Introduction to SPSS

Instructor: Jeromy Anglim

TABLE OF CONTENTS

1. OVERVIEW OF THE COURSE

Aims of the course

- To enable participants to use SPSS to solve practical problems in their work or research
- To enable participants to conduct interpret and report the most common statistical analyses in SPSS
- Insert more aims

About the Author

My name is Jeromy Anglim. I lecture and tutor statistics in psychology at the University of Melbourne. I am completing my PhD developing mathematical models for describing processes of skill acquisition and training. I am also completing my masters of industrial and organisational psychology. I work as a consultant for a number of organisations (i.e., market research, employee opinion surveys, psychological test construction, selection and recruitment, policy evaluation, academic research advice) where I provide statistical advice and perform analyses. I am passionate about statistics as I see it as a powerful tool for bridging the gap between observations of the world and knowledge.

Your background

What motivated you to take this course?

How do you plan to use SPSS in your job?

What kinds of research questions do you have?

Approach of the course

The approach of the course is to give you a basic introduction to performing statistical analyses in SPSS. The aim is to teach you how to use the most common statistical analyses in SPSS. Detailed formulas and hand calculations will rarely be discussed.

The course aims to achieve a balance between teaching a wide range of techniques and giving you enough information to effectively use these techniques. In most cases, more time could be spent on each technique in order to give a deeper and richer understanding of the underlying statistical concepts. If you are particularly interested in a statistical technique, then you may want to read more extensive information about this. The emphasis here is on how to use SPSS in terms of running analyses and appropriately interpreting output.

The course material is based on SPSS version 14. A number of improvements have been made with each successive version, although the majority of this book can be applied with little or no modification to earlier versions at least going back to version 10.

Course Content

- Introduction to statistics with practical advice on application to industry and academic settings
- Overview of SPSS software environment
    - Opening & saving data
    - Setting up the data file
    - Defining variables

    - Data entry
    - Importing data from Excel and other data formats
    - Overview of syntax and output windows
- Descriptive statistics (means, standard deviations, frequencies, etc.)
- Dealing with missing data
- Graphs (scatter plots, histograms, bar and line graphs, etc.) and Table creation and interpretation
- Customising output for reports (Pivot Trays, table formatting)
- Data manipulation (creating scales, restructuring data, merging data files, recoding variables, filtering, split files)
- Training (running in SPSS, testing assumptions, interpreting output, writing up and presenting in report format, basic theory) in the use of the following statistical procedures:
    - ANOVA
    - T-tests (one sample, independent groups, repeated measures)
    - Correlation
    - Multiple Regression
    - Factor Analysis
    - Principal Components Analysis
    - Chi-Square
    - Reliability Analysis
- Combining SPSS Output with Microsoft Excel and Word for generating reports
- Tips and tricks to:
    - Make running analyses more efficient
    - Learn more advanced statistical techniques
    - Avoid errors in analyses, etc.
- Automation of running analyses (e.g., weekly reports)

2. INTRODUCTORY CONTENT

General Advice

Overview

What is statistics? Why do we use SPSS? What questions do we hope to answer? How should we approach data analysis?

At a basic level data analysis can be seen as a set of techniques for answering specific questions. As you become more skilled at data analysis, commonalities between techniques will become more apparent. Good data analysis skills emerge as the particular techniques become integrated into a broader statistical approach. Some of these broader issues are discussed below.

Data analysis is a tool to answer questions

It is generally desirable to have specific questions that you wish to answer from your data analyses.

Garbage in – Garbage Out

Sometimes, it does not matter how skilled you are at data analysis. If the data that is collected has errors or is unreliable, it does not matter how good you are at SPSS. The results will be bad. As a data analyst, it is important to take an interest in the quality of data collection. Using a quality research design with good measurement instruments and process that maintains the integrity of the data is critical.

Exploratory versus confirmatory analyses

When we conduct data analyses, there are two broad approaches we can adopt.

Confirmatory approach: We can conduct the analyses with particular expectations and hypotheses in mind. We may want to see whether a particular relationship exists or whether particular groups differ on some outcome.

Exploratory approach: We can examine our data set with few expectations. We can explore the data and try to extract meaningful relationships. We might conduct many tests to see what is statistically significant

In most data sets there are many relationships and questions that could be explored. The confirmatory approach attempts to limit these to a smaller set of possibilities. The benefit of the confirmatory approach is that you are less likely to find statistically significant relationships by chance. The benefit of the exploratory approach is that you can really get to know your data and you may discover meaningful relationships that you were not expecting. As a general rule we will have more confidence in findings obtained based on a confirmatory approach as the findings are less likely to be a function of chance.

In reality when analysing data we commonly use a combination of the two approaches. We may have expectations but be open to exploring the data. An important consideration is to know when you are in a confirmatory mode of analysis and when you are in an exploratory mode and acknowledge this in your reporting.

Think about your client

When performing statistical analyses, it is important to think about your audience. What level of statistical understanding do they have? How will you present your analyses in a way that they will understand? What information is most useful?

If you are presenting information to an academic audience, there may be very different expectations to when you present to another business professional or if you present to member of the general public.

Improving your intuition

At first statistics can appear like a cookbook. A technique can appear like a set of steps to be followed. While this is an important first step, it is equally important to move beyond this and start to think about what the numbers really mean. Try to draw links between your everyday understanding of the world and the data. Compare your expectations of the data to what the SPSS output is saying. Examine graphical representations that make the understanding more intuitive.

Industry applications

In industry the most common analyses that are performed are relatively simple. Displaying graphs and presenting descriptive tables are very powerful tools for summarising data.

Skill Acquisition

The number one predictor of people's level of performance in specialised domains is the amount and kind of practice. Learning data analysis is no different. Learning how to use a particular statistical technique requires practice with feedback. Repetition in solving data analytic problems is critical.

SPSS is designed to make it easier to explore different options. Using help files, taking courses and having the right guide materials all speed up the process of learning.

SPSS is also structured to allow gradual development of skills. Once you get your handle on the basic techniques discussed in this book, there are many more advanced topics that can be explored. These include different statistical techniques, using syntax to automate techniques, and using more advanced data manipulation.

Variables

Level of measurement

Overview

Variables can be categorised into different types. The type of the variable has implications for the type of analysis you perform.

Nominal

Categorical or nominal variables are discrete, unordered categories.

Examples include race, favourite food, political preference, favourite television show,

Ordinal

Limited number of ordered categories where the relative distance between the categories is not necessarily equal. If you think about the order someone comes in a running race. The difference in completion times between first and second are not necessarily the same as the difference between second and third. With ordinal variables all we know is that a score is higher or lower than another, but not the relative distance between scores.

Examples include 5-point rating scales, rankings in a race.

It should be noted that ordinal variables are frequently used in analyses which assume interval data. For example, when we use a 5-point strongly disagree to strongly agree item as the dependent variable in a t-test.

Ratio and Interval

Ratio and interval variables are ordered and the distance between two data points is assumed to be equal. The difference between interval and ratio variables is that ratio variables assume that zero is inherently meaningful. With ratio scales you can speak of someone being 20% higher on a variable than someone else. This is not possible with interval scales.

Examples of interval scales include temperature in degrees.

Examples of ratio scales include height, time, frequency,

Binary

Binary variables are those that take two values. These are sometimes thought of as nominal, but for many contexts they can be treated differently.

Examples include Yes/No, gender, high/low, good/bad, old/young.

Independent vs Dependent

Independent / Predictor / Factor

In an experimental context the independent variable is the variable that we use to explain a particular outcome. Examples include whether someone has received training, the country they come from, or gender. This variable is used to explain differences on a dependent variable. The independent variable is often also referred to as a factor.

In the context of multiple regression variables used to explain a dependent variable are typically called predictor variables. While there are different ways of describing a variable that is used to predict another variable, the terms can be used interchangeably.

Dependent / Outcome

The dependent variable is what we are trying to explain.

Discrete vs continuous

Another distinction made between variables is whether they are discrete or continuous

Variables are discrete when they can take on a limited set of possible values.

Variables are continuous when there are an infinite number of possible values that can occur between two points. For example, between 1 minute and 2 minutes are an infinite number of time points. As long as we had a sufficiently precise clock we could measure time to many decimal places.

3. Overview of SPSS Software Environment

Opening & saving data

When starting an SPSS session you can either open an existing data file or you can enter data yourself.

To open an existing data file: File >> Open >> Data…

The default SPSS data file ends with a ".sav" file extension.

Setting up the data file

Defining variables

SPSS

In SPSS there are two ways of viewing the data: variable view and data view. You can toggle between the two views by clicking on one of the tabs at the bottom of the data editor.

Data view is used to enter the data. Each row is a case and each column is a variable. If you want to enter data just start typing it into the cells. Pressing the Enter Key will move you to the next row and pressing the tab key will move you to the next variable.

Variable view is used to define the variables in your data file. Each row is a variable and each column is a property of the variable. Typically you would define your variables before entering the data.

Type

Most data entered into SPSS can be represented as 'Numeric'.

Sometimes you need to enter 'String' data, such as when you wish to record someone's name.

Width

For string variables it specifies the maximum number of characters that the variable can take.

Decimals

This determines the number of decimal places that are displayed in the data view and in output. If you are dealing with data that should have no decimal places, it is often clearer to set this to zero.

Names

Variable names can not contain spaces

Use variable names that are simple to work with e.g., s1 to s20 for a 20 item questionnaire. It is best to stick to a set of conventions that are meaningful to you.

Labels

Labels are useful descriptions of each variable. These are typically shown in output to make the meaning of the variable clearer.

Value Labels

If you have a categorical variable it is better to assign a number to each category.

Values are used to indicate the relationship between the numbers you enter into the data editor and the descriptive names they have.

These value labels are then displayed in SPSS Output.

When in data view, you can select: View >> Value Labels. This toggles between displaying data labels or the original numbers.

Missing Data Codes

There are two strategies for indicating missing values. You can either just leave the cell blank or you can assign a special number to be a missing value code. When choosing a missing value code it is important to select a value that could not actually occur (e.g., -9 is common).

Columns

This defines the number of columns that are used to display the information.

Data entry

SPSS prefers to work with numbers. Thus, if the variable was gender with values of male and female, you would code male = 1 and female = 2.

Pressing Tab moves the entry point into the next column

Importing data from Excel and text files

Overview

When working with others you will frequently be given data files in a range of different formats. It is useful to be able to import the data into SPSS from different formats.

Excel

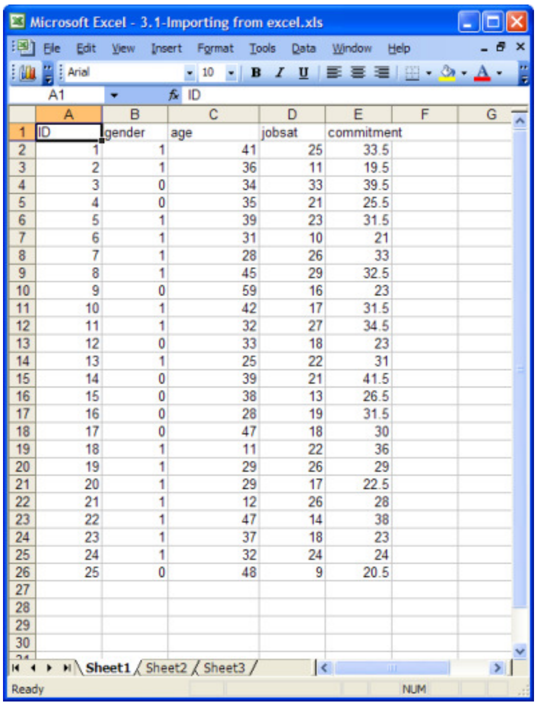
Overview

Working with data in Microsoft Excel is very common. It is relatively easy to import data from Excel into SPSS.

Process

Step 1. Prepare the excel file for importation

The simplest way to import data from excel is to have all the data in a separate worksheet in Excel. The first row has the variable name and the rows that follow have the data. When it is set up, save and close the Excel file.

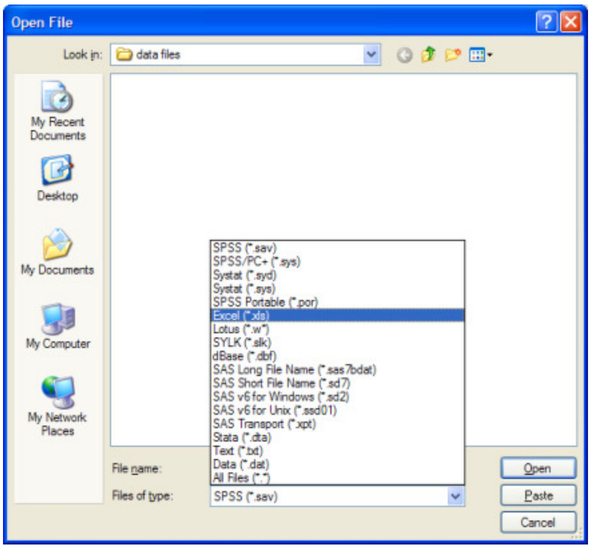

Microsoft Excel - 3.1-Importing from excel.xls
File Edit View Insert Format Tools Data Window Help
Arial 10
A1 ID
Sheet1 Sheet2 Sheet3
Ready NUM

ID	gender	age	jobsat	commitment
1	1	41	25	33.5
2	1	36	11	19.5
3	0	34	33	39.5
4	0	35	21	25.5
5	1	39	23	31.5
6	1	31	10	21
7	1	28	26	33
8	1	45	29	32.5
9	0	59	16	23
10	1	42	17	31.5
11	1	32	27	34.5
12	0	33	18	23
13	1	25	22	31
14	0	39	21	41.5
15	0	38	13	26.5
16	0	28	19	31.5
17	0	47	18	30
18	1	11	22	36
19	1	29	26	29
20	1	29	17	22.5
21	1	12	26	28
22	1	47	14	38
23	1	37	18	23
24	1	32	24	24
25	0	48	9	20.5

Step 2. Open in SPSS

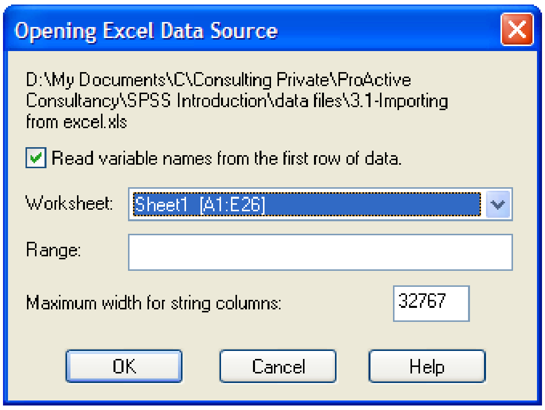
2.1 File >> Open >> Data

2.2 Select "Files of Type" and click excel



2.3 Select the Excel file of interest and Open

2.4 Specify the worksheet that you have the data in and whether the variable names are shown in the first row. Note that it is possible to import only a range of data. Click OK.



D:\My Documents\C\Consulting Private\ProActive Consultancy\SPSS Introduction\data files\3.1-Importing from excel.xls

☑ Read variable names from the first row of data.

Worksheet: Sheet1 [A1:E26]

Range:

Maximum width for string columns: 32767

If there was nothing unusual, the excel file will open in SPSS. You may then want to save it as an SPSS file. You may also want to customise the decimal places, labels, value labels, etc to make the data ready for subsequent analyses.

Common errors arise from not closing the original Excel file, having variable names that are not compatible with SPSS, and choosing the wrong worksheet.

Text

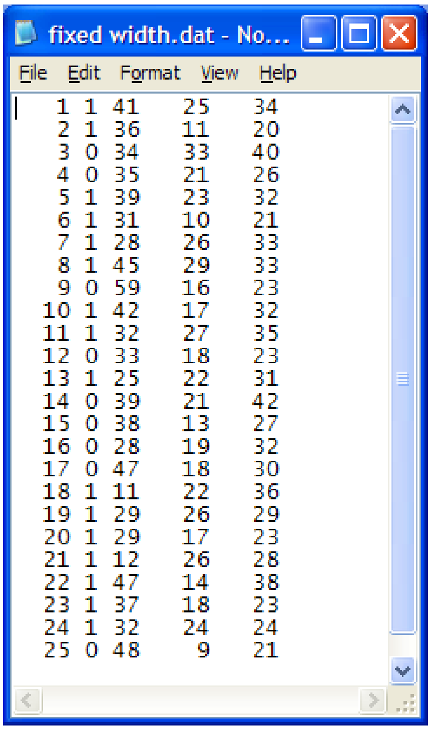
Overview

Data can be represented in many different file formats. You may find that you are obtaining data from people who use different software, databases and operating systems. One format for representing information common to almost every program is text. Although SPSS programs can not open every file format directly, most programs will export data to a text format.

Fixed width

This format allocates a certain number of spaces to each variable. In the present example these widths were:

ID (3), gender (2), age (3), jobsat (5), commitment (5)

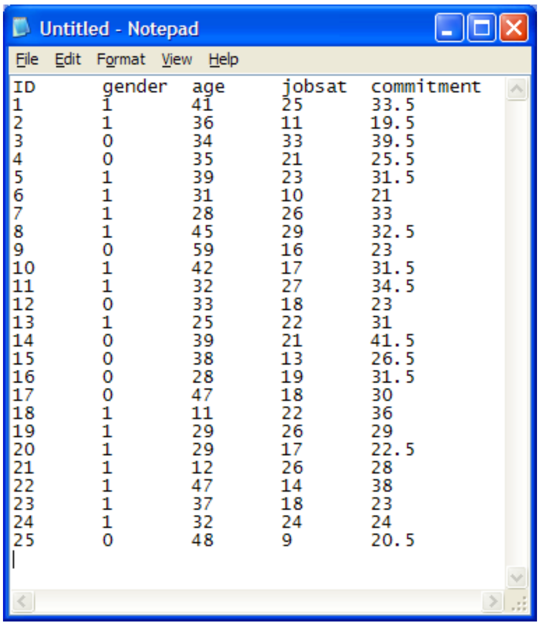


```
  1 1 41   25   34
  2 1 36   11   20
  3 0 34   33   40
  4 0 35   21   26
  5 1 39   23   32
  6 1 31   10   21
  7 1 28   26   33
  8 1 45   29   33
  9 0 59   16   23
 10 1 42   17   32
 11 1 32   27   35
 12 0 33   18   23
 13 1 25   22   31
 14 0 39   21   42
 15 0 38   13   27
 16 0 28   19   32
 17 0 47   18   30
 18 1 11   22   36
 19 1 29   26   29
 20 1 29   17   23
 21 1 12   26   28
 22 1 47   14   38
 23 1 37   18   23
 24 1 32   24   24
 25 0 48    9   21
```

Figure: Fixed width

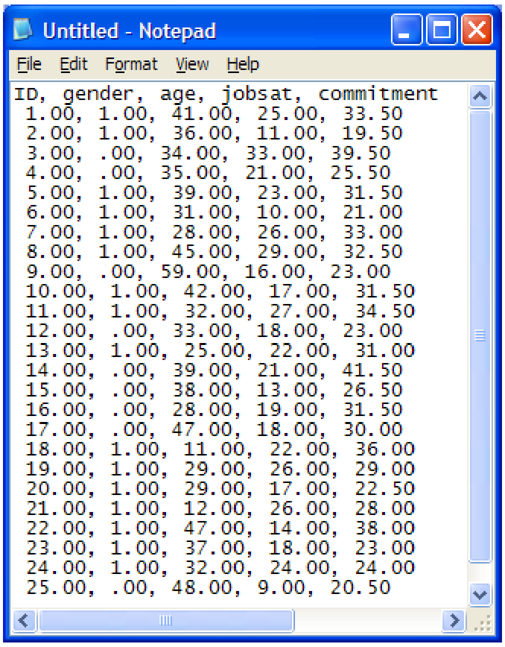
Tab or comma delimited

Tab delimited data



```
ID	gender	age	jobsat	commitment
1	1	41	25	33.5
2	1	36	11	19.5
3	0	34	33	39.5
4	0	35	21	25.5
5	1	39	23	31.5
6	1	31	10	21
7	1	28	26	33
8	1	45	29	32.5
9	0	59	16	23
10	1	42	17	31.5
11	1	32	27	34.5
12	0	33	18	23
13	1	25	22	31
14	0	39	21	41.5
15	0	38	13	26.5
16	0	28	19	31.5
17	0	47	18	30
18	1	11	22	36
19	1	29	26	29
20	1	29	17	22.5
21	1	12	26	28
22	1	47	14	38
23	1	37	18	23
24	1	32	24	24
25	0	48	9	20.5
```

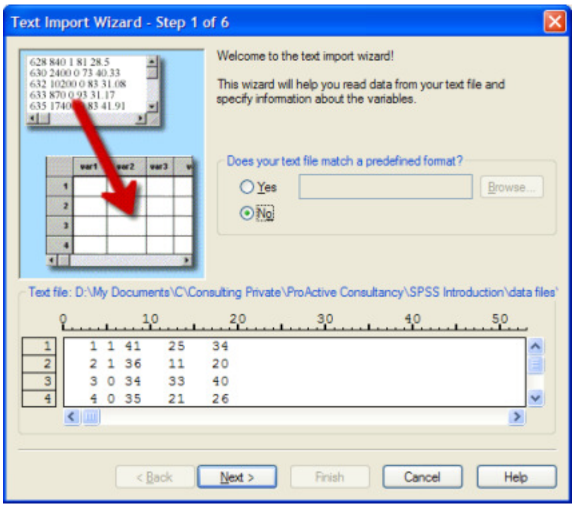
Comma delimited data



```
ID, gender, age, jobsat, commitment
 1.00, 1.00, 41.00, 25.00, 33.50
 2.00, 1.00, 36.00, 11.00, 19.50
 3.00, .00, 34.00, 33.00, 39.50
 4.00, .00, 35.00, 21.00, 25.50
 5.00, 1.00, 39.00, 23.00, 31.50
 6.00, 1.00, 31.00, 10.00, 21.00
 7.00, 1.00, 28.00, 26.00, 33.00
 8.00, 1.00, 45.00, 29.00, 32.50
 9.00, .00, 59.00, 16.00, 23.00
 10.00, 1.00, 42.00, 17.00, 31.50
 11.00, 1.00, 32.00, 27.00, 34.50
 12.00, .00, 33.00, 18.00, 23.00
 13.00, 1.00, 25.00, 22.00, 31.00
 14.00, .00, 39.00, 21.00, 41.50
 15.00, .00, 38.00, 13.00, 26.50
 16.00, .00, 28.00, 19.00, 31.50
 17.00, .00, 47.00, 18.00, 30.00
 18.00, 1.00, 11.00, 22.00, 36.00
 19.00, 1.00, 29.00, 26.00, 29.00
 20.00, 1.00, 29.00, 17.00, 22.50
 21.00, 1.00, 12.00, 26.00, 28.00
 22.00, 1.00, 47.00, 14.00, 38.00
 23.00, 1.00, 37.00, 18.00, 23.00
 24.00, 1.00, 32.00, 24.00, 24.00
 25.00, .00, 48.00, 9.00, 20.50
```

SPSS Text Import Wizard

File >> Read Text Data…



Text Import Wizard - Step 1 of 6

Welcome to the text import wizard!

