variable is different from zero. To test if the mean is different than some other value, subtract that value from each observation and the test to see if the mean of the new values is zero.

In the independent samples t-test, the purpose of randomly assigning the people to groups is to control for the effect of other differences between people that might have affected the effect we're measuring. Using the matched pairs has the same goal, but now the theory is that we have matched the subjects on the concomitant variables.

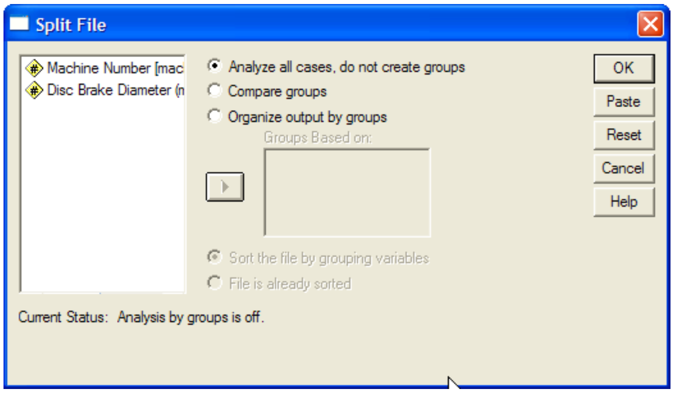
What if you can't randomly assign people to groups and you can't match them? For example, if you want to examine salary differences based on gender? One thing you can do is assume that the only difference between people that could result in a difference in income is gender. With that assumption in hand, we can do an independent sample t-test to see if there's a gender difference in income.

If you're one of those folks who think that assumption doesn't pass the "smell" test, you can go back to our earlier friend the correlation, put gender in as a binary variable (where 1 is female and 0 is not female), and look at the correlation. This does not mean that gender has *caused* the difference in salaries, but you can measure how well gender is associated with income level. More sophisticated analyses could be done by using partial correlations or multiple regression to control for concomitant factors.

How to: In this exercise, you'll split a file based upon a single variable and test the set of cases for each value of the variable as a separate sample. First you'll instruct SPSS to treat each group (in this case, each result from a specific machine) as a separate sample; this procedure is called *splitting the file.* Then you'll run the T-test to determine whether all the machines are meeting the production specifications.

1. From the menu, select File > Open > Data.
2. Navigate to the folder named SPSSTutorialData and open brakes.sav.
3. From the menu, select Data > Split File to open the Split File window (Figure 20).

FIGURE 20. Split file window



Split File

Machine Number [mac
Disc Brake Diameter (n

● Analyze all cases, do not create groups
○ Compare groups
○ Organize output by groups

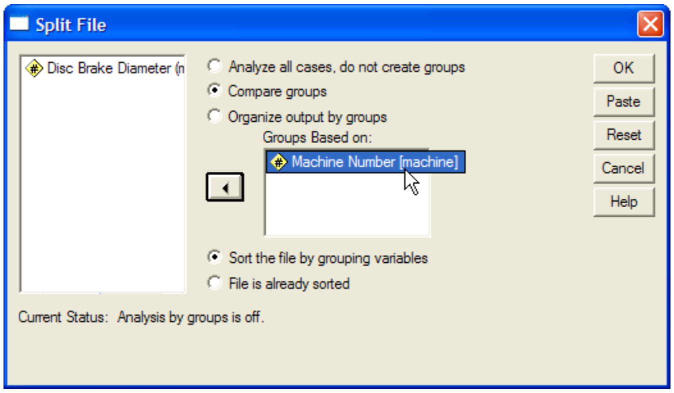
Groups Based on:

● Sort the file by grouping variables
○ File is already sorted

Current Status: Analysis by groups is off.

4. Click **Compare Groups.**
5. Select **Machine Number** and move it to the **Groups Based on** pane. (Figure 21)

FIGURE 21. Completed Split File window



Split File

Disc Brake Diameter (n

○ Analyze all cases, do not create groups
● Compare groups
○ Organize output by groups

Groups Based on:

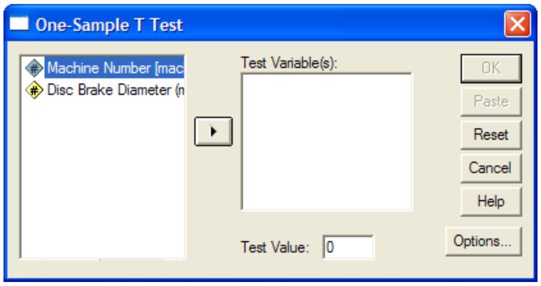
Machine Number [machine]

● Sort the file by grouping variables
○ File is already sorted

Current Status: Analysis by groups is off.

6. Click **OK.** Even though you haven't created separate files, you have instructed SPSS to treat each *group* within the file as if it were a separate sample, with a group being defined by the machine number.
7. From the menu, select Analyze > Compare Means > One-Sample T Test (Figure 22). You're going to test against a known value which in this case is the diameter of the disc brake.