This wizard will help you read data from your text file and specify information about the variables.

Does your text file match a predefined format?

○ Yes

● No

Text file: D:\My Documents\C\Consulting Private\ProActive Consultancy\SPSS Introduction\data files'

```
   1 1 41    25    34
   2 1 36    11    20
   3 0 34    33    40
   4 0 35    21    26
```

Choices: Fixed width or delimited

Variable names: are they included in the first row line of the file

If the data is fixed width, SPSS will attempt to guess where the breaks are, but you may need to adjust these. If the data is delimited, you need to let SPSS know what character separates each variable (e.g., comma, space, tab, etc.).

Syntax Window

Syntax is a powerful tool

While some users new to SPSS find dealing with syntax files intimidating, there are many benefits to be gained from getting acquainted with SPSS syntax.

- It makes re-running analyses more reliable and efficient.
- It documents the analysis process.
- It makes more explicit the options selected.
- It allows others to review the work done.
- Some options and analyses can only be done with syntax.

Working with syntax

The following tips and tricks should make SPSS syntax more accessible.

- To bring up a new syntax window, go to: File >> New >> Syntax
- Almost all SPSS dialog boxes have a paste command. Once you have set up the options you can paste the SPSS syntax and then run it.
- All SPSS commands start with one or more keywords followed by options and end with a dot "."
- To run a command highlight it and press the run button or select: Run >> Selection.
- To document your syntax, start the line with a "*" and end it with a "."
- I recommend changing the font to 'courier' as it makes it easier to identify problems.
- Syntax files can be saved for later use. They have a '.sps' file extension.
- Pressing F1 after typing a command brings up the SPSS syntax template for that command.

Output Window

Navigating Output

SPSS displays the results of analyses in an “Output Window”. It is composed of a navigation pane on the left and the output on the right. To select a particular piece of output it is generally more efficient to use the navigation pane on the left.

Exporting Output

SPSS Output Files

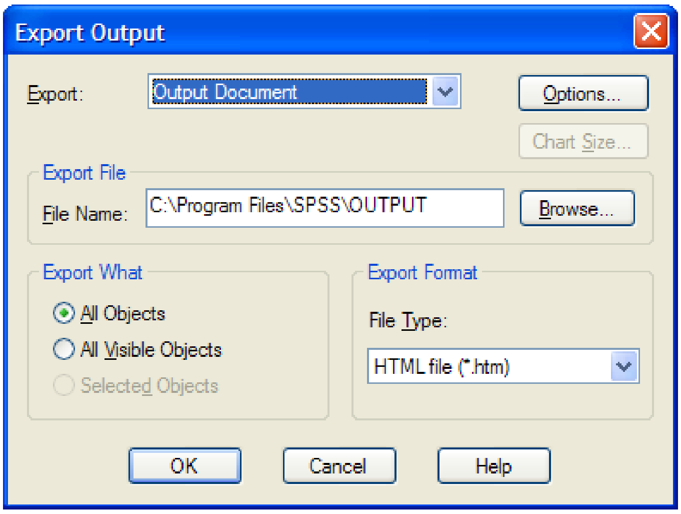
You can save an SPSS Output file by going to File >> Save within an Output window. SPSS Output can be saved as a “.spo” format. This format is good when you know you will be working with SPSS. However, the output will only open in SPSS. In addition, SPSS Output files from one version usually do not open in previous versions of SPSS.

Different Output File formats

If you are working with other people who do not have SPSS, you will probably wish to save the SPSS Output in a different file format. SPSS allows you to save output as HTML, text, Word (RTF), Excel, or PowerPoint.

You can export Output to a different format by going to: File >> Export

Choose the elements you want to export (all output, charts, not charts); file location; which output (all or just those selected; and file format (Word is often most useful).



Export Output

Export: Output Document

Export File

File Name: C:\Program Files\SPSS\OUTPUT

Export What

● All Objects

○ All Visible Objects

○ Selected Objects

Export Format

File Type: HTML file (*.htm)

Copying and Pasting

Alternatively, you can copy and paste individual tables and charts from SPSS output into other programs.

SPSS Options

To change SPSS options, go to: Edit >> Options

In general most of the default options are reasonable for conducting analyses.

A few things that you may sometimes want to do include the following:

GENERAL TAB: click Variable List – Display names: This option changes the way variables are displayed in SPSS dialog boxes to only show the variable names and not the labels. This can make it easier to search large lists of variables.

GENERAL TAB: click Open syntax window at start-up. This can be useful if you use syntax a lot.

VIEWER TAB: click Display commands in log. If you want to get familiar with SPSS syntax, this will display the syntax used for each command in the output window.

VIEWER TAB: Width custom (255); Length infinite. This makes output from some SPSS procedures better for displaying on the screen, but not so good for printing.

OUTPUT LABELS TAB: This allows you to decide whether to show values or value labels and whether to show variables or variable labels in SPSS Output.

Help in SPSS

SPSS is a massive product. The more you start using it, the more you realise it can do. It is not possible to know every feature and option. A better strategy is to know how to empower yourself to get the information you need, when you need it.

SPSS provides many sources of help to assist you with conducting analyses.

Right clicking on a dialog box

This should give an explanation of what the particular option means.

Right clicking on Output tables

You can then select Results Coach or Case Studies to have the output explained to you.

General Help System

Going to: Help >> Topics

This will bring up the SPSS general help files, which are getting better and more accessible with each release of SPSS.

External Help

There are many books and websites that can help you. In terms of the internet, a good starting point is to type in the statistical technique you are using into Google. Add the term SPSS for some more SPSS specific information.

4. DESCRIPTIVE STATISTICS

Theory

Central Tendency

Mean

This is the most commonly reported measure of central tendency. It involves adding up all the scores and dividing by the number of scores. It is appropriate for continuous data

$$\overline{X} = \frac{\sum X}{N}$$

$\overline{X}$ is the mean of all X scores

$\sum X$ is the sum of all X scores

N is the number of scores

Median

The median is the middle score. If all scores were ordered from highest to lowest, it is the middle score. The median is the score at the 50th percentile. It is particularly useful for describing ordinal data and continuous data with skewed distributions. It is more resistant to the effect of outliers than the mean.

Mode

The mode is the most frequently occurring category. It is most appropriate for describing categorical data. If you ask people, what their favourite television channel is, the modal response would be the most frequently cited channel. To calculate the mode, you calculate the frequencies for all response categories and identify the most frequently occurring category.

Measures of spread

Variance

Variance is the mean of squared deviations from the mean. A deviation from the mean is just the difference of a score from the mean. Squaring the difference from the mean has the effect of removing the sign associated with the difference (e.g., -3 squared = 9; 3 squared = 9). Explaining variance is a recurring theme in statistics.

Population Variance = $\sigma^2 = \frac{SS}{N}$

Sample Variance = $s^2 = \frac{SS}{N-1}$

SS = Sums of Squares = $\sum (X - \overline{X})^2$

X = each score

$\overline{X}$ = the mean of all X scores

N = number of scores

Standard deviation

The standard deviation characterises the typical deviation from the mean. It is the square root of variance. This has arguably more intuitive meaning than variance.

Population Standard deviation = $\sigma = \sqrt{\frac{SS}{N}}$

Sample Standard deviation = $s = \sqrt{\frac{SS}{N-1}}$

SS = Sums of Squares = $\sum (X - \overline{X})^2$

N = number of scores

Interquartile Range

The interquartile range represents the width of the middle 50% of scores. It is the score for the 75th percentile minus the score for the 25th percentile.

Semi-interquartile range

The semi-interquartile range is half the interquartile range.

Range

The range of scores is the difference between the smallest and largest score.

Range = maximum - minimum

Frequencies

Frequencies describe the number of scores of a particular value. Frequency tables can be expressed in raw counts or as a percentage.

Z-scores

Theory

Z-scores are a useful way of describing an individual's score in a standardised way. A distribution of z-scores has a mean of 0 and a standard deviation of 1. On a normal item, when someone gets a score of 7 out of 10, we do not necessarily know what this means. We need to compare this score to some frame of reference or benchmark in order to understand it. If I said that someone has driven for 30 years and never had an accident, we would agree this is good (or lucky), because we assume that the normal person is likely to have an accident or two over that period of time. Z-scores tell use where someone stands in relation to the mean. Thus, a z-score of 1 indicates that some one is one standard deviation above the mean.

$$Z = \frac{X - \overline{X}}{s}$$

Z is the standardised score

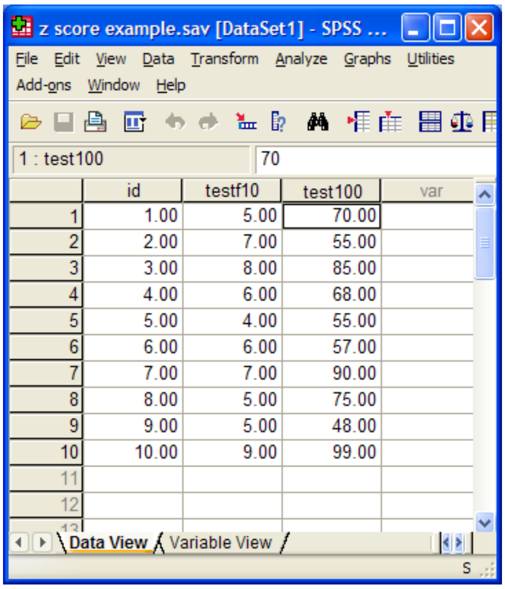
X is the individual's score

$\overline{X}$ is the mean for the variable

s is the standard deviation

SPSS

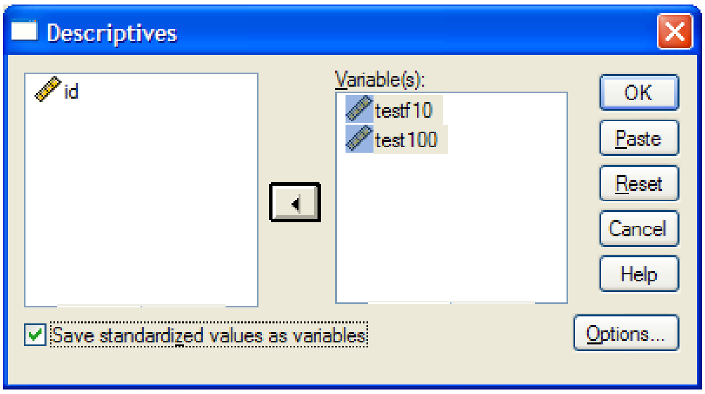
In the present example we are assuming we have scores for 10 employees on two tests. One measure was out of 10 and measured social skills another was out of 100 and measured technical skills. Assume that you want to know which participants did better than average on both so that you can promote them.


z score example.sav [DataSet1] - SPSS ...
File Edit View Data Transform Analyze Graphs Utilities
Add-ons Window Help
1 : test100
70
Data View
Variable View

	id	testf10	test100	var
1	1.00	5.00	70.00	
2	2.00	7.00	55.00	
3	3.00	8.00	85.00	
4	4.00	6.00	68.00	
5	5.00	4.00	55.00	
6	6.00	6.00	57.00	
7	7.00	7.00	90.00	
8	8.00	5.00	75.00	
9	9.00	5.00	48.00	
10	10.00	9.00	99.00	
11				
12				

Analyze >> Descriptive Statistics Descriptives

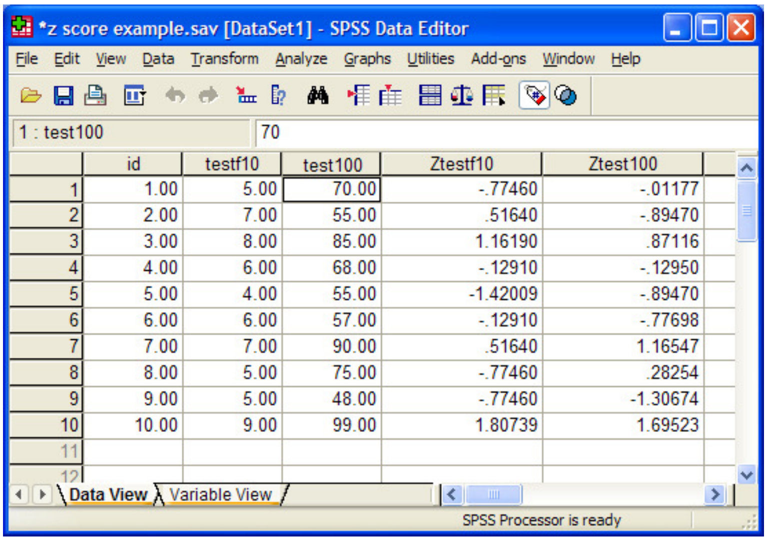
Place the two tests in the variable box and select "Save standardized values as variables"



Variable(s):
testf10
test100

☑ Save standardized values as variables

This creates two new variables in our data file that represent z-scores of the original tests. These test scores are then more comparable


*z score example.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : test100
70
Data View
Variable View
SPSS Processor is ready

	id	testf10	test100	Ztestf10	Ztest100
1	1.00	5.00	70.00	-.77460	-.01177
2	2.00	7.00	55.00	.51640	-.89470
3	3.00	8.00	85.00	1.16190	.87116
4	4.00	6.00	68.00	-.12910	-.12950
5	5.00	4.00	55.00	-1.42009	-.89470
6	6.00	6.00	57.00	-.12910	-.77698
7	7.00	7.00	90.00	.51640	1.16547
8	8.00	5.00	75.00	-.77460	.28254
9	9.00	5.00	48.00	-.77460	-1.30674
10	10.00	9.00	99.00	1.80739	1.69523
11					
12					

Distributions

Types of Distributions

Unimodal vs Bimodal

A distribution is unimodal if it only has one mode or one peak. A distribution is bimodal if it has two peaks.

Symmetric vs Asymmetric

A distribution is symmetric if when you draw a line through the middle of the distribution, the left side is a mirror image of the right side. When a distribution is symmetric, its mean and median will be the same.

Rectangular

In a rectangular or uniform distribution, the distribution covers a range of values. Every value within the range of possible values is equally likely to occur.

Normal

The normal distribution is a frequently assumed distribution. It has a characteristic bell shape. It is unimodal and symmetric.

Positively Skewed

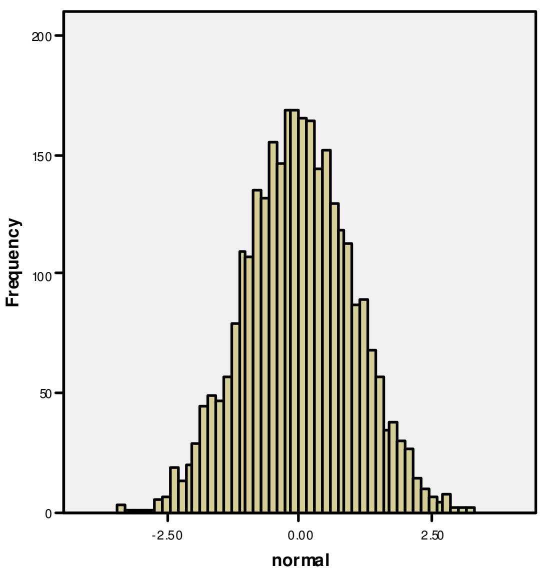
A distribution is positively skewed when its tail points to the right towards positive numbers. A common example of a positively skewed variable is income in the general population. Most people earn a little while a few earn a lot and fewer still earn a huge amount. In positively skewed distributions the mode is to the left of the median and the median is to the left of the mean. The mean gets pulled out by the extreme scores. In really skewed data, the median may be a better measure of central tendency.

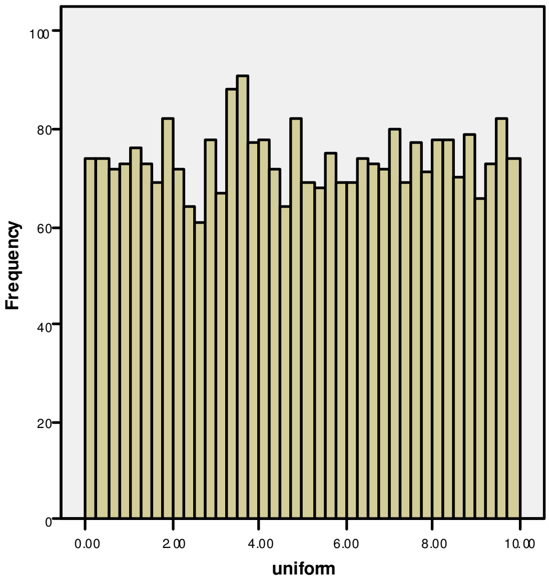
Negatively Skewed

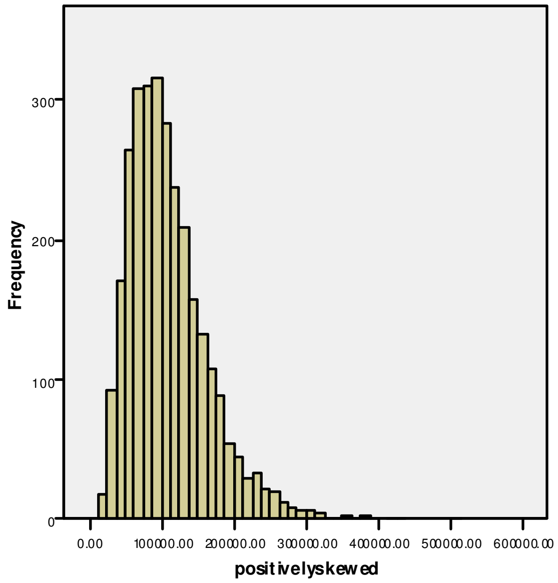
A distribution is negatively skewed when its tail points to the left towards negative numbers. The same rules apply to it as to positively skewed distribution, but in reverse.

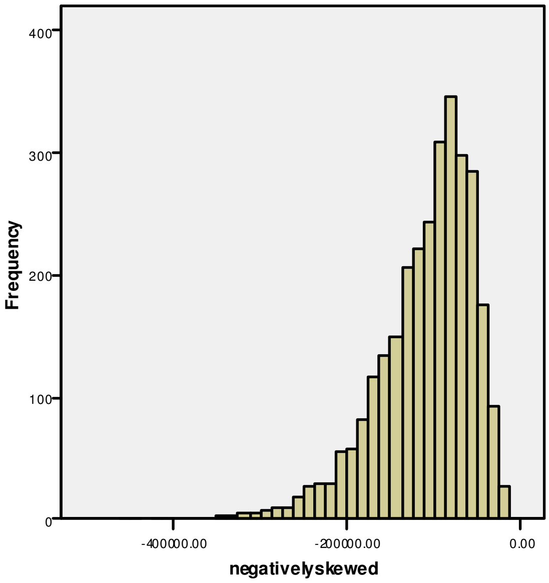
Images of Distributions

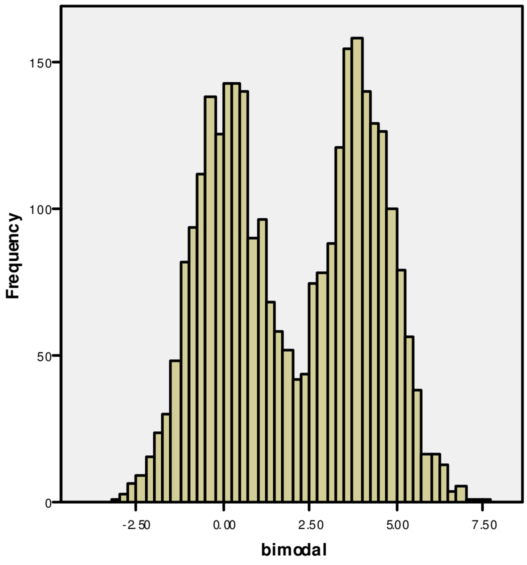
The following graphs display data for approximately 3000 cases drawn from population distributions characterized by the following shapes.







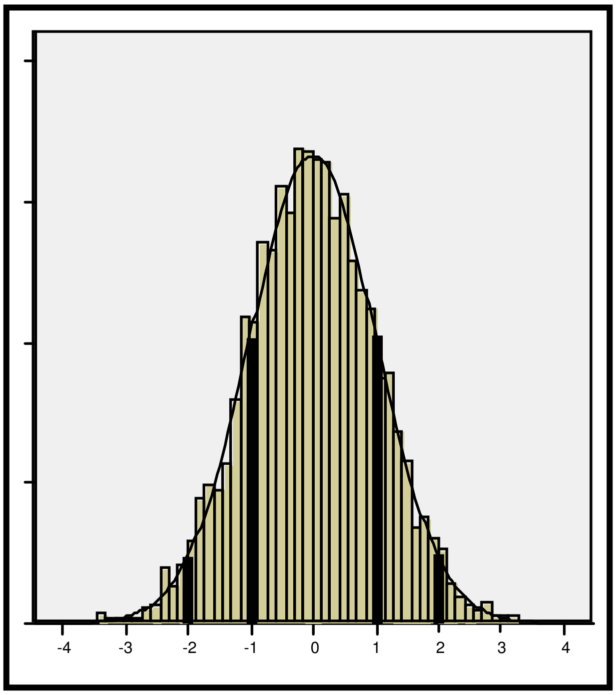




Normal Distribution

The normal distribution is reflected by a bell shaped curve. It is assumed to arise in many settings in contexts where many random processes are operating. For example, the size of noses, the height of females, shyness might all be assumed to exhibit a normal distribution. The normal distribution is an assumption of many statistical tests. In reality your data is often not normally distributed, and the question becomes what analyses should I perform. Often the tests that assume normality are relatively robust to violation of the assumption. It is often sufficient to show that our data is relatively symmetric and hope that the test is statistically robust.

When the normal distribution is composed of z-scores it is called the Z normal distribution. It has a mean of 0 and a standard deviation of 1. Based on knowledge of the normal distribution we can state that 68% of scores will be within 1 standard deviation of the mean, 95% within 2 standard deviations and 99.7% of scores within 3 standard deviations.



Assessing Distributions in SPSS

There are two main ways to explore distributional properties. You can assess them graphically or you can assess them with statistics. Graphical assessment of distributions is often better particularly if your sample size is above 100. This is because the statistical tests are often too sensitive in detecting violations of normality. To graphically assess the distribution, bring up a histogram of the variable of interest.

There are a number of ways to assess a distribution statistically. Two common summary measures are skewness and kurtosis. Skewness describes the degree to which a distribution's tails go off in one example (see the examples of positive and negatively skewed distributions above). Kurtosis refers to the degree to which the distribution is peaked or flat. Through SPSS Analyze >> Frenquencies procedure you can bring up skewness and kurtosis information. When the value is greater than 3 times the standard error of the statistic, this may suggest a significant violation of normality.

Inferential statistics

General

Imagine we wanted to draw some conclusions about the nature of employees in a particular country. On average how many hours a week do they work? How much money do they earn on average? How many weeks holiday a year do they get? If we were going to research these questions, it is rarely feasible to obtain data from every person in the population of interest. Thus, we draw a smaller sample of people and assess them on how many hours they work, how much they earn and how many weeks holiday they get each year. We then may attempt to infer the characteristics of the broader population from our sample.

Samples and population

Population (parameters)

The individuals we are trying to draw inferences about population parameters

Sample (statistics)

A selection of individuals drawn from the population that provide sample statistics to estimate population parameters

Hypothesis testing

H0: null hypothesis

H1: alternative hypothesis

p value

The probability of obtaining a result as large as that observed in the sample if the null hypothesis were true.

Alpha

The probability of falsely rejecting the null hypothesis

Typically, we talk about alpha being .05 or .01

Hypothesis testing logic

- if the p value is less than alpha (e.g., .05),
- the probability of the null hypothesis being true is low
- we reject the null hypothesis
- and accept the alternative hypothesis

Effect size

Effect size measures show the degree of relationship or extent of relationship. With a sufficiently large sample size, even small effects can be statistically significant. When evaluating the practical significance of a research finding, whether we are concerned with group differences or the relationship between variables, it is desirable to report a measure of effect size.

Many different effect size measures exist. Two common ones we will encounter during this course are cohen's d and r.

Cohen's d = $\dfrac{\overline{X}_1 - \overline{X}_2}{s}$

This can be expressed as the difference between the two group means divided by the standard deviation.

The following table shows the relationship between Cohen's d and r. Cohen provided some rules of thumb that might guide the practical understanding of obtained effect sizes. These are also displayed on the table. It should be noted that the actual importance of an effect size will vary across contexts.

Cohen's d	*r*	*Cohen's convention*
2	.71	
1.9	.69	
1.8	.67	
1.7	.65	
1.6	.63	
1.5	.60	
1.4	.57	
1.3	.55	
1.2	.51	
1.1	.48	
1	.45	
0.9	.41	
0.8	.37	Large
0.7	.33	
0.6	.29	
0.5	.24	Medium
0.4	.20	
0.3	.15	
0.2	.10	Small
0.1	.05	
0	.00	

Power

In the logic of hypothesis testing, there are two possible states in the world and we can draw two possible conclusions. If the null hypothesis is true and conclude that this is the case, then we have made a correct decision.

Power is the probability of correctly rejecting the null hypothesis when it is false in the population.

		Conclusion about the null hypothesis	
		True	False
Null Hypothesis in the world	True	Correct decision	Alpha (Type I Error)
	False	Beta (Type II Error)	Power (1-Beta)

To state the case strongly, a study that with insufficient power is not worth doing.

From a practical perspective, it is important to know what power is and what increases it.

Power increases with bigger samples.

Power increases with bigger effect sizes.

Frequencies

Overview

"Frequencies" is a tool in SPSS that is designed to give you frequency counts and percentages for your variables. It can also provide summary statistics and basic graphs for your variables. It can take any type of variable. Frequency counts and percentages are an excellent way of understanding the spread of scores on a variable. This is particularly true for nominal and ordinal data.

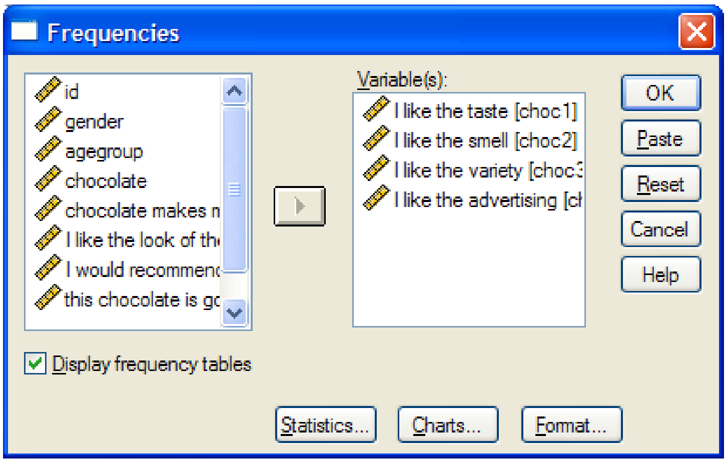
Case Study

Imagine you were exploring attitudes towards a new brand of chocolate. You asked participants a range of different questions about the chocolate on a five point scale where higher scores indicated greater agreement with the statement.

Running

Analyze >> Descriptive Statistics >> Frequencies

This tool allows you obtain frequencies, descriptive statistics and charts for variables in your data set.



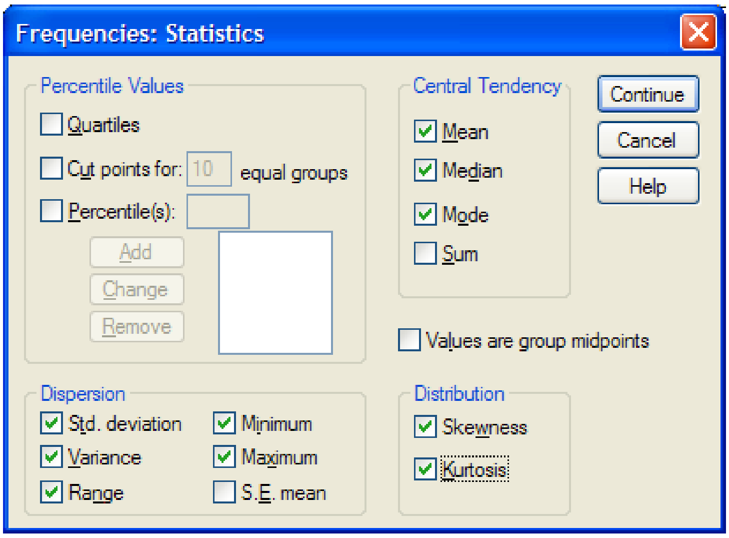
Frequencies

id
gender
agegroup
chocolate
chocolate makes n
I like the look of th
I would recommend
this chocolate is gc

Variable(s):
I like the taste [choc1]
I like the smell [choc2]
I like the variety [choc3
I like the advertising [cl

☑ Display frequency tables

Under the "statistics" option you can select a range of descriptive statistics about the variable. Remember that some of these summary measures are only relevant for certain types of data.



Frequencies: Statistics

Percentile Values
☐ Quartiles
☐ Cut points for: 10 equal groups
☐ Percentile(s):

Central Tendency
☑ Mean
☑ Median
☑ Mode
☐ Sum

☐ Values are group midpoints

Dispersion
☑ Std. deviation
☑ Variance
☑ Range
☑ Minimum
☑ Maximum
☐ S.E. mean

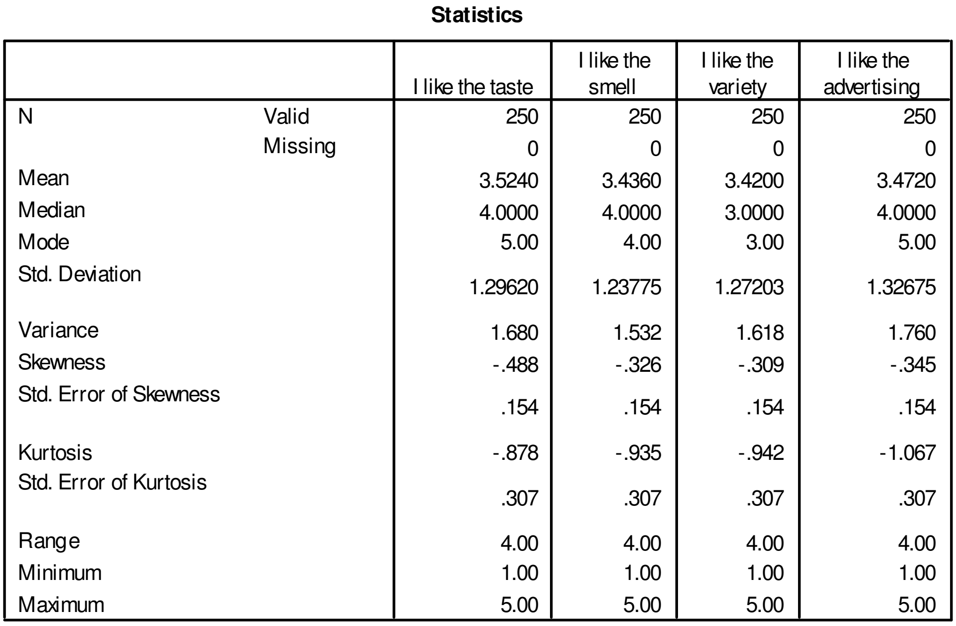
Distribution
☑ Skewness
☑ Kurtosis

Under the "charts" option you can select from Bar Charts, Pie Charts and Histograms

Below is some sample output showing descriptive statistics for four items asking participants about chocolate. Each item was on a five points scale from 1 to 5.

Output

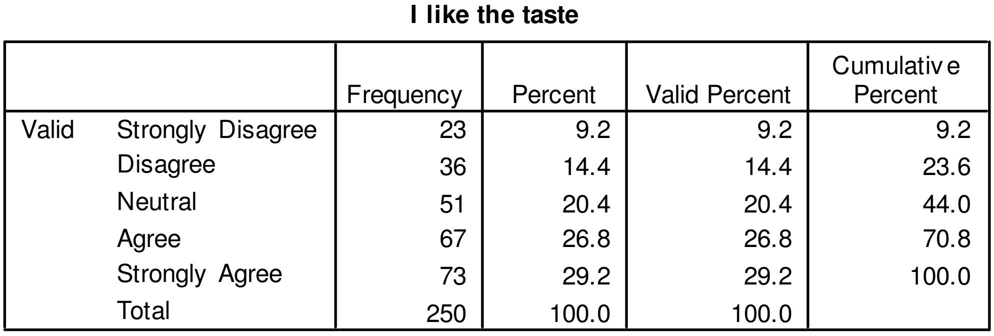
The first table provides descriptive statistics for the variables you have entered. We can find out how participants successfully answered the question (N), and how many respondents did not answer the question (Missing). You would typically take the relevant parts of this table and present this in a report.



Statistics

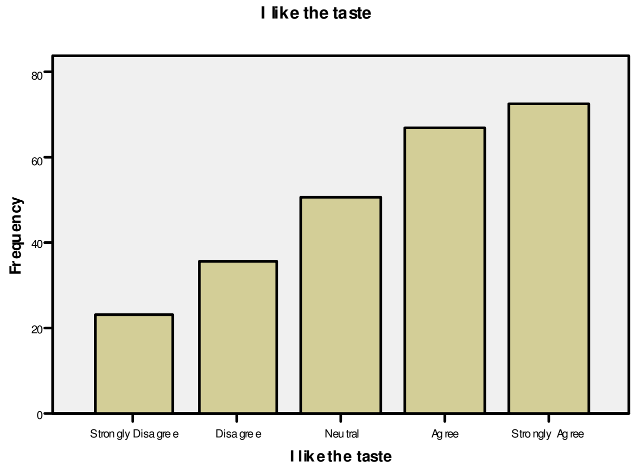
		I like the taste	I like the smell	I like the variety	I like the advertising
N	Valid	250	250	250	250
	Missing	0	0	0	0
Mean		3.5240	3.4360	3.4200	3.4720
Median		4.0000	4.0000	3.0000	4.0000
Mode		5.00	4.00	3.00	5.00
Std. Deviation		1.29620	1.23775	1.27203	1.32675
Variance		1.680	1.532	1.618	1.760
Skewness		-.488	-.326	-.309	-.345
Std. Error of Skewness		.154	.154	.154	.154
Kurtosis		-.878	-.935	-.942	-1.067
Std. Error of Kurtosis		.307	.307	.307	.307
Range		4.00	4.00	4.00	4.00
Minimum		1.00	1.00	1.00	1.00
Maximum		5.00	5.00	5.00	5.00

You also get a table with the frequency, percentage and cumulative percentage for each category. This allows you to make the claim 23 people in the strongly disagreed with the statement "I like the taste". The cumulative percentage allows you to say 23.6% of participants either strongly disagreed or disagreed with the statement, "I like the taste". The distinction between 'Percent' and 'Valid Percent' becomes important when you have missing data. 'Percent' is the percentage of all data including missing data. 'Valid Percent' is the percentage only of those that actually responded to the question.



I like the taste

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Strongly Disagree	23	9.2	9.2	9.2
	Disagree	36	14.4	14.4	23.6
	Neutral	51	20.4	20.4	44.0
	Agree	67	26.8	26.8	70.8
	Strongly Agree	73	29.2	29.2	100.0
	Total	250	100.0	100.0	



Descriptives

Overview

The SPSS tool 'Descriptive' is used for ordered variables to get measures of central tendency, spread and distributional properties, like the mean, median, standard deviation and the range. It can also be used to create standardised versions of variables (i.e., z-scores). Its functionality large overlaps with 'Frequencies'. This is often the case with SPSS that there are multiple ways of getting similar output.

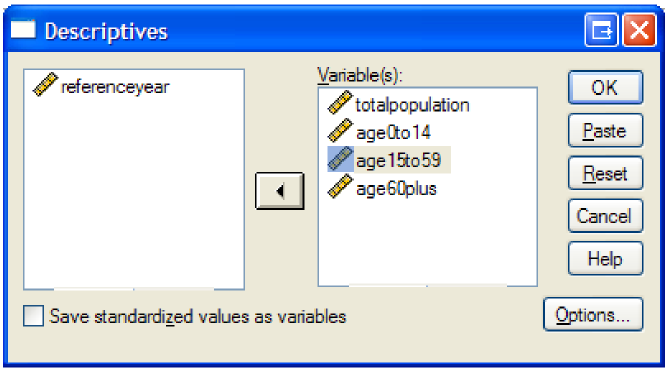
Case Study

The data is drawn from 18 countries. The variables show the percentage of the population in each country that is in each of three age groups (0 to 14, 15 to 59, 60 and above). Imagine we wanted to get some descriptive statistics around these variables.

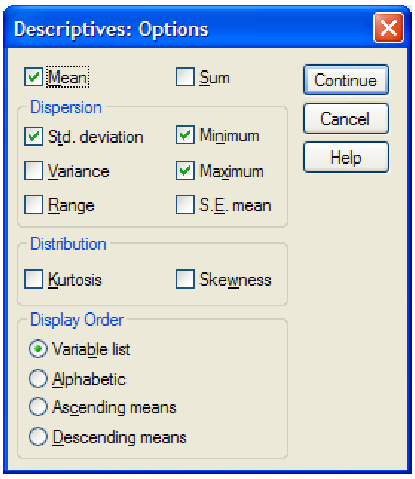
Running

Analyze >> Descriptive Statistics >> Descriptives

Place the variables of interest into 'Variables'. In this case we have three variables representing percentage of population in particular age groups and a total population (in thousands) variable.



Clicking on 'Options' gives you choices about which descriptive statistics you would like produced. The default options shown below are often adequate.



Descriptives: Options

☑ Mean ☐ Sum

Dispersion

☑ Std. deviation ☑ Minimum

☐ Variance ☑ Maximum

☐ Range ☐ S.E. mean

Distribution

☐ Kurtosis ☐ Skewness

Display Order

● Variable list

○ Alphabetic

○ Ascending means

○ Descending means

Output

The output shows the number of valid cases, minimum and maximum values, the mean and the standard deviation for each variable. From this, observations could be made that the smallest country in the sample had only about 4 million people, whereas the largest had 1.3 billion people. The average country in the sample had 16 percent of their population over 60 years of age.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
totalpopulation	18	4028000	1315844000	133024889	305853118.4
age0to14	18	14	40	21.11	7.070
age15to59	18	56	70	63.00	3.630
age60plus	18	4	26	16.00	6.660
Valid N (listwise)	18				

Explore

Overview

The SPSS tool called 'Explore' serves a number of purposes. It provides the same summary statistics as 'Descriptives' and 'Frequencies'. In addition, it provides additional estimators of central tendency that are robust to non-normality and outliers. It can be used to identify outliers in your data set. It also has graphing procedures such as box-plots, histograms, and stem and leaf plots. Arguably, its most unique feature is that it can be used to get statistics for groups. Thus, if you are about to run analyses comparing groups, you may want to get measures of central tendency, spread and distributional properties for each group.

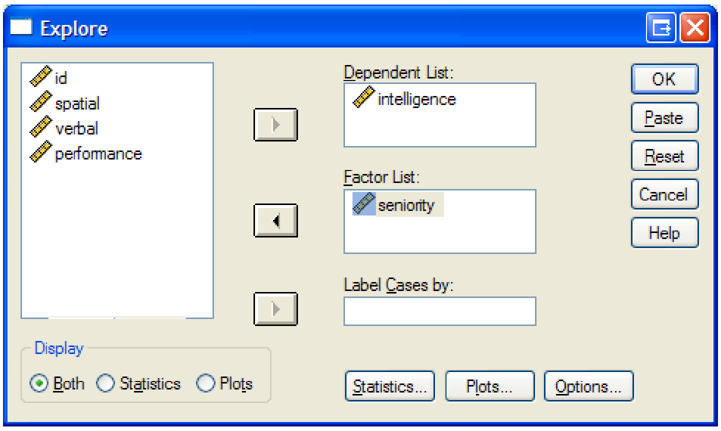
Case Study

Imagine you had collected information about the intelligence of employees in an organisation and some had been classified as high seniority and others as low seniority. You want to know descriptive statistics and the distributional properties for each group.

Running

1. Analyze >> Descriptive Statistics >> Explore

2. Place the grouping variables in 'Factor List' and the outcome variables in 'Dependent List'. It is also possible to still use the procedure without any grouping variables present.



id
spatial
verbal
performance

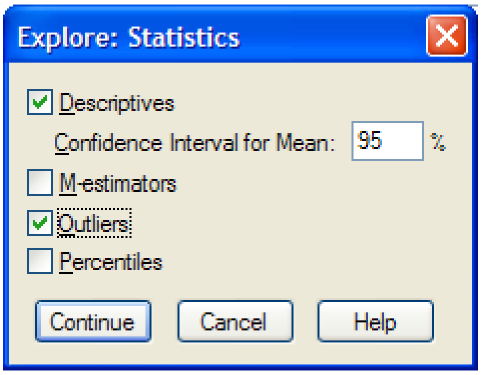
Dependent List: intelligence

Factor List: seniority

Label Cases by:

Display: ● Both ○ Statistics ○ Plots

3. Select the statistics you want. 'Descriptives' is almost always useful. 'M-estimators' give more robust measures of central tendency when there are violations of normality or outliers. 'Outliers' identifies the highest and lowest scores in the data set.



☑ Descriptives

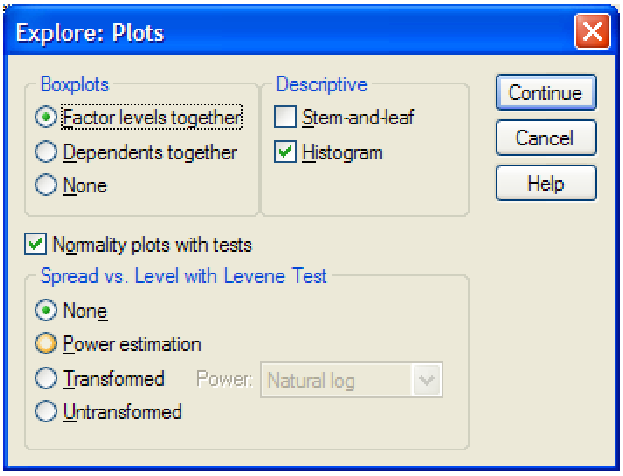
Confidence Interval for Mean: 95 %

□ M-estimators

☑ Outliers

□ Percentiles

4. Select the desired plots. Each of these is beneficial for examining normality and the general distribution. 'Normality plots with tests' are particularly useful for examining the normality assumption.