FIGURE 22. One-sample T-Test window



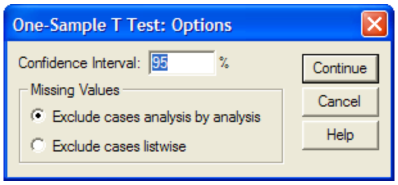
Test Value: 0

8. Select Disc Brake Diameter and move it to the Test Variables pane.
9. Select the text in the Test Value field and type:

322

10. Click **Options** to set the confidence level. (Figure 23)

FIGURE 23. One-sample T-Test: Options window



Confidence Interval: 95 %

Missing Values

● Exclude cases analysis by analysis

○ Exclude cases listwise

11. Change the confidence interval to 90.
12. Click **Continue.**
13. Click **OK.** The test results are displayed in the output window.

Homework: The exercise above is based on the SPSS tutorial. To find the tutorial, select from the menu Help > Topics. In the keyword field, type *One-Sample T Test.* Click **Show Me.**

ANOVA

If we have more than one dependent continuous variable or more than two values across the categorical independent variables or we have both categorical and continuous independent variables, we need to use an analysis of variance to measure differences. There are a number of particularly useful special cases for the general analysis of variance model.

If we have only one dependent continuous variable and one independent categorical variable we can use a One-Way Analysis of Variance or One-Way ANOVA. If we have only one dependent continuous variable, but more than one independent categorical variable we can use a general Univariate Analysis of Variance or Univariate ANOVA. If we have one or more independent continuous variables, we can use the Univariate Analysis of Covariance or Univariate ANCOVA

More advanced models are available for more than one dependent continuous variable. If the model has one or more independent categorical variables, we would use a Multivariate Analysis of Variance or MANOVA. If the model also included one or more continuous independent variables, we would use a Multivariate Analysis of Covariance or MANCOVA.[1]

One-Way ANOVA

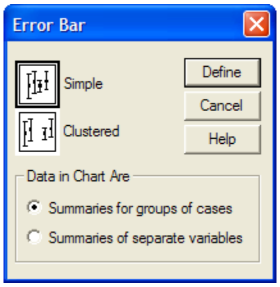
One-way analysis of variance is an extension of the t-test in that, in a t-test you have two groups, one that received a treatment and one that did not, while in a one-way analysis of variance you have more than two groups where the groups received different variation of the same treatment. For example, a treatment factor could be whether you fry donuts in vegetable oil, butter, or lard.

How to: In this exercise, you’re going to identify the most effective number of training hours to achieve a given result, with results scored between zero and 100, 100 being the optimal score.

1. From the menu, select File > Open > Data. Navigate to the folder named SPSS-TutorialData and open Training.sav.
2. First you’ll graph the means and standard error, so select Graphs > Error Bar (Figure 24).

[1] For more on determine the statistical procedure to use with your data, consult DataStep’s Statstical Selection Guide, included with this tutorial.

FIGURE 24. Error bar window



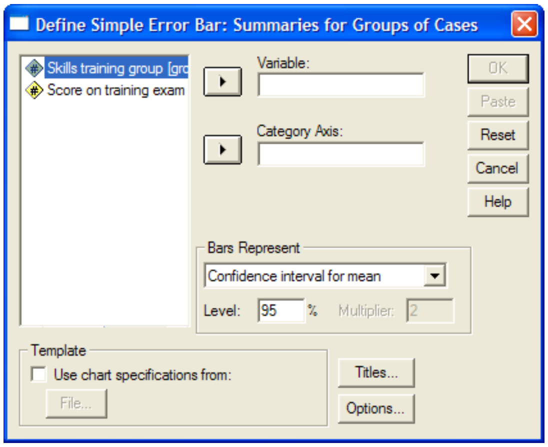
Data in Chart Are

● Summaries for groups of cases

○ Summaries of separate variables

3. You’ll use the default choices, so click **Define. (Figure 25)**

FIGURE 25. Error bar graph: defining summaries for groups of cases


Define Simple Error Bar: Summaries for Groups of Cases
Skills training group [grc
Score on training exam
OK
Paste
Reset
Cancel
Help
Titles...
Options...
File...

Variable:

Category Axis:

Bars Represent: Confidence interval for mean

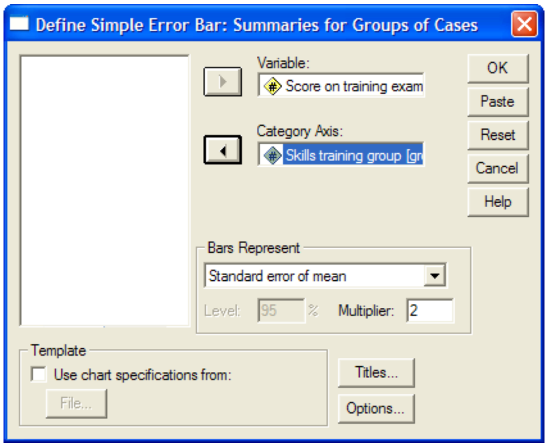
Level: 95 %   Multiplier: 2

Template

□ Use chart specifications from:

4. Select **Skills Training group** and move it to the Category Axis field.
5. Select **Score on training exam** and move it to the Variable field.
6. In the **Bars represent** field, open the drop-down list and select **Standard error of mean.** (Figure 26)

FIGURE 26. Completed Error Bar window



Define Simple Error Bar: Summaries for Groups of Cases

Variable: Score on training exam

Category Axis: Skills training group [gr

Bars Represent: Standard error of mean

Level: 95 %

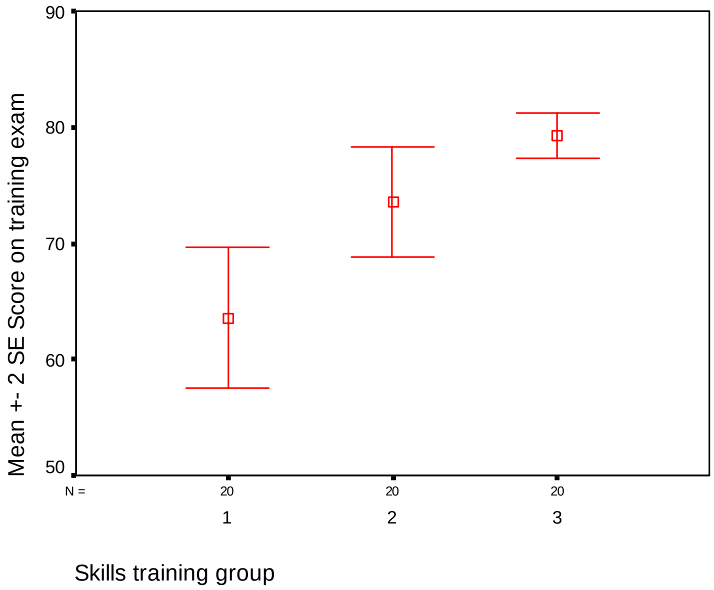
Multiplier: 2

Template

□ Use chart specifications from:

7. Click **OK.** The results are displayed in the output window. (Figure 27)

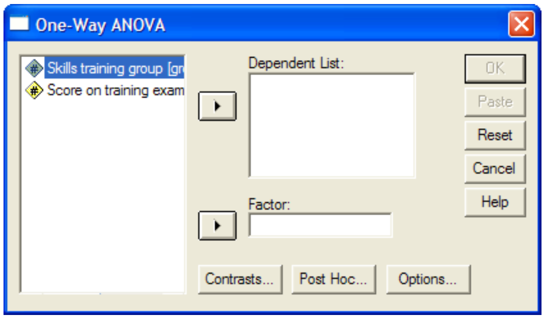
FIGURE 27. Standard error graph for skills training groups



Note that the three bars do not represent equal variance among the three groups; instead, variance decreases as days of training increase. These data may not be appropriate for ANOVA. In the next step, you'll run the ANOVA to test the data.

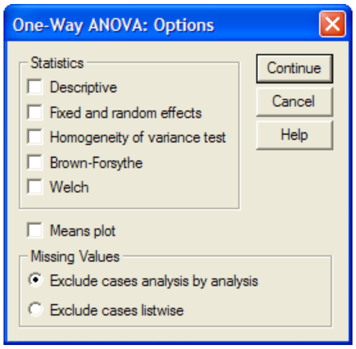
8. From the menu, select Analyze > Compare Means > One-Way ANOVA. (Figure 28)

FIGURE 28. One-Way ANOVA window



9. Move **Score on training exam** to the Dependent List.
10. Move **Skills training group** to the Factor list.
11. Click **Options.** (Figure 29)

FIGURE 29. One-Way ANOVA Options window



Statistics
☐ Descriptive
☐ Fixed and random effects
☐ Homogeneity of variance test
☐ Brown-Forsythe
☐ Welch

☐ Means plot

Missing Values
● Exclude cases analysis by analysis
○ Exclude cases listwise

12. Select **Descriptive** and **Homogeneity of variance test** by clicking their check boxes.
13. Click **Continue.**
14. Click **OK.** The results appear in the output window.

Homework: This exercise is based on the One-Way ANOVA test in the SPSS tutorial. The SPSS tutorial explains the meaning of the results and the displayed output. To use the tutorial, select Help > Topics. In the keyword field, type **One-Way ANOVA**. In the Topics Found window, double-click One-Way ANOVA. In the Help window, click **Show Me** to view the tutorial.

Summary

This chapter has provided you with a brief introduction to some of the capabilities of the SPSS application. You should now feel comfortable navigating the interface, creating contingency tables (crosstabs), conducting some of the most common statistical tests, and working with charts to present your data in a graphial format. Most importantly, you should now know how to learn more about the SPSS functions. You might even want to check out some of those 2,032 books at Amazon.com. Yes, you have noticed correctly. Sometime between last week and this week we lost two books in the search. Who knows, perhaps you'll write the next one.

Summary