Boxplots

● Factor levels together
○ Dependents together
○ None

Descriptive

□ Stem-and-leaf
☑ Histogram

☑ Normality plots with tests

Spread vs. Level with Levene Test

● None
○ Power estimation
○ Transformed Power: Natural log
○ Untransformed

Output

The first table sets out the number of cases in each group. Here we have 49 in the high seniority group and 51 in the low seniority group

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	seniority	N	Percent	N	Percent	N	Percent
intelligence	High Seniority	49	100.0%	0	.0%	49	100.0%
	Low Seniority	51	100.0%	0	.0%	51	100.0%

The next table shows summary statistics for each group. For example, you could observe that mean intelligence for the high seniority group was 105.36, and for the low seniority group it was 100.5.

Descriptives

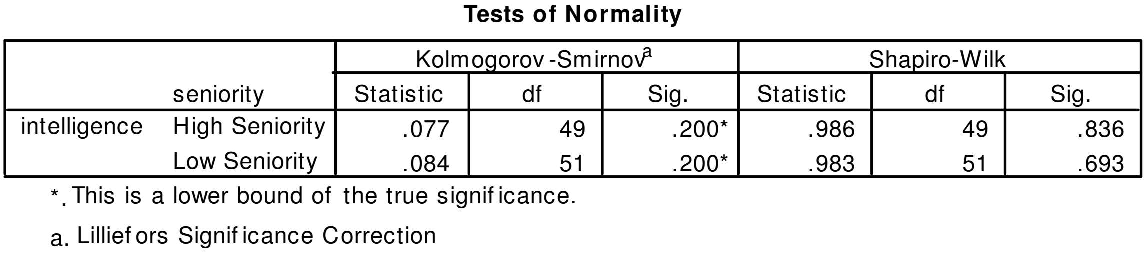
	seniority			Statistic	Std. Error
intelligence	High Seniority	Mean		105.36	2.278
		95% Confidence Interval for Mean	Lower Bound	100.78	
			Upper Bound	109.94	
		5% Trimmed Mean		105.05	
		Median		107.33	
		Variance		254.197	
		Std. Deviation		15.944	
		Minimum		73	
		Maximum		146	
		Range		73	
		Interquartile Range		21	
		Skewness		.175	.340
		Kurtosis		-.016	.668
	Low Seniority	Mean		100.50	2.287
		95% Confidence Interval for Mean	Lower Bound	95.91	
			Upper Bound	105.09	
		5% Trimmed Mean		100.59	
		Median		98.53	
		Variance		266.763	
		Std. Deviation		16.333	
		Minimum		65	
		Maximum		137	
		Range		72	
		Interquartile Range		24	
		Skewness		.006	.333
		Kurtosis		-.632	.656

This table shows the cases with the highest and lowest scores in each group. This can be useful for identifying errors in data entry. It may also identify cases that should not be included in analysis, because they are too unusual. In the present example, we see that the most intelligent person in the high seniority group had an intelligence of 146.

Extreme Values

	seniority			Case Number	Value
intelligence	High Seniority	Highest	1	99	146
			2	33	138
			3	56	133
			4	21	128
			5	95	126
		Lowest	1	76	73
			2	70	77
			3	53	79
			4	71	80
			5	86	84
	Low Seniority	Highest	1	61	137
			2	55	128
			3	48	123
			4	34	123
			5	74	123
		Lowest	1	11	65
			2	6	71
			3	45	74
			4	87	76
			5	27	78

This table displays two different tests of whether the data significantly departs from normality. It is generally better to focus on the plots, because in small samples this test may not be powerful to detect violations of normality and in large samples, it may detect small violations from normality that do not adversely affect your statistical tests.



Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	seniority	Statistic	df	Sig.	Statistic	df	Sig.
intelligence	High Seniority	.077	49	.200*	.986	49	.836
	Low Seniority	.084	51	.200*	.983	51	.693

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

You then get a range of graphs (Histogram, tests of departure from normality). This output has been abbreviated for the sake of space.

5. GRAPHS

Overview

Graphs can be a powerful and efficient way of communicating information to your audience. The graph chosen depends on a number of factors, including the kinds of variables being graphed (nominal, ordinal interval) and the questions being asked.

SPSS has a number of ways of bringing up graphs. It has a Graph menu. This allows you to select the graph you wish to run. It has two ways of running graphs. The traditional way will be shown here, but be aware that there is also a graphing module called "interactive graph". This allows you to set up your graph and change features in real-time. SPSS also has graphing procedures distributed across its main analysis modules.

Histograms

Theory

Histograms are used to show the distribution of scores for continuous variables. Histograms groups scores into discrete blocks and show the count or percentage of cases obtaining that particular block. For example, age is a continuous variable, which could be grouped into blocks of 20 years (0-19; 20-39; 40-59; 60-79).

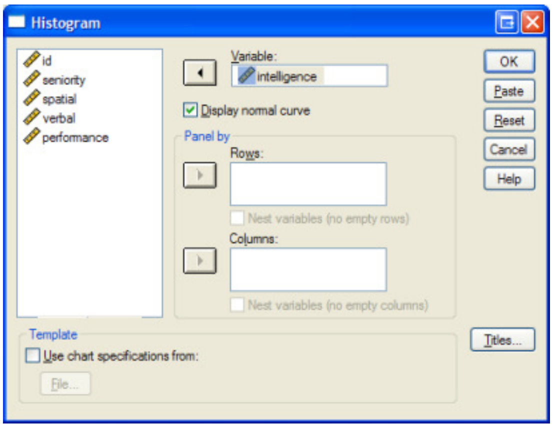
SPSS

Running

Graphs >> Histogram

Place variable into 'variable box'

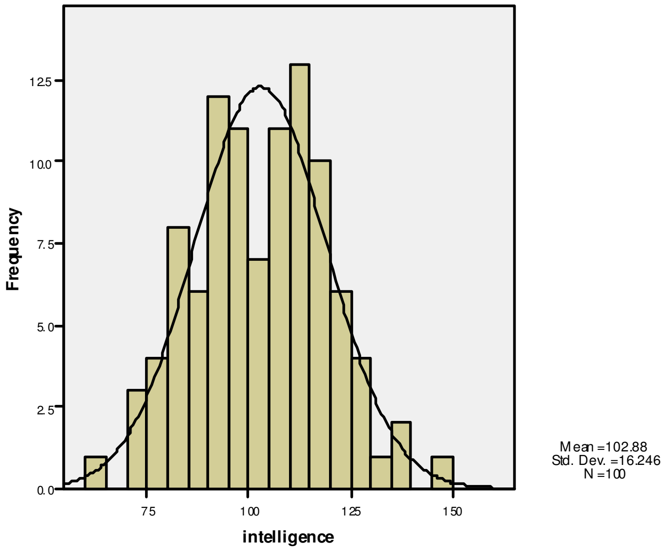
Sometimes it is desirable to overlay a normal curve, if you are assessing the variable's normality.



Variable: intelligence

☑ Display normal curve

Output



Output shows the frequency for different ranges of intelligence score. An examination of the graph shows that it is relatively normally distributed as the raw distribution matches closely to the normal curve. Mean, standard deviation and sample size ('N') are also displayed.

Bar Charts

Theory

Bar charts can be used for a range of purposes. They are effective at presenting percentages and counts for data with discrete categories, such as ordinal and nominal data. Bar charts can also be used to compare means between groups.

SPSS

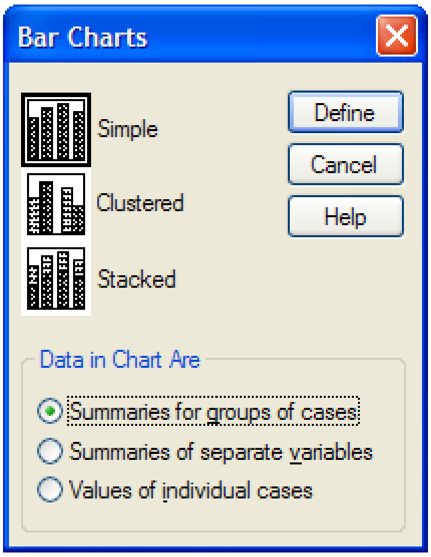
Running

Graphs >> Bar

Choose between simple, clustered and stacked bar charts.

Choose between the options for what the data in the chart will be.

A common option to choose is: Simple – Summaries for groups of cases: This will produce the frequencies or percentages for a set of categories on a single variable. It can also be used to produce the means on a second variable for each level of a categorical variable.



Data in Chart Are

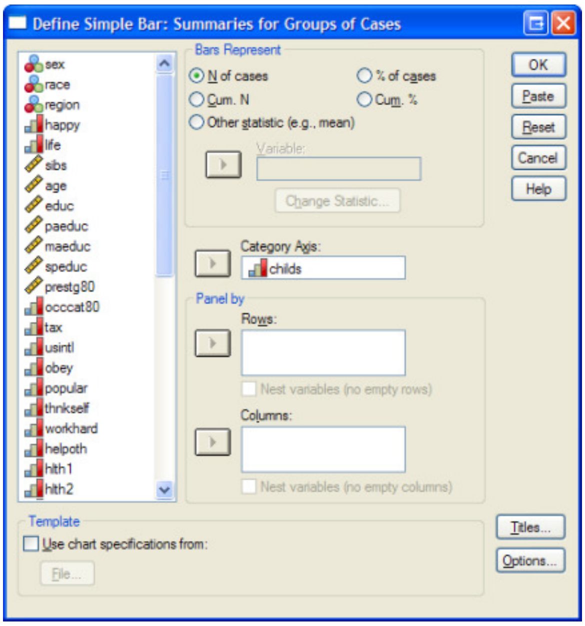
● Summaries for groups of cases

○ Summaries of separate variables

○ Values of individual cases

Place the categorical variable into ‘Category Axis’

Select what you want the bars to represent. The default setting is ‘N of cases’, which will provide frequency counts. ‘% of cases’ is also useful if percentages would be a more meaningful metric. At other times it may be useful to report the mean (or some other summary statistic such as the median) of some other variable for each of the categories. In the first example below we will obtain a bar chart of the number of people reporting having various amounts of children. In the second example, we get the average years of education for each of the categories of number of children.



Bars Represent

● N of cases ○ % of cases

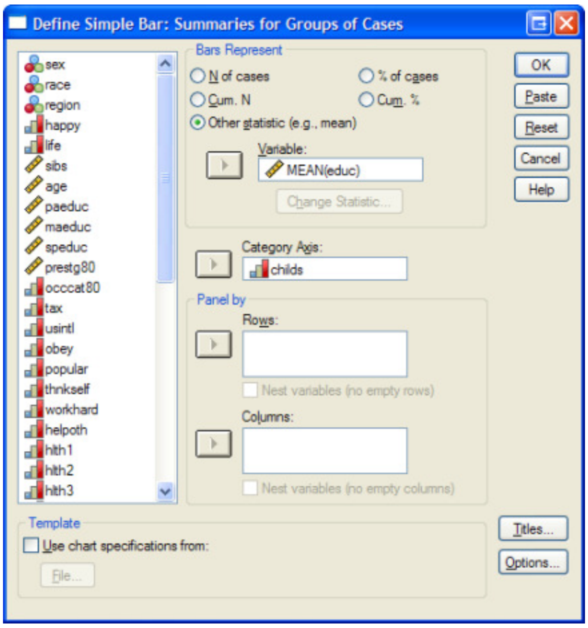
○ Cum. N ○ Cum. %

○ Other statistic (e.g., mean)

Variable:

Category Axis: childs

□ Use chart specifications from:



Bars Represent

○ N of cases ○ % of cases

○ Cum. N ○ Cum. %

● Other statistic (e.g., mean)

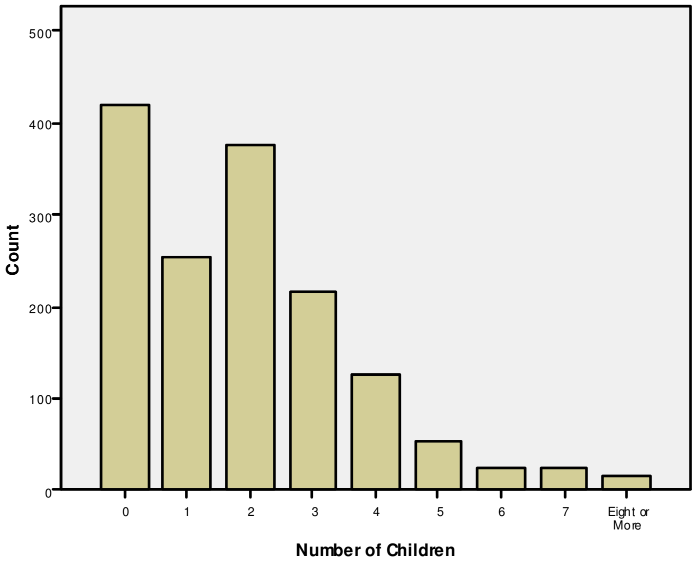
Variable: MEAN(educ)

Category Axis: childs

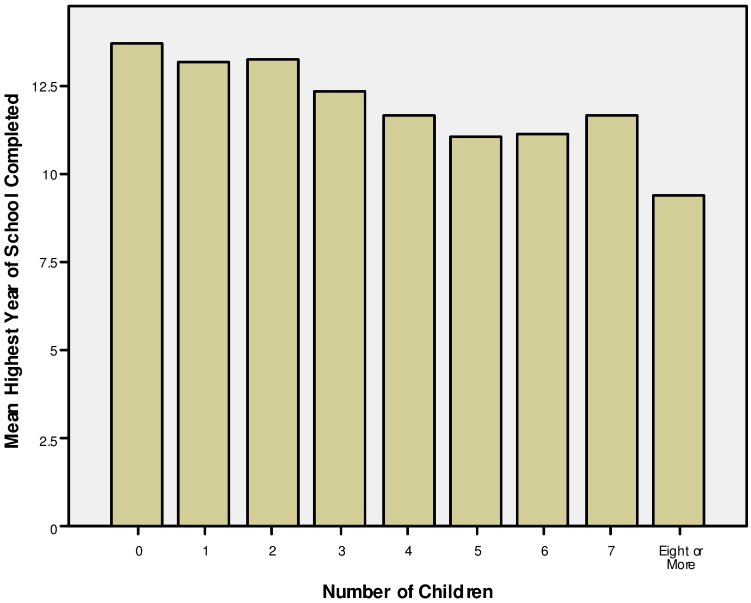
□ Use chart specifications from:

Output

In the output below we first see the frequency counts represented in a bar chart of the frequencies in the sample of number of children. The bar chart quickly shows that most of the sample has between 0 and 4 children. ‘No children’ is the most common response. Of the people with children, two is the most common. The bar chart also allows the determination of the raw frequencies. We can also see that the frequency counts for each category are quite large (e.g., over 400 in the 0 category).



In the second graph we see the mean number of years of high school education for each count of number of children. It would suggest a trend whereby people with more children have had slightly less education on average.



Line Charts

Theory

Line charts can be used for a range of different purposes. They can be used in many of the same ways as bar charts. They can be used to show frequency and count information for particular values of a categorical or ordinal variable. They can also be used to show summary statistics, such as the mean, on a second variable across the levels of another variable. They can be particularly good for showing summary statistics when there are two or more categorical grouping variables (e.g., sales revenue in different locations and in different product lines).

Line charts are also particularly good for showing changes over in a variable over time. Plotting stock prices, sales, number of customers, and other variables over time can be very useful for exploring trends over time and examining seasonal cycles.

SPSS

Running

Graphs >> Line

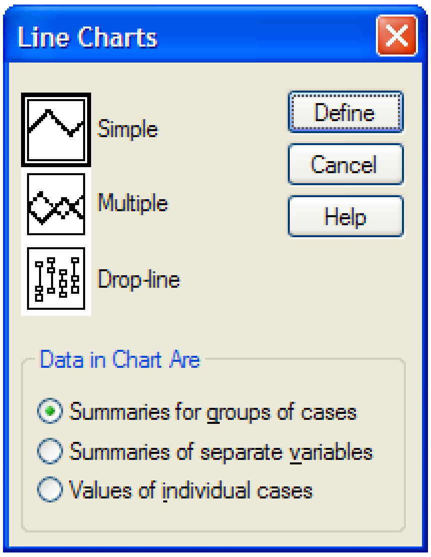
Choose the type of chart: Simple, multiple, drop-line

Choose what the data in the chart is

Common selections would include:

Simple – Summaries for groups of cases: This can be used to show the frequencies, percentages or summary statistics for a second variable across the categories of categorical variable.

Simple – Values of individual cases can be used to plot data for individual cases. This is particularly useful when you each row is a time point in some time-series.



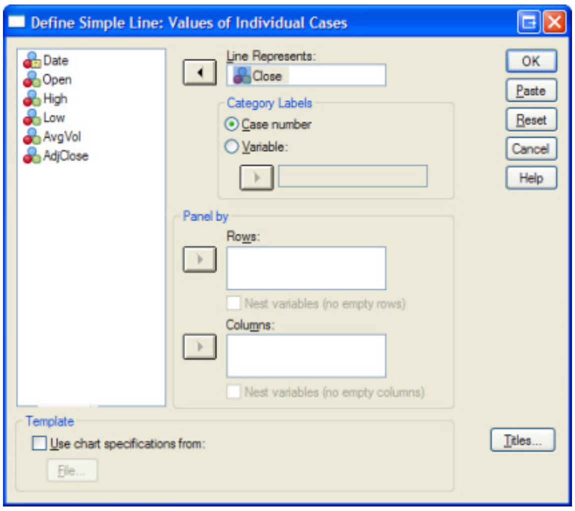
Data in Chart Are

● Summaries for groups of cases

○ Summaries of separate variables

○ Values of individual cases

This example is based on a data file of Google's stock price from August 2004 to May 2006. Each row of the data file is a month. If we wanted to plot the stock price in a graph, we would select: Simple – Values of individual cases. Then in the next dialog box, we place the variable representing the stock price (in this case, it is called 'close') in the 'Line Represents' box.



Line Represents: Close

Category Labels

● Case number

○ Variable:

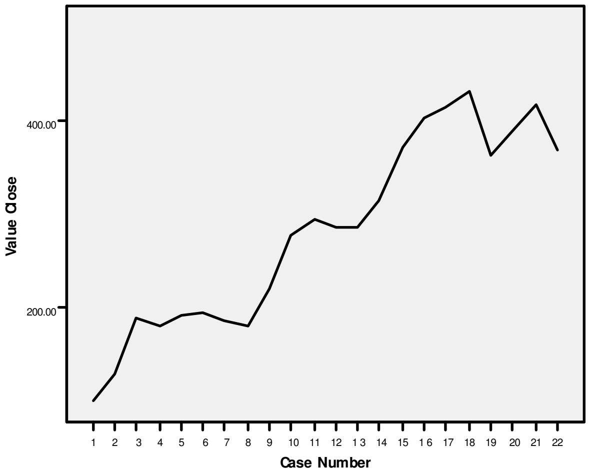
□ Nest variables (no empty rows)

□ Nest variables (no empty columns)

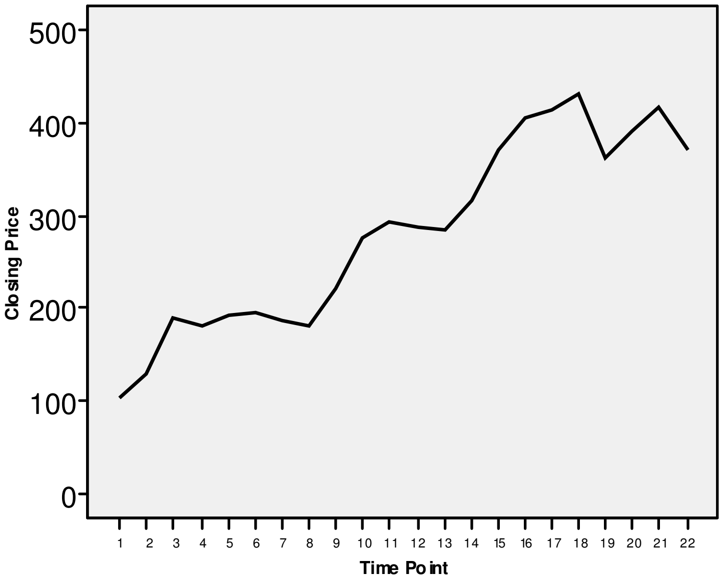
□ Use chart specifications from:

Output

The output shows the way Google's stock price has increased over time but has also gone through periods of stability.



It should be noted that the graph is somewhat misleading in that the x axis does not cross the y axis at zero. By Double clicking on the graph, it is possible to change a range of characteristics of the graph. By selecting 'Edit >> Y-axis' we could adjust the scale of the graph making the minimum value '0' and increasing the number of increments to every 100. By double clicking on the Y and X axis labels we could change the labels to something more appropriate. Selecting the Axis it is also possible to change the font size and the number of decimal places shown. The options are fairly self-explanatory. It is a good strategy to just explore the different options. SPSS allows for substantial customisation of graphs to suit your particular purpose. Below we see the result of the graph after some customisation.



Pie Charts

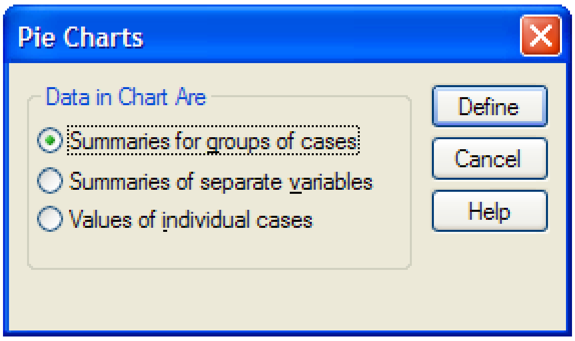
Theory

Pie charts are used for data with a limited number of categories, typically nominal or ordinal data to show the relative percentage of each category. The size of a segment of the chart reflects its percentage size.

SPSS

Graphs >> Pie

Choose what the data represents: Summaries for groups of cases would be the most common option



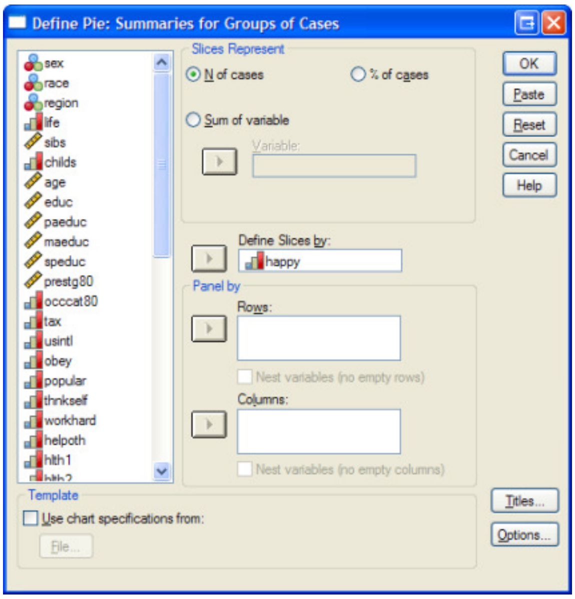
Pie Charts

Data in Chart Are

● Summaries for groups of cases
○ Summaries of separate variables
○ Values of individual cases

In the U.S. Social Survey data file, participants were asked to rate their general happiness (very happy, pretty happy, not too happy).

The variable of interest (happiness) goes into the 'define slices by' box.



Define Pie: Summaries for Groups of Cases

Slices Represent

● N of cases ○ % of cases
○ Sum of variable
Variable:

Define Slices by: happy

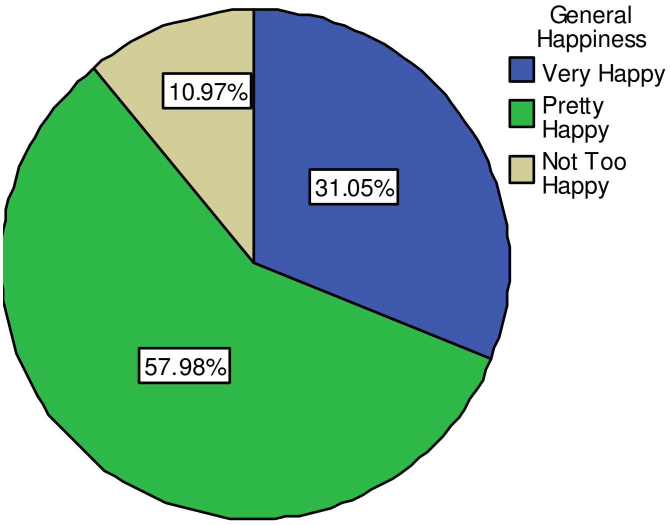
Panel by
Rows:
□ Nest variables (no empty rows)
Columns:
□ Nest variables (no empty columns)

Template
□ Use chart specifications from:

This will bring up a pie chart of the relative counts of each category. Double clicking on the graph allows further useful customisation such as changing font sizes. In particular, placing the actual counts or percentages associated with each segment of the pie can be useful. Clicking on 'Elements >> Show Data Labels' will bring up the counts. You can then change the option to display 'percent' and click on 'text style' to increase the font size if required.

Output

The results after some customisation are shown below. It can be seen that a majority of the sample report being 'pretty happy'.



Box Plots

Theory

A box plot is typically used to explore the distribution of one or more continuous variables. The box plot marks a number of points on the distribution. The middle black line represents the median. The two points above and below the median, which define the box, represent the 25th and 75th percentile. The tails of the box plot which extend from the box represent the highest and lowest values within 3 semi-interquartile ranges of the median. Circles represent outliers and crosses represent extreme scores.

Box plots are useful in assess whether a variable is normally distributed and in identifying potential outliers that might be having excessive influence on analyses.

SPSS

Running

Graphs >> Boxplot

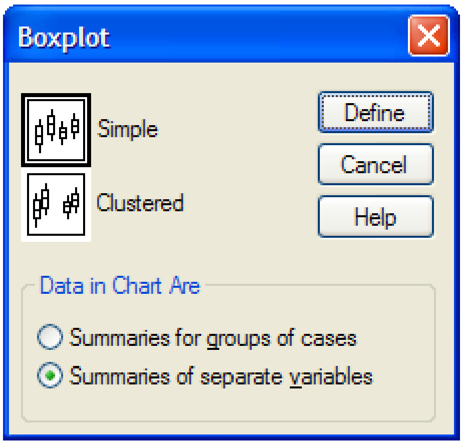
Choose between simple and clustered

Choose what the data in the chart should be

Common options would include:

Simple - Summaries for groups of cases: This enables the creation of box plots on a single variable for each level of a categorical variable.

Simple – summaries of separate variables: This can be used to show box plots for single or multiple variables without a categorical variable.

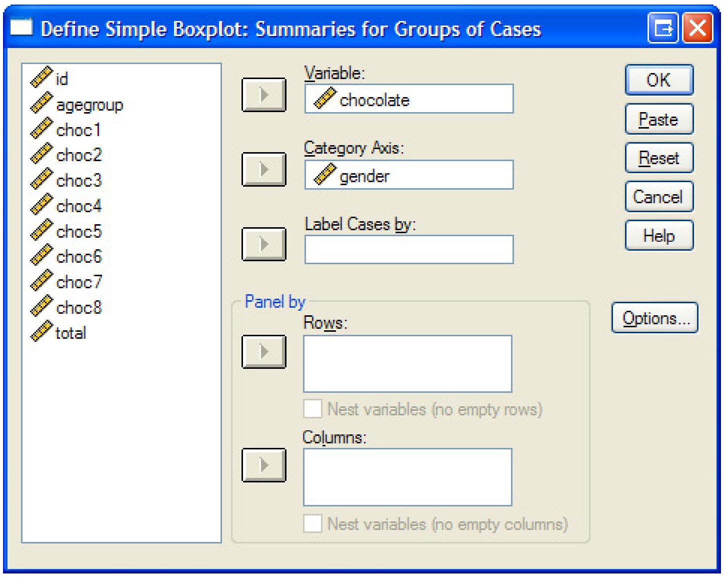

Boxplot

Data in Chart Are

○ Summaries for groups of cases

● Summaries of separate variables

In example one (Simple – Summaries for groups of cases) we examine the distribution of mean ratings of liking for chocolate in males and females. Gender goes into the ‘category axis’ box and chocolate goes into the ‘variable’ box.


Define Simple Boxplot: Summaries for Groups of Cases

Variable: chocolate

Category Axis: gender

Label Cases by:

Panel by

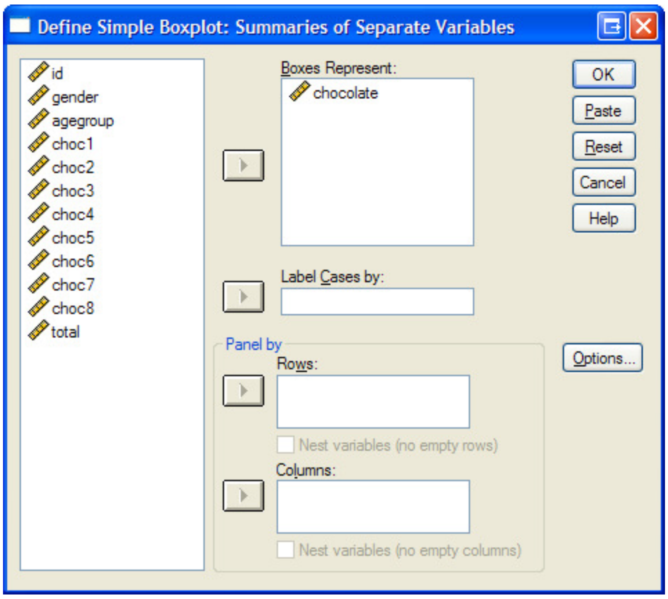
Rows:

□ Nest variables (no empty rows)

Columns:

□ Nest variables (no empty columns)

In the second example (Simple – Summaries of separate variables) we just want the distribution of mean ratings for liking of chocolate without any grouping variable. We place the variable of interest (‘chocolate’) into the ‘boxes represent’ box.


Define Simple Boxplot: Summaries of Separate Variables

Boxes Represent: chocolate

Label Cases by:

Panel by

Rows:

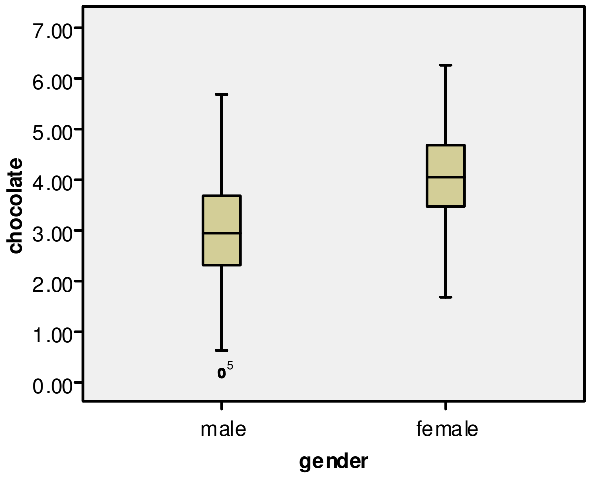
□ Nest variables (no empty rows)

Columns:

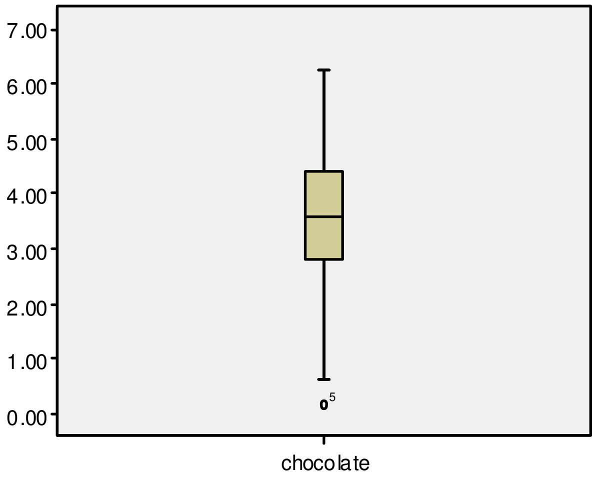
□ Nest variables (no empty columns)

Output

In example one, we see the distribution of chocolate liking ratings for males and females. The median liking rating is higher for females than males. Both variables look relatively normally distributed; the median is in the middle of the box and the tails extend relatively evenly either side of the median. In the male sample there was one outlier with a case number of 5.



In the second example we see the distribution of chocolate ratings for the entire sample.



Scatter plots

Theory

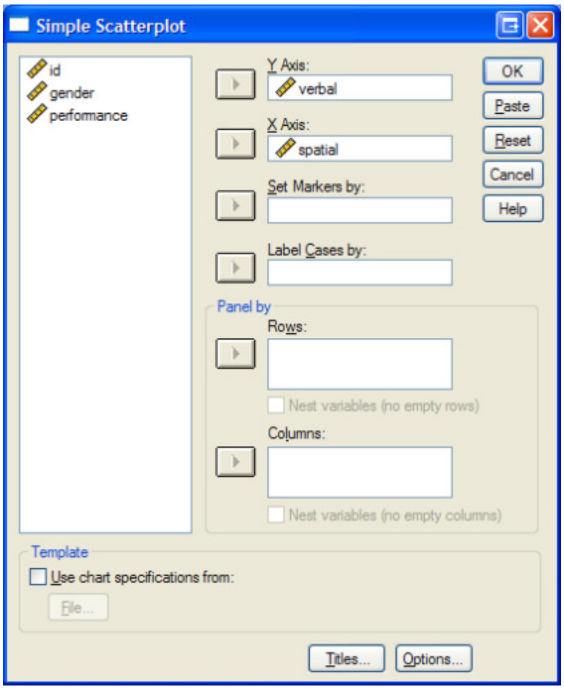
Scatter Plots are used to show the relationship between two continuous variables. They are particularly useful in the context of correlation coefficients. Examination of scatter plots can assist in determining whether a relationship is linear or not. SPSS allows you to attach data labels and colour code data points. SPSS also allows you to plot lines of best fit.

Basic scatter plot

Graphs >> Scatter/Dot

Simple Scatter

Place one variable on X axis and one on Y axis



Simple Scatterplot

Y Axis: verbal

X Axis: spatial

Set Markers by:

Label Cases by:

Panel by

Rows:

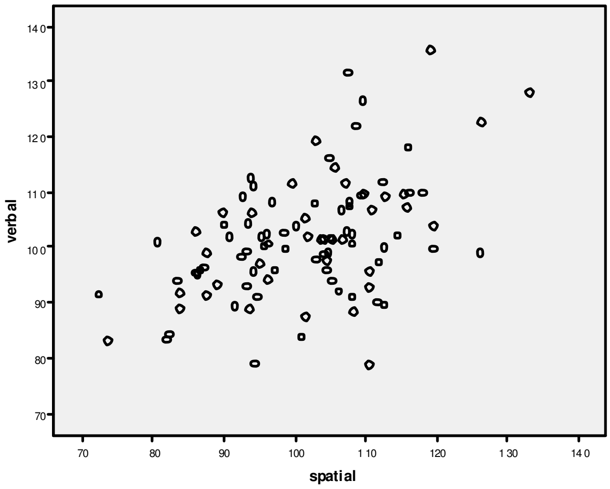
□ Nest variables (no empty rows)

Columns:

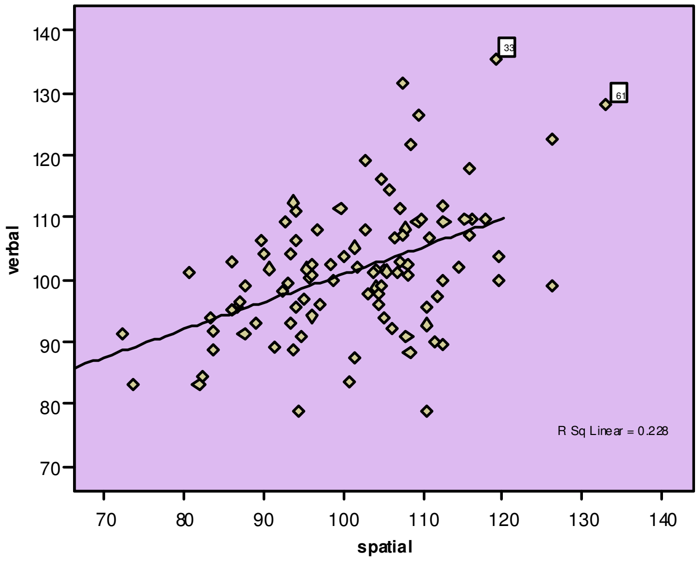
□ Nest variables (no empty columns)

Template

□ Use chart specifications from:

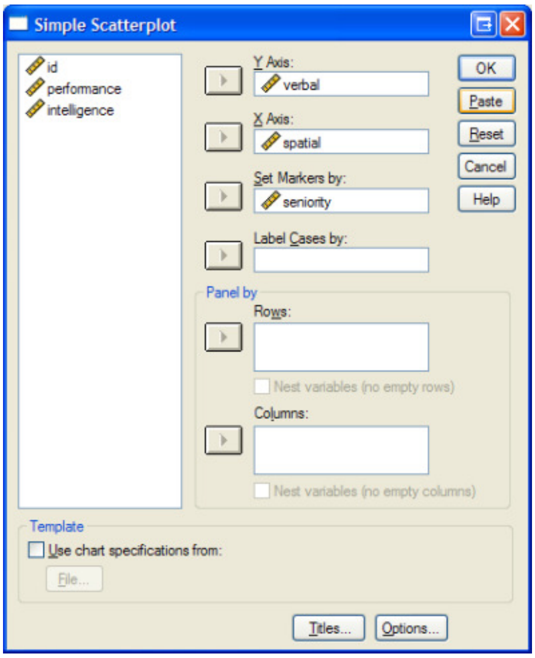


Double clicking on the graph allows you to customise the graph in a number of ways. You can add a line of best fit (Elements >> Fit Line at Total). Identify the case number of a particular data point (Elements >> Data Label Mode; then click on point). You can also modify the appearance of the points, the axis labels and tick points. In the plot you see an example where a line of best fit has been added, two cases of interest have been highlighted, the background colour has been changed, and the font size of the axis points has been increased.



Scatter plot with group markers

Imagine you wanted to see the relationship between two or more groups. In the present example we explore the relationship between intelligence and performance in employees with high and low seniority. SPSS allows you to mark certain groups based on a grouping variable. The process is the same as with a standard scatter plot except that you add a variable to "Set Markers by" which indicates the grouping variable.

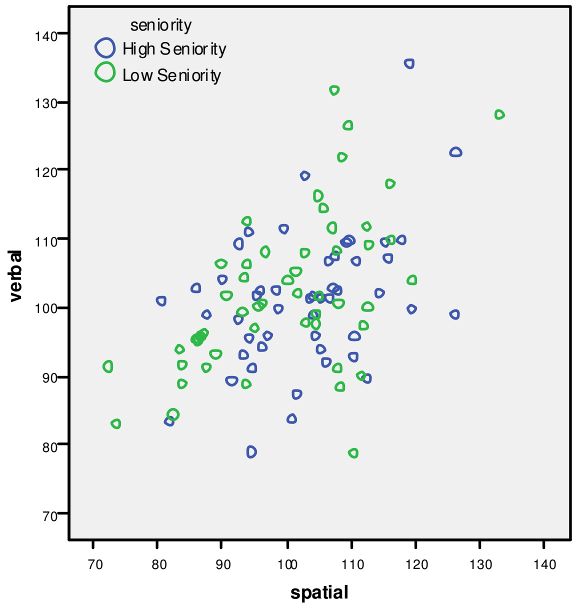


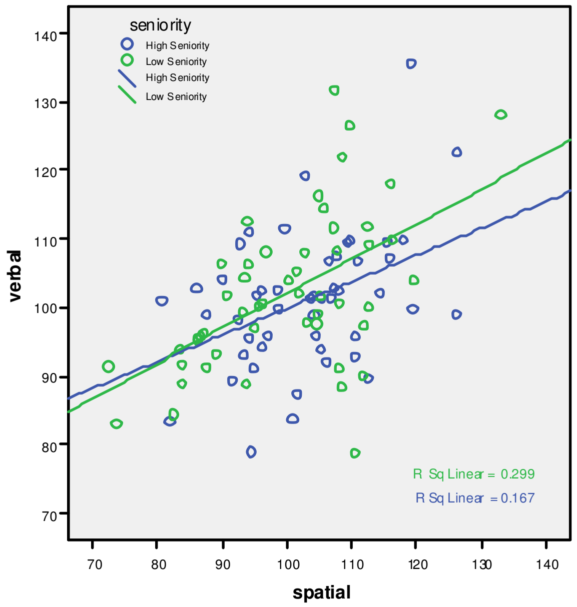
Y Axis: verbal

X Axis: spatial

Set Markers by: seniority

Label Cases by:





Scatter plot with data labels

Imagine that we wanted to see the relationship between the percentage of young people and the percentage of old people in selected countries. The following data is taken from the United Nations 2005, World Population Prospects: The 2004 Revision, http://www.un.org/esa/population/publications/wpp2004wpphighlightsfinal.pdf> accessed 31 March 2005.

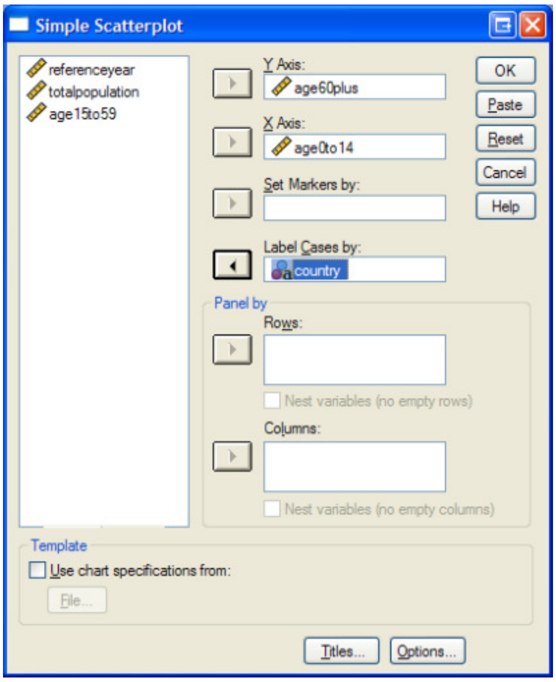
In this context we are actually interested in the individual data points as well as the overall pattern. Imagine we wanted to see the relationship between 'percentage of 0 to 14 year olds' with 'percentage of people over 60' in different countries.

Graph >> Scatter/Dot

Simple Scatter

Place the variable that contains the written names of the country in 'label cases by'

Click 'Options' and Select 'Display chart with case labels'


Simple Scatterplot

referenceyear
totalpopulation
age15to59

Y Axis: age60plus

X Axis: age0to14

Set Markers by:

Label Cases by: country

Panel by

Rows:

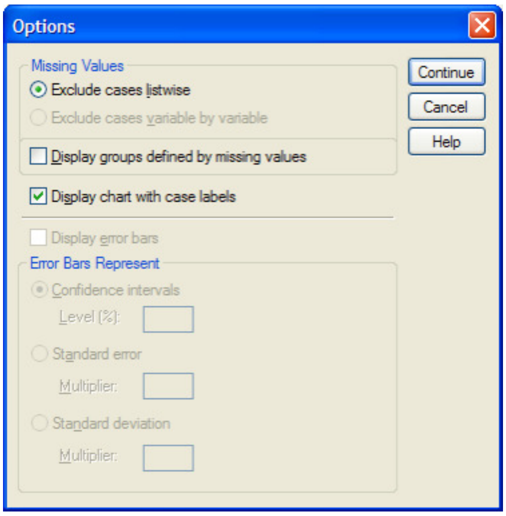
□ Nest variables (no empty rows)

Columns:

□ Nest variables (no empty columns)

Template

□ Use chart specifications from:


Options

Missing Values

● Exclude cases listwise

○ Exclude cases variable by variable

□ Display groups defined by missing values

☑ Display chart with case labels

□ Display error bars

Error Bars Represent

● Confidence intervals

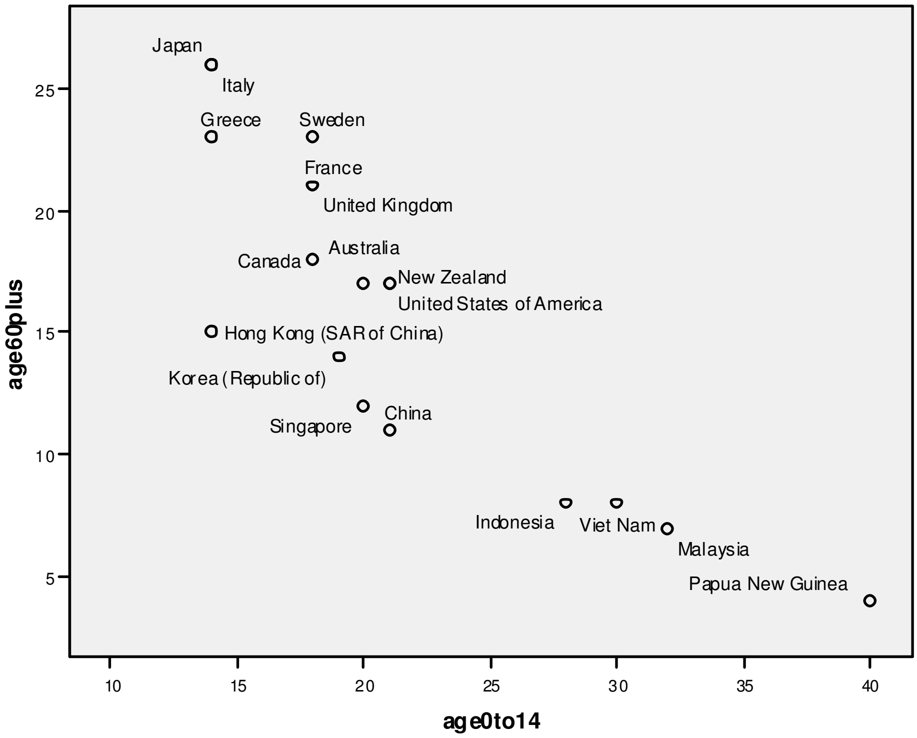
Level (%):

○ Standard error

Multiplier:

○ Standard deviation

Multiplier:



Matrix scatter plot

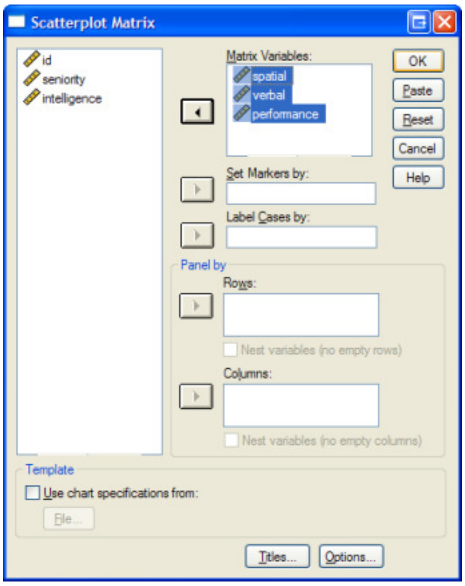
You can also get scatter plots for more than two variables. The output is in the form of a matrix of scatter plots. This is very useful for exploring the relationship between a set of variables.

Graphs >> Scatter / Dot

Select Matrix Scatter

Place all variables into “Matrix Variables”

In the example here, two measures of ability were entered with a measure of performance. The matrix scatter plot shows that there is a positive relationship between all the variables.



Scatterplot Matrix

id
seniority
intelligence

Matrix Variables:
spatial
verbal
performance

Set Markers by:

Label Cases by:

Panel by

Rows:

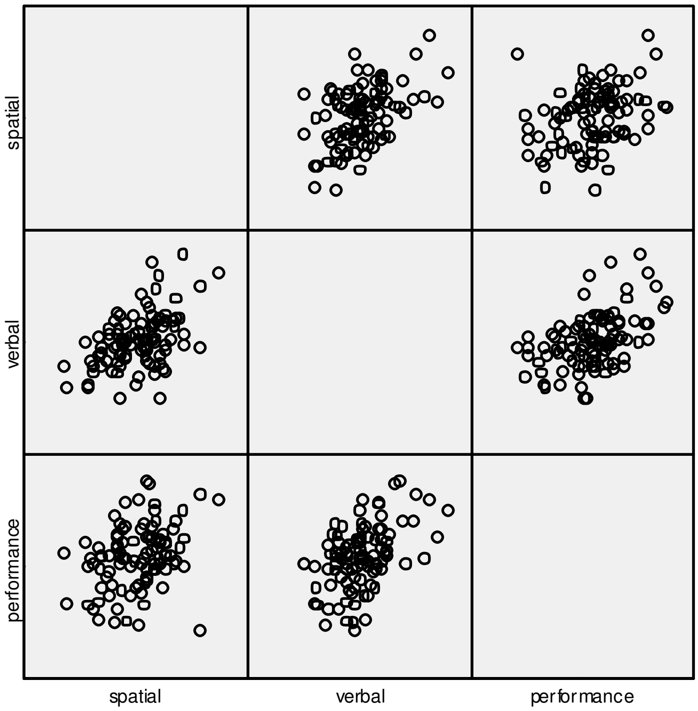
□ Nest variables (no empty rows)

Columns:

□ Nest variables (no empty columns)

Template

□ Use chart specifications from:



6. TABLE CREATION AND INTERPRETATION

Overview

Frequently, your task will be to present a table of values in a report. The default format that SPSS displays is not generally the final format you want to present to your audience. SPSS provides a number of tools to customise the formatting of your tables. These tools are generally accessed by double clicking on the table.

Pivot Trays

General

'Pivot trays' is a tool in SPSS that allows the re-arranging of table elements to improve presentation. Common tasks include re-arranging rows and columns, and hiding certain information in layers.

SPSS

Hiding Data in Layers

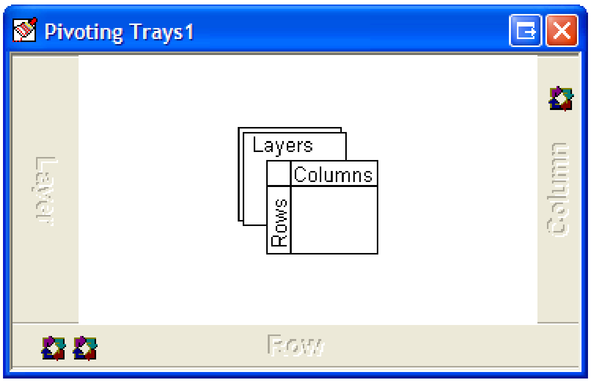
If you obtain a correlation matrix between a set of variables, you will be presented with correlations, significance levels and sample size. Often is sufficient and easier to read, if you just show the correlations. In the example below, correlations between four items related to liking chocolate are presented. This was obtained using 'Analyze >> Correlate >> Bivariate'. The table appears initially as shown below, but what if we wanted to show only Pearson's correlation coefficient.

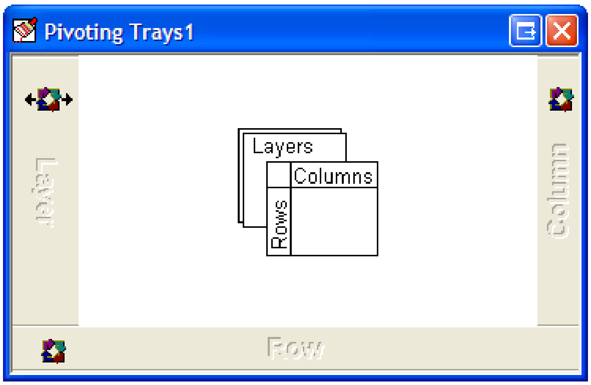
Correlations

		I like the taste	I like the smell	I like the variety	I like the advertising
I like the taste	Pearson Correlation	1	.550**	.550**	.575**
	Sig. (2-tailed)		.000	.000	.000
	N	250	250	250	250
I like the smell	Pearson Correlation	.550**	1	.549**	.503**
	Sig. (2-tailed)	.000		.000	.000
	N	250	250	250	250
I like the variety	Pearson Correlation	.550**	.549**	1	.544**
	Sig. (2-tailed)	.000	.000		.000
	N	250	250	250	250
I like the advertising	Pearson Correlation	.575**	.503**	.544**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	250	250	250	250

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

If we double click on the table, the 'Pivot' menu will appear. Selecting 'Pivot >> Pivot Trays' brings up the pivot trays. Each arrowed square represents an element in the table. We can then drag and drop these elements into different parts of the table, such as layer, row or column.





The result of this movement is that the different statistics get hidden in separate layers and we see only the information we want to see.

Correlations

Pearson Correlation

	I like the taste	I like the smell	I like the variety	I like the advertising
I like the taste	1	.550**	.550**	.575**
I like the smell	.550**	1	.549**	.503**
I like the variety	.550**	.549**	1	.544**
I like the advertising	.575**	.503**	.544**	1

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

Transposing rows and columns

Imagine we were examining to see whether all responses to a series of eight questions on a 1 to 5 scale were within this range. The default output we would obtain using 'Analyze >> Descriptive Statistics >> Frequencies' and selecting minimum and maximum would look like this:

Statistics

		I like the taste	I like the smell	I like the variety	I like the advertising	chocolate makes me feel good	I like the look of the chocolate	I would recommend this chocolate	this chocolate is good value
N	Valid	250	250	250	250	250	250	250	250
	Missing	0	0	0	0	0	0	0	0
Minimum		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Maximum		5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00

While this works okay for eight variables as the number of variables increases it might be better to have the questions forming the rows and the statistics forming the columns. To do this we can double click on the table and select 'Pivot >> Transpose Rows and Columns'. The result is the table below. It contains the same information but may sometimes be a preferred format for presentation.

Statistics

	N			
	Valid	Missing	Minimum	Maximum
I like the taste	250	0	1.00	5.00
I like the smell	250	0	1.00	5.00
I like the variety	250	0	1.00	5.00
I like the advertising	250	0	1.00	5.00
chocolate makes me feel good	250	0	1.00	5.00
I like the look of the chocolate	250	0	1.00	5.00
I would recommend this chocolate	250	0	1.00	5.00
this chocolate is good value	250	0	1.00	5.00

Cell Formats

It is possible to adjust a range of formatting settings of cells, including font type, alignment, number of decimal places, and number format.

These options are available if you double click on a table and highlight the cells you want to format, and then go to ‘Format >> cell properties’. In a particular, it is often preferable to have fewer decimal places than SPSS provides by default.

You can edit or delete cell information in SPSS by double clicking on the cell.

If you want to delete an entire row or column, you need to hold down the ‘Alt’ key and click on the row or column label.

Table Formats

Double clicking on a table allows for further editing. Going to ‘Format >> Table Properties’ allows you to format things like font size, text alignment, background colour, and border formatting. Under ‘Format >> Table Looks’ there are a range of templates that you can use to adjust all of these features at once. Using ‘TableLooks’ it is possible to save your table customisations as a template and reapply the template to other tables in the future.

Interacting with Excel and Word

SPSS provides a number of ways to format tables. Nonetheless, there is substantial overlap between the capacities of Word, Excel and SPSS. If you prefer, you can paste the table from SPSS into Word or Excel and adjust the cell and table formatting in these programs.

In particular Excel is good at setting number formats, borders and row and column widths. Word is good at adjusting the table to fit to the page size.

7. DATA VALIDATION STRATEGY

Overview

Quality control is critical in data analysis. If data becomes severely corrupted, all the time and money spent on collecting, analysing data can amount to nothing. In fact, reporting on corrupted data is often worse than nothing, because it can strengthen beliefs in baseless claims. For these reasons it is critical to have a data validation strategy. The traditional computer science idea of 'garbage in, garbage out' is very relevant here.

Strategies

Basic checking of data entry

If the data has been entered from paper based questionnaires and tests, it is worth taking a subsample of the tests and verifying that the values in the data file are correct. If critical decisions in relation to individual cases, such as when entering exam grades, or entering data that will effect promotions, hiring decisions and similar things, it may be worth considering a double data-entry methodology. In this case, all data is entered twice and checks are performed to verify that have been entered in the same way.

Frequency and range checks

After entering data and setting up your data file in SPSS, it is important to verify that the data is correct. There are many reasons why errors can arise. When doing data entry, errors can be made (e.g., someone type 44 instead of 4).

It is important to examine the minimum and maximum column to verify that all values are within the range of valid values. For example, if you have a 5 point scale and you have a score of '45', you can presume that there was a data entry error. Number of valid responses should be assessed to see that the number of cases with missing data is not smaller or larger than you expect.

Frequency counts should be assessed to see that the frequencies correspond with theoretical expectations. This is particularly the case for nominal data and ordered categorical data with small number of categories (e.g., less than 15).

Additional checks for continuous data

For continuous data means, standard deviations and histograms should be examined. Compare the means to what you would expect. For example if you are sampling people from the adult population, you might expect the mean age to be somewhere between 30 and 45. If the mean age was outside this range, you might want to think why that might be.

Univariate Outliers

A univariate outlier is a case that is particularly high or low on a continuous variable. It is often defined as a score that is more than a certain number of standard deviations away from the mean. Values larger than 2.5 in a small sample or 3.0 in a large sample may be considered as potential outliers. When outliers are encountered, it is worth considering whether they have been produced by a data entry error or some other error.

Bivariate checks

Further checks can be performed by looking at crosstabulations of categorical data and correlations for continuous data. If your variables are expected to be related in particular ways, it is worth seeing whether the particular relationship is present. Relationships may be a theoretical question, but a failure to find a particular relationship may indicate data entry problems. Sometimes you will find that where there should be a positive correlation there is a negative correlation. In this case, the variable may need to be reversed. Some questionnaires have conditional questions such that only participants who answer a particular way to question 1 are asked about question 2. By examining the crosstabulations between question 1 and 2 you can verify that all those and only those people who were meant to respond to question 2 did in fact respond to question 2.

Concluding Comments

The most important rule to follow for maintaining integrity and quality in data integrity is: Compare theoretical expectations to obtained results. Sometimes your expectations will not be correct. But other times such differences may reflect problems with the data. It is important to always be vigilant to this possibility.

An additional benefit of the data validation procedure is that you get to know your data. You get familiar with the basic descriptive statistics and distributional properties of the data. It also forces you to think about what the data would be expected to look like.

SPSS

Case Study

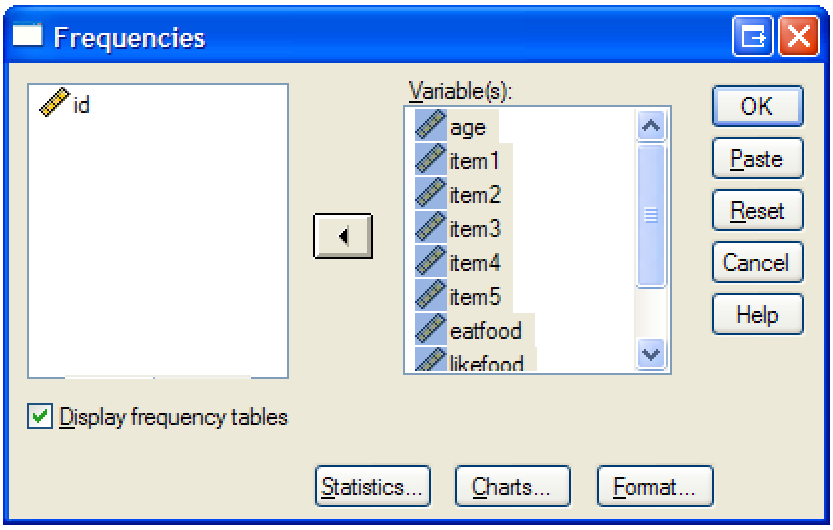
Imagine a study was conducted looking at a number of variables: age, 5 items on attitudes to different types of food rated on a 5 point scale, a question assessing whether they have eaten at a particular restaurant and if so what they think of the restaurant on a 5 point scale. The task is to assess whether the data has been entered in correctly and the data looks valid.

Analyses

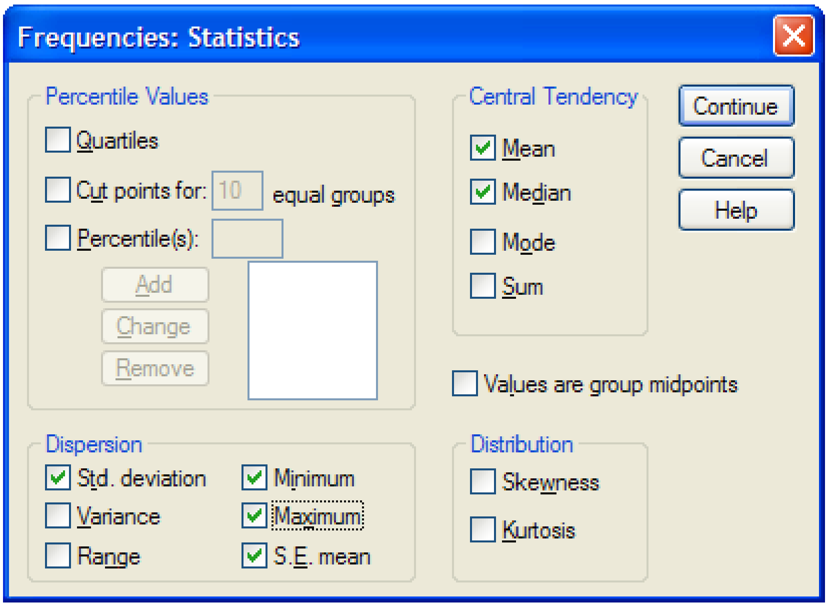
To assess that minimum and maximum values are within the appropriate range, SPSS frequencies can be used.

Analyze >> Descriptive Statistics >> Frequencies

All substantive variables are placed in the 'variable' box



A range of statistics are selected from under the 'statistics' sub-dialog box, the most important being the minimum, maximum, mean and standard deviation. Others might also be of interest. It would also be useful to obtain histograms by clicking on the charts button.



Percentile Values

□ Quartiles

□ Cut points for: 10 equal groups

□ Percentile(s):

Central Tendency

☑ Mean

☑ Median

□ Mode

□ Sum

□ Values are group midpoints

Dispersion

☑ Std. deviation

□ Variance

□ Range

☑ Minimum

☑ Maximum

☑ S.E. mean

Distribution

□ Skewness

□ Kurtosis

After transposing rows and columns, we get the following output. In looking at the minimum and maximum for each of the variables we can see that the value of 99 for an age may be a bit surprising in our sample. Also, item 1 has a maximum of 44, which is larger than the 1 to 5 scale that was being used. It can also be noted that there is a bit more missing data on item3 than the other items.

Statistics

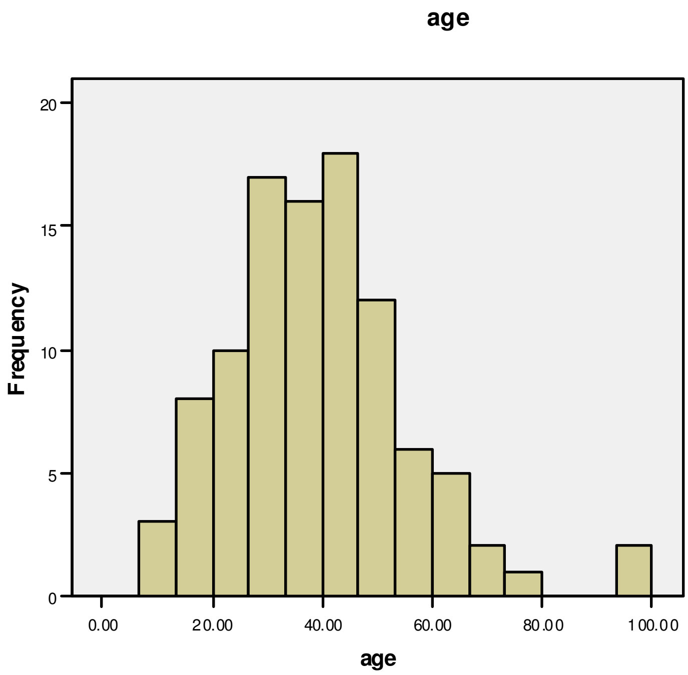
	N		Mean	Std. Error of Mean	Median	Std. Deviation	Minimum	Maximum
	Valid	Missing						
age	100	0	39.3200	1.63583	39.0000	16.35829	9.00	99.00
item1	100	0	3.4600	.47235	3.0000	4.72351	1.00	44.00
item2	100	0	3.2500	.11924	3.0000	1.19236	1.00	5.00
item3	94	6	3.0426	.11411	3.0000	1.10633	1.00	5.00
item4	98	2	3.0816	.13347	3.0000	1.32130	1.00	5.00
item5	98	2	2.9898	.12005	3.0000	1.18839	1.00	5.00
eatfood	100	0	1.5700	.04976	2.0000	.49757	1.00	2.00
likefood	58	42	3.0690	.08821	3.0000	.67179	2.00	4.00

Examination of the item1 frequency distribution shows that there are two cases that are outside the range of valid values. These two values should be removed. It would then be appropriate to go back to the original questionnaires to ascertain what the actual values were.

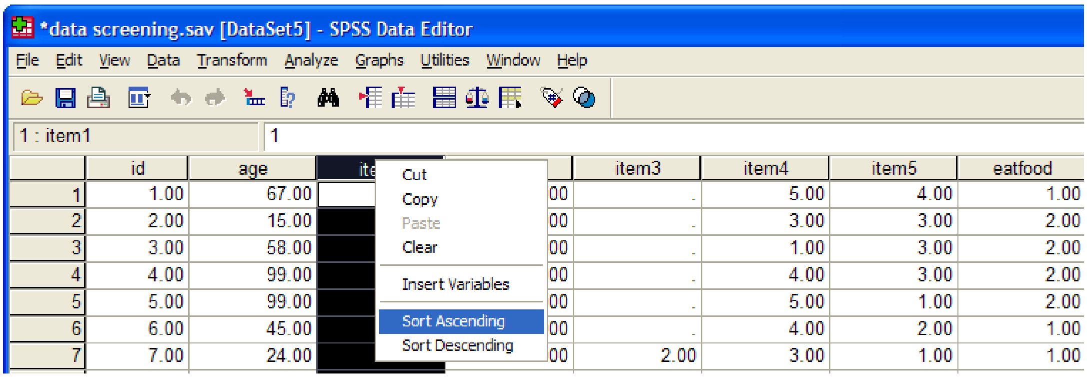
item1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	17	17.0	17.0	17.0
	2.00	24	24.0	24.0	41.0
	3.00	21	21.0	21.0	62.0
	4.00	29	29.0	29.0	91.0
	5.00	7	7.0	7.0	98.0
	23.00	1	1.0	1.0	99.0
	44.00	1	1.0	1.0	100.0
	Total	100	100.0	100.0	

An examination of the age distribution shows 2 cases with values that appear somewhat outside the normal range of scores.



These cases can be identified in the raw data by sorting the variable from highest to lowest. This is done by right clicking the relevant column and clicking sort ascending or sort descending.



1 : item1 | 1

	id	age	ite		item3	item4	item5	eatfood
1	1.00	67.00		00	.	5.00	4.00	1.00
2	2.00	15.00		00	.	3.00	3.00	2.00
3	3.00	58.00		00	.	1.00	3.00	2.00
4	4.00	99.00		00	.	4.00	3.00	2.00
5	5.00	99.00		00	.	5.00	1.00	2.00
6	6.00	45.00		00	.	4.00	2.00	1.00
7	7.00	24.00		00	2.00	3.00	1.00	1.00

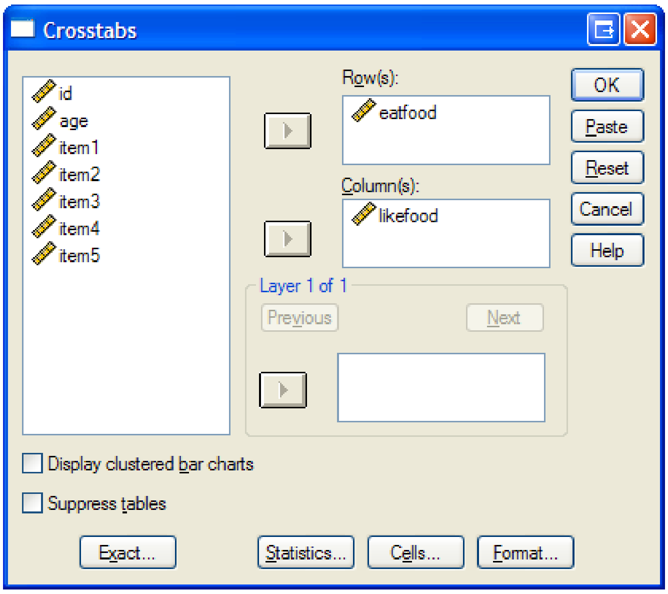
These values can then be deleted.

A second data check would be to verify that the conditional question was effective. The condition was the only people who said yes to the variable ‘eatfood’ would be asked the question on ‘likefood’. Thus, a cross tabulation should show whether this condition was satisfied.

This is obtained in SPSS by going to: Analyze >> Descriptive Statistics >> Crosstabs

Place the condition variable into rows and the answer variable in columns

It is also beneficial to see the output in conjunction with the raw frequencies of the condition variable.



Row(s): eatfood

Column(s): likefood

eatfood * likefood Crosstabulation

Count

		likefood			Total
		2.00	3.00	4.00	
eatfood	1.00 no	0	0	2	2
	2.00 yes	11	32	13	56
Total		11	32	15	58

eatfood

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 no	43	43.0	43.0	43.0
	2.00 yes	57	57.0	57.0	100.0
	Total	100	100.0	100.0	

The output above shows that 57 respondents stated that they had eaten at the restaurant and should be asked the question. However, from the crosstabulation it can be seen that only 56 of the 57 who answered the question actually responded to the question. More concerning is that 2 respondents who had claimed not to have eaten at the restaurant rated how much they liked the restaurant.

These two values should probably be deleted from the data file. It would once again be beneficial to return to the raw data to verify whether it was a data entry or an error in how the interviews were conducted to obtain the data.

8. DEALING WITH MISSING DATA

Overview

Missing data is a common occurrence in data collection. Participants drop out of studies. Machines break down. Pages from questionnaires get lost. Participants forget or refuse to answer particular questions. For these reasons and many more, the final data file can often contain a large amount of missing data. Missing data can present problems for analysis. How should missing data be treated in data analysis?

Dealing with missing data is a relatively advanced topic. This discussion will just begin to deal with assessment and some possible resolutions. None of the solutions are ideal and the most important rule to follow is to try to minimise the occurrence of missing data by using good research design. Online tests can force participants to respond to every item. Questionnaires can be pilot tested to verify that all questions have valid responses. Drop-outs in longitudinal data can be minimised using a range of strategies. The point is that it is difficult to completely resolve missing data issues after a study has been completed.

Assessing Missing Data

The first task is to assess the scope of missing data. By placing all variables in (Analyze >> Frequencies) it is possible to examine the number of missing values for each variable. It is useful to know whether some variables appear to have more missing data than others. It may also be that some variables have missing data because of research design, such as when particular items are only to be answered by a subset of the sample. It would also be appropriate to see if there are any particular cases with more missing data than others. More sophisticated analyses could explore whether there between those with and without missing data for particular items on other important variables.

Missing Data procedures

Overview

Once missing data has been assessed, the analyst must decide what to do with the missing data. The important point to note is that we always do something. Even, if we choose to do nothing, SPSS will choose a particular strategy by default. The question for us is whether this default strategy is the most appropriate. Often it is the simplest but not necessarily the best.

Listwise deletion

Listwise deletion is the standard missing values procedure in SPSS. Listwise deletion removes any case that is missing data on any variable used in a particular analysis.

Pairwise deletion

Pairwise deletion is a missing values procedure which is often an alternative option in such analyses as regression, factor analysis and correlations. It only deletes a case with missing data for the elements of the analysis that rely on that variable. For example, in a correlation matrix, if a case has data on variable X and Y, but not on variable Z, it will be included in correlations between X and Y, but not in correlations between X and Z or Y and Z.

Replace with mean

This procedure replaces the missing data with the mean value for the variable. This procedure is not a particularly respected procedure for missing data replacement as it tends to reduce the variance of the variable. It can be done automatically in certain SPSS procedures such as factor analysis. It can also be used by going to: Transform >> Replace missing values; and only placing one variable in the 'new variable' box.

Replace with series mean

This is a more sophisticated technique than replace with mean. It is appropriate when you have a number of items in a questionnaire that are all measuring the same thing and are on the same scale. This procedure will replace a missing item with the value of the data that is present. This can be performed in SPSS by going to: Transform >> Replace missing values. Then, add the variables that form the set of similar items making sure to specify the method as 'series mean'.

Replace with best guess

This is not a particular sophisticated technique, but sometimes we have enough information about our data to estimate or know what the particular case would have received. This technique is not always appropriate and it depends on the knowledge of the analyst. There is evidence for bias to enter into the data, if this is performed without due care.

Advanced Techniques

Several more advanced techniques exist for replacing missing values. These include regression, EM, and imputation. Regression attempts to make a prediction for the missing data for a particular case based on what values the case has for other variables. EM is similar to regression in that it uses information from other variables to make a prediction, but does this in an arguably more sophisticated way. One form of imputation involves finding a case that closely matches the existing case and replacing the missing value with the value for this closely matching case. These methods are not available in the base module of SPSS. They are available with the SPSS missing data analysis module.

SPSS Missing Data Analysis Module

If you have this add-on module, you are given a number of powerful tools for examining the nature of missing data in your data file and tools for replacing missing data. This course does not assume that you have this add-on module.

9. DATA MANIPULATION

Overview

Converting data from the form it is originally entered into a form that is more appropriate for analysis is a common task in statistics. SPSS provides a range of tools to assist this process. This section outlines some of the most important of these tools with examples of how they are applied.

Transforming variables and creating scales

Theory

There are many times when we want to create scales and transform variables. You may want to create a total or average of a set of other items. You might want to transform a variable. You might to see the difference between two time points.

Examples: getting the average of four questions on job satisfaction to get an overall measure of job satisfaction; asking participants to rate which of a series of stressful events they have experienced and adding that up to form a total number of stressful events; obtaining weekly sales for each store over the year and adding them up to form a year total.

SPSS

Case Study

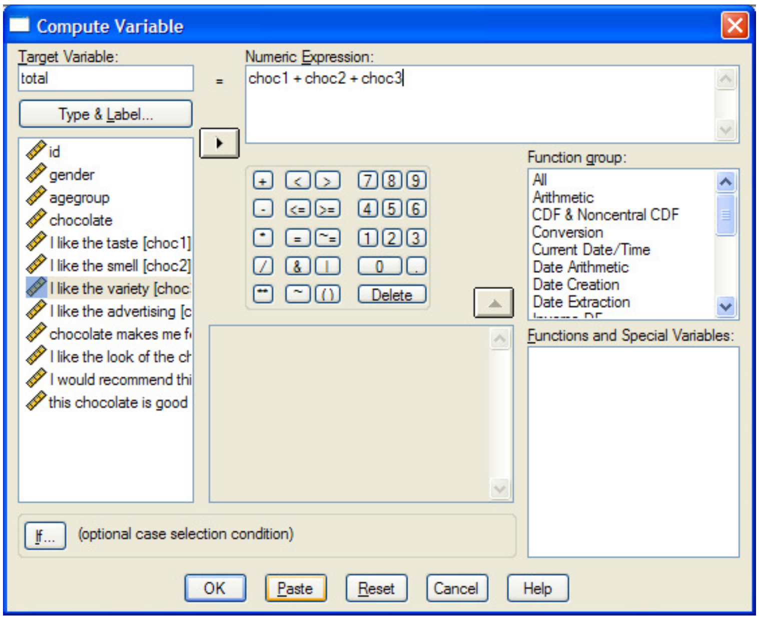
Imagine that we were conducting a market research study on attitudes towards chocolate and we had asked people their agreement with a series of statements about chocolate. Each statement was related to liking the particular brand of chocolate (e.g., I like the taste; I like the smell; I like the variety) on a five-point scale where higher scores indicate more liking. You might decide that you want to create an overall measure of liking for the chocolate. This could be the average of an item or the total of the items.

Running

Transform >> compute

1. Enter the name of the new variable that is going to be created in “Target Variable”. I have called this one ‘total’.

2. Enter the equation that will be used to create the new variable into the field called “Numeric Expression”. You can use standard mathematical symbols such as plus, minus, and multiplication. Usually your new variable will be a function of previous variables. These can be copied in by clicking them on the left and then pressing the rightward arrow. In the present example, I have simply created an equation that adds item 1, 2 and 3 to form total. There are also a number of functions that are grouped into categories. Some of the most important functions are Mean(), sum(), sqrt(), ln(), which give the mean, total, square root and natural log respectively. Each function has a name followed by brackets. Each function takes one or more arguments, which are separated by commas. When you are finished you can click ‘OK’ or you can ‘Paste’ the command into Syntax for future use.

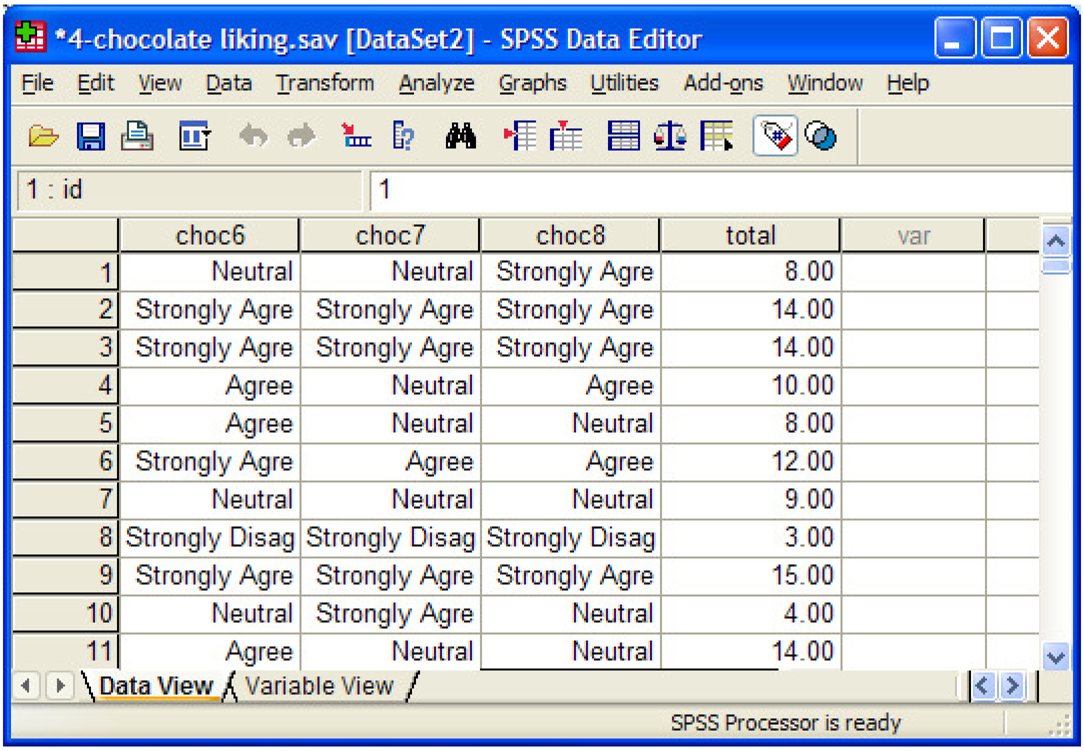


Compute Variable

Target Variable: total

Numeric Expression: choc1 + choc2 + choc3

The result of running these analyses is that a new column will be created in SPSS that contains the values for the transformed variable. We can see this below with the new variable called total.



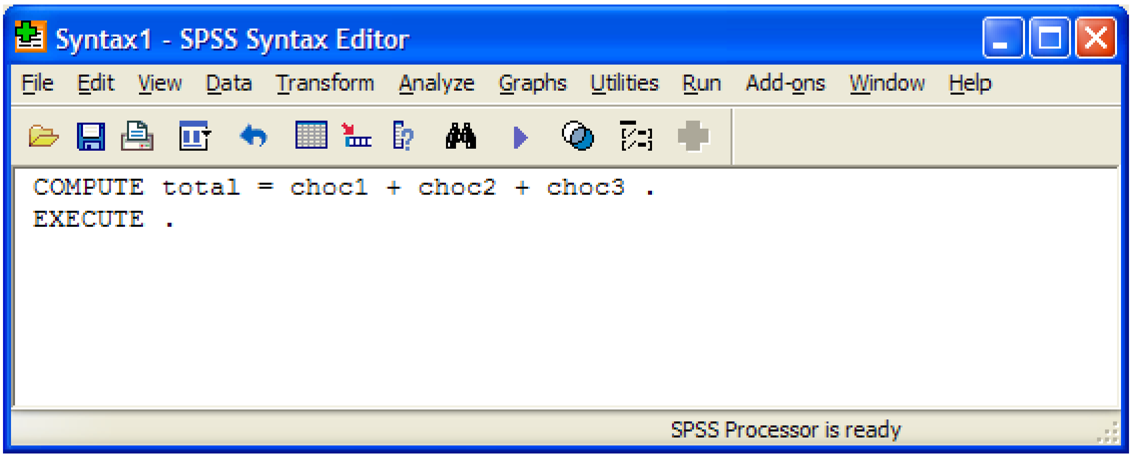
*4-chocolate liking.sav [DataSet2] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : id
1

	choc6	choc7	choc8	total	var
1	Neutral	Neutral	Strongly Agre	8.00	
2	Strongly Agre	Strongly Agre	Strongly Agre	14.00	
3	Strongly Agre	Strongly Agre	Strongly Agre	14.00	
4	Agree	Neutral	Agree	10.00	
5	Agree	Neutral	Neutral	8.00	
6	Strongly Agre	Agree	Agree	12.00	
7	Neutral	Neutral	Neutral	9.00	
8	Strongly Disag	Strongly Disag	Strongly Disag	3.00	
9	Strongly Agre	Strongly Agre	Strongly Agre	15.00	
10	Neutral	Strongly Agre	Neutral	4.00	
11	Agree	Neutral	Neutral	14.00	

There are many benefits to pasting transformations into Syntax. Transformations are something that you often want to re-run at a future time. For example, if you are running analyses and then collect some additional participants, you would need to re-run the transformations to apply them to the new cases. Variables created by transformations do not automatically update themselves when you add new cases or change the variables that were used to create the new variable.

The syntax includes the `COMPUTE` command followed by the name of the new variable, an equals sign and the equation that has been entered. The next line includes an `EXECUTE` command which tells SPSS to perform the transformations. To run this syntax highlight the syntax and click the run button or go to Run >> Selection.


Syntax1 - SPSS Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Run Add-ons Window Help

```
COMPUTE total = choc1 + choc2 + choc3 .
EXECUTE .
```

Some other commands that you might consider running could include

Calculate the average of a set of items

```
COMPUTE meanchoc = mean(choc1, choc2, choc3).
EXECUTE.
```

Calculate the total of a set of items

```
COMPUTE totchoc = sum(choc1, choc2, choc3).
EXECUTE.
```

Reverse an item on a 5 point scale so that 1 becomes 5 and 5 becomes 1. The rule for doing this is that the new score should equal: minimum + maximum – original score.

```
COMPUTE reversedchoc1 = 1 + 5 - choc1.
EXECUTE.
```

Merging data files

Theory

Frequently, you will have two data files on the same participants but with different variables. How do you combine these into a single data file?

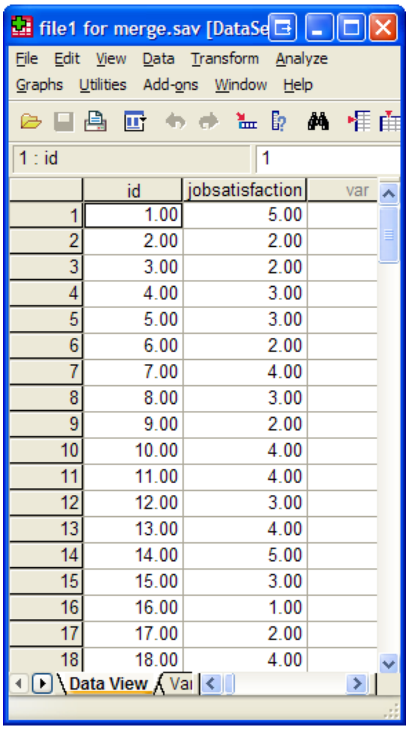
SPSS

Case study

Imagine that you have administered two questionnaires. One measures job satisfaction and another measures leadership. You want to be able to analyse the relationship between job satisfaction and leadership but they are in different data files. You need to be able to merge them together.

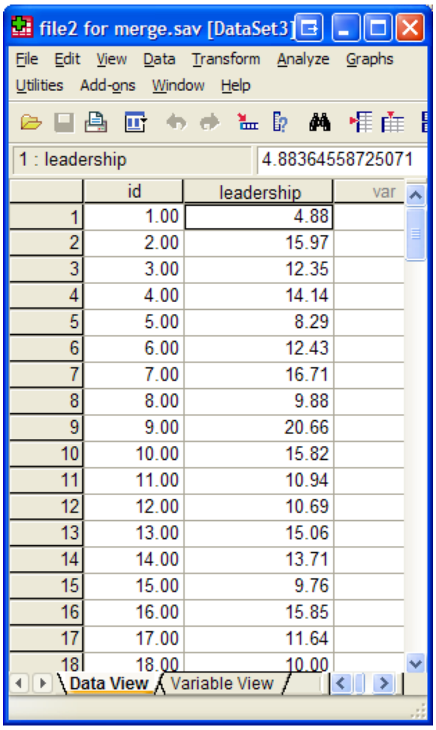
Running

1. Make sure that you have a unique ID variable in each data file. It is important that it has the same variable name in both files.



file1 for merge.sav [DataSe
File Edit View Data Transform Analyze
Graphs Utilities Add-ons Window Help
1 : id
1
Data View
Var

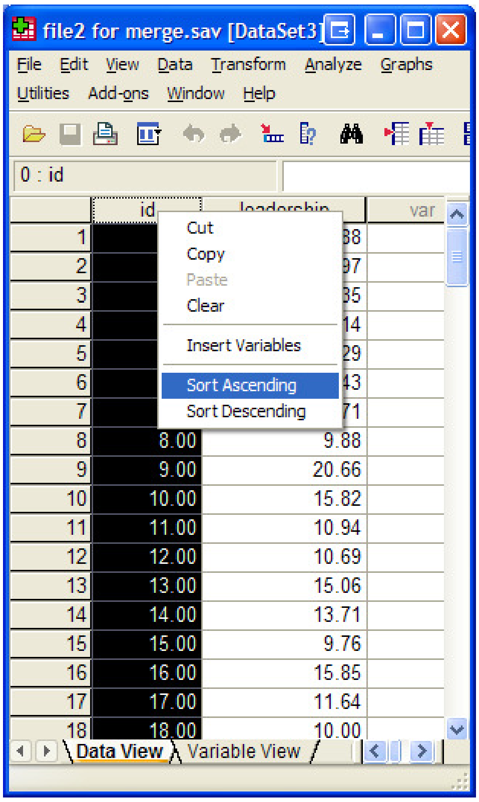
	id	jobsatisfaction	var
1	1.00	5.00	
2	2.00	2.00	
3	3.00	2.00	
4	4.00	3.00	
5	5.00	3.00	
6	6.00	2.00	
7	7.00	4.00	
8	8.00	3.00	
9	9.00	2.00	
10	10.00	4.00	
11	11.00	4.00	
12	12.00	3.00	
13	13.00	4.00	
14	14.00	5.00	
15	15.00	3.00	
16	16.00	1.00	
17	17.00	2.00	
18	18.00	4.00	



file2 for merge.sav [DataSet3]
File Edit View Data Transform Analyze Graphs
Utilities Add-ons Window Help
1 : leadership
4.88364558725071
Data View
Variable View

	id	leadership	var
1	1.00	4.88	
2	2.00	15.97	
3	3.00	12.35	
4	4.00	14.14	
5	5.00	8.29	
6	6.00	12.43	
7	7.00	16.71	
8	8.00	9.88	
9	9.00	20.66	
10	10.00	15.82	
11	11.00	10.94	
12	12.00	10.69	
13	13.00	15.06	
14	14.00	13.71	
15	15.00	9.76	
16	16.00	15.85	
17	17.00	11.64	
18	18.00	10.00	

2. Make sure that each data file is sorted in ascending order by ID. If this is not the case, right click the ID field and select ‘Sort Ascending’. Do this for both data files and save them.

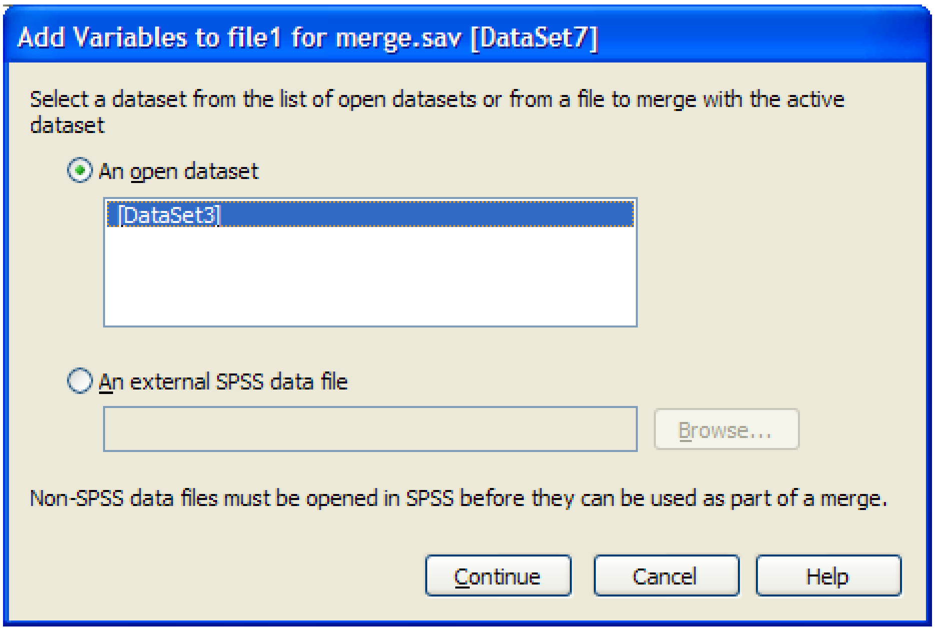


file2 for merge.sav [DataSet3]
File Edit View Data Transform Analyze Graphs
Utilities Add-ons Window Help
0 : id
Cut
Copy
Paste
Clear
Insert Variables
Sort Ascending
Sort Descending
Data View
Variable View

	id	leadership	var
1		88	
2		97	
3		35	
4		14	
5		29	
6		43	
7		71	
8	8.00	9.88	
9	9.00	20.66	
10	10.00	15.82	
11	11.00	10.94	
12	12.00	10.69	
13	13.00	15.06	
14	14.00	13.71	
15	15.00	9.76	
16	16.00	15.85	
17	17.00	11.64	
18	18.00	10.00	

3. Go to the file you want to merge data into. Go to Data >> Merge Files >> Add Variables

4. Select the file that contains the second data file. If the file is already open you can select that.



Add Variables to file1 for merge.sav [DataSet7]

Select a dataset from the list of open datasets or from a file to merge with the active dataset

● An open dataset

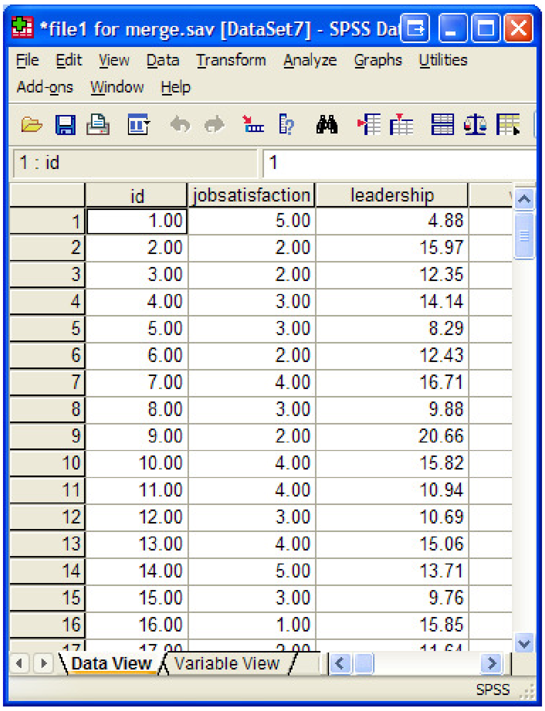
[DataSet3]

○ An external SPSS data file

Non-SPSS data files must be opened in SPSS before they can be used as part of a merge.

5. Click on ‘Match cases on key variables in sorted files’ Place the ID variable into ‘Key Variables’

The result should be a single merged file:



*file1 for merge.sav [DataSet7] - SPSS Da

1 : id 1

	id	jobsatisfaction	leadership
1	1.00	5.00	4.88
2	2.00	2.00	15.97
3	3.00	2.00	12.35
4	4.00	3.00	14.14
5	5.00	3.00	8.29
6	6.00	2.00	12.43
7	7.00	4.00	16.71
8	8.00	3.00	9.88
9	9.00	2.00	20.66
10	10.00	4.00	15.82
11	11.00	4.00	10.94
12	12.00	3.00	10.69
13	13.00	4.00	15.06
14	14.00	5.00	13.71
15	15.00	3.00	9.76
16	16.00	1.00	15.85
17	17.00	2.00	11.64

Recoding variables

Theory

There are many reasons for wanting to recode a variable. These can include:

- Reverse a variable such that high values become low values and low values become high values
- Collapse a set of categories into a smaller set of categories
- Convert a continuous variable into a set of categories (e.g., High and Low)

In SPSS you can either recode into a different variable or change the existing variable. It is generally better to change into a different variable as this is less prone to errors.

SPSS

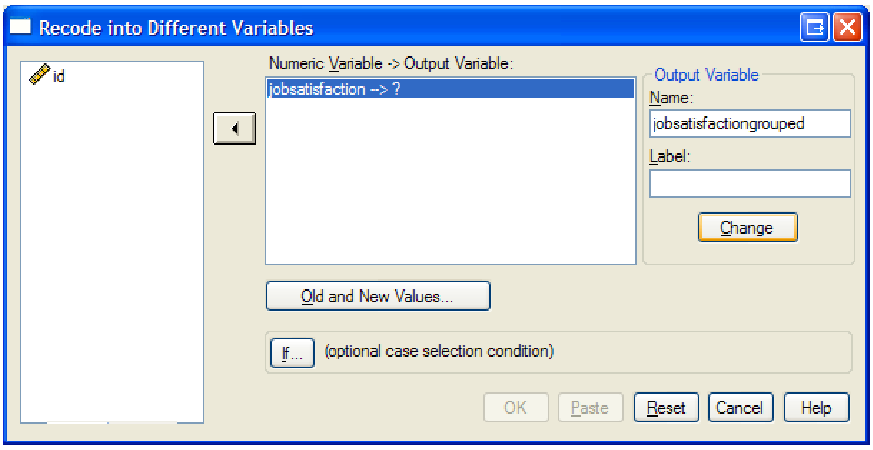
Case Study

Imagine that you had collected job satisfaction information in your organisation. You had collected the data on a 5 points scale. The five points were: Very Dissatisfied (1), Somewhat Dissatisfied (2), Neutral (3), Somewhat Satisfied (4), and Very Satisfied (5). Imagine that you wanted to simplify the results when you reported it to management such that you collapsed very and somewhat dissatisfied into one group and collapsed somewhat and very satisfied into another group.

Running

1. Transform >> Recode >> Recode into Different Variables

2. Enter the original variable into 'Numeric Variable'. Type in the name of the new variable in 'Output Variable - Name'. Click 'Change'.

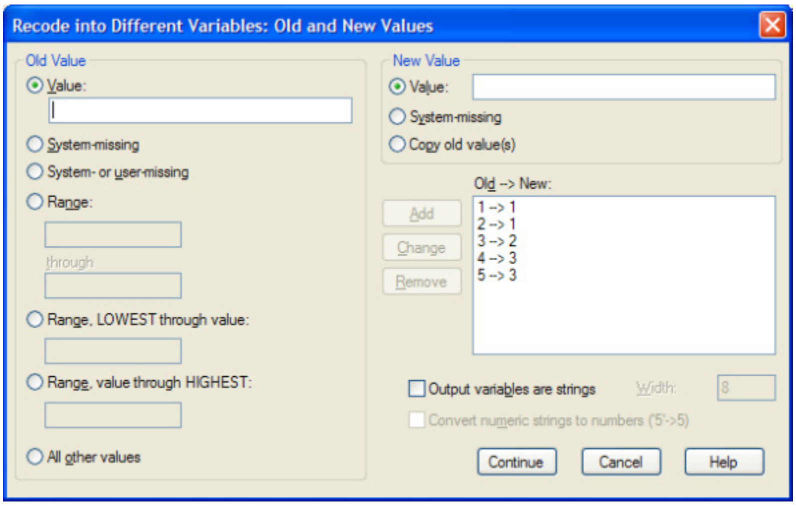


Recode into Different Variables

Numeric Variable -> Output Variable: jobsatisfaction --> ?

Output Variable
Name: jobsatisfactiongrouped
Label:

3. Select Old and New values. In this screen you have to be clear about what the values of the existing variable are and what you want the new values to be. In this example we are converting 5 values into 3. Thus, the first two values 1 and 2 are becoming a single category, the neutral category will stay its own category and values 4 and 5 will become a new value. To specify this you put the old value in 'Old Value' and the New Value in 'New Value' and click 'Add'. You can also specify ranges of old values to be converted into new values.



Recode into Different Variables: Old and New Values

Old Value
● Value:
○ System-missing
○ System- or user-missing
○ Range:
through
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
○ All other values

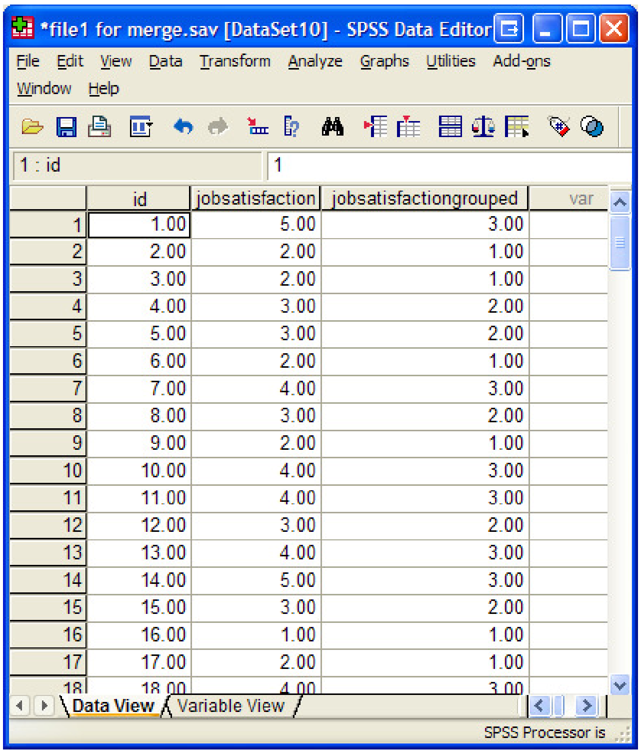
New Value
● Value:
○ System-missing
○ Copy old value(s)

Old --> New:
1 --> 1
2 --> 1
3 --> 2
4 --> 3
5 --> 3

□ Output variables are strings Width: 8
□ Convert numeric strings to numbers ('5'->5)

Output

This will create a new variable in the data file. You will probably then want to assign new value labels to this variable reflecting its new meaning.



*file1 for merge.sav [DataSet10] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons
Window Help
1 : id
1
Data View
Variable View
SPSS Processor is

	id	jobsatisfaction	jobsatisfactiongrouped	var
1	1.00	5.00	3.00	
2	2.00	2.00	1.00	
3	3.00	2.00	1.00	
4	4.00	3.00	2.00	
5	5.00	3.00	2.00	
6	6.00	2.00	1.00	
7	7.00	4.00	3.00	
8	8.00	3.00	2.00	
9	9.00	2.00	1.00	
10	10.00	4.00	3.00	
11	11.00	4.00	3.00	
12	12.00	3.00	2.00	
13	13.00	4.00	3.00	
14	14.00	5.00	3.00	
15	15.00	3.00	2.00	
16	16.00	1.00	1.00	
17	17.00	2.00	1.00	
18	18.00	4.00	3.00	

Converting strings to numbers

Theory

When you have a categorical variable, SPSS prefers to think of the variable as numbers rather than characters and words (strings). The preferred way of entering data when there are a limited set of categories is to assign a number to each category. Many SPSS procedures will not work if the categorical data is represented by a string variable. SPSS provides a command to automatically convert string variables into a series of numbers where each number represents each category. The new variable then has the value labels that match the original string values.

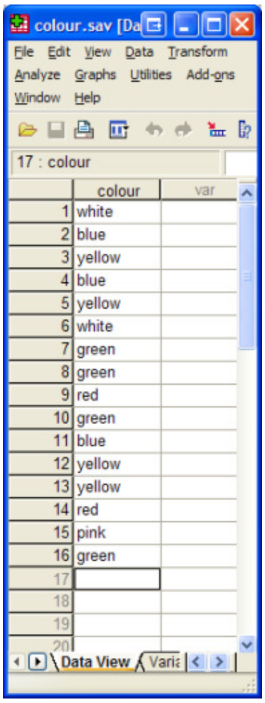
SPSS

Case Study

Imagine you asked a set of participants what was their favourite colour. When it was initially entered, the response of each participant was typed in directly as text. Now you want to convert it to numbers with the labels representing the text.

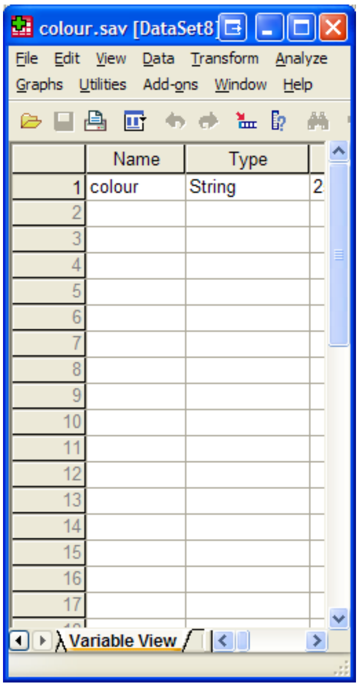
Running

Below we see the data in its original format both in data view and variable view.


colour.sav [Da
File Edit View Data Transform
Analyze Graphs Utilities Add-ons
Window Help
Data View
Varia

17 : colour

	colour	var
1	white	
2	blue	
3	yellow	
4	blue	
5	yellow	
6	white	
7	green	
8	green	
9	red	
10	green	
11	blue	
12	yellow	
13	yellow	
14	red	
15	pink	
16	green	
17		
18		
19		
20		

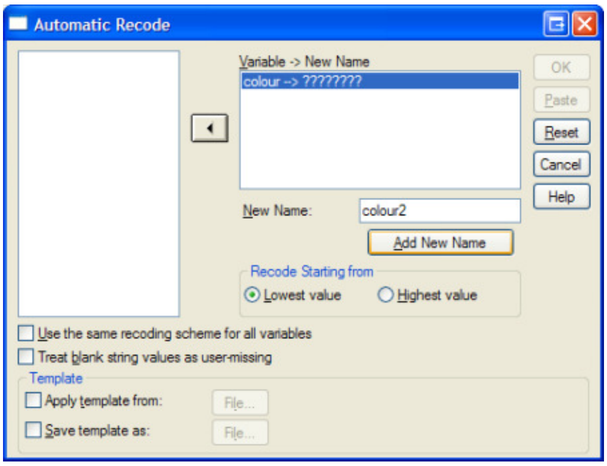

colour.sav [DataSet8]
File Edit View Data Transform Analyze
Graphs Utilities Add-ons Window Help
Variable View

	Name	Type	
1	colour	String	2
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			

Go to Transform >> Automatic Recode

Place the string variable into Variable

Type the new name for the numeric version of the variable in 'New Name' and click 'Add New Name'



Variable -> New Name:
colour --> ????????

New Name: colour2

Recode Starting from
● Lowest value ○ Highest value

□ Use the same recoding scheme for all variables
□ Treat blank string values as user-missing
□ Apply template from:
□ Save template as:

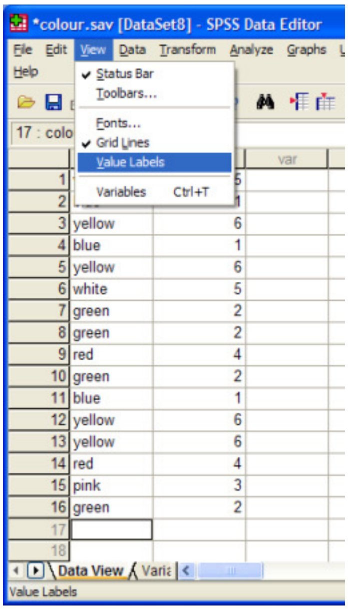
Output

The output displayed by SPSS shows the old and new values for each colour.

```
colour into colour2
Old Value                         New Value  Value Label

blue                                      1  blue
green                                     2  green
pink                                      3  pink
red                                       4  red
white                                     5  white
yellow                                    6  yellow
```

The data file will now have a new variable that called 'colour2', which will be numbers with labels for each colour. When you look at 'Data View' you may see number of words depending on whether you have selected to show value labels (Data >> Value Labels).



			var
1	[illegible]	5	
2	[illegible]	1	
3	yellow	6	
4	blue	1	
5	yellow	6	
6	white	5	
7	green	2	
8	green	2	
9	red	4	
10	green	2	
11	blue	1	
12	yellow	6	
13	yellow	6	
14	red	4	
15	pink	3	
16	green	2	
17			
18			

Filtering

Theory

Sometimes, you will want to run an analysis on just a sub-set of your data, such as just males, or just those over 50 years of age. This comes up often when you have problematic cases, which have strange data and you are concerned that they may be overly influencing your results. You can run the analyses with and without them to see whether you get the same the results. The strategy for dealing with all these cases is to apply a filter on the data.

Instead of manually deleting the cases that do not meet the criteria, SPSS provides a feature to allow you to temporarily filter out cases. The result is that while the filter is applied, any analyses you run will only be applied to the non-filtered cases.

SPSS

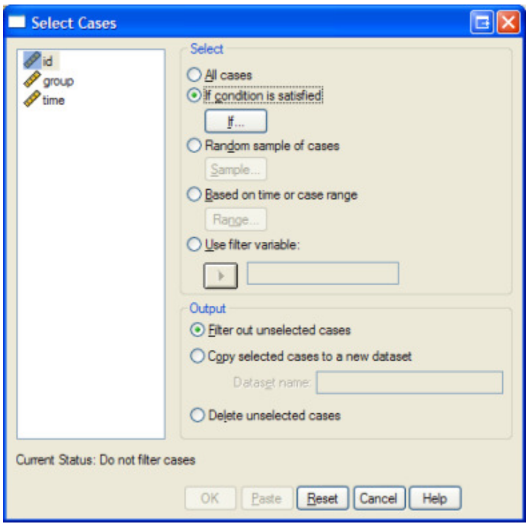
Case Study

In an experiment looking at differences between four groups on a dependent variable, imagine you wanted to run an analysis only on groups 1 and 2.

Running

Data >> Select Cases

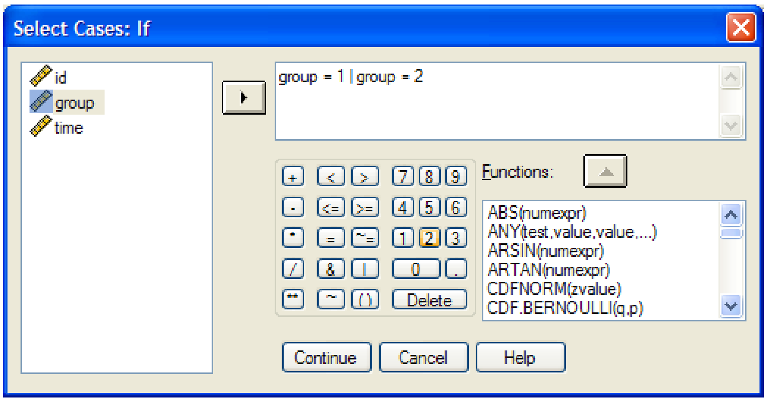
Under this dialog box are a number of options for selecting cases. The one we will use here is “If condition is satisfied”. The default option is to filter out the cases, but it is also possible to delete the unselected cases.



Current Status: Do not filter cases

Set up the logical condition for selecting cases. In the present case we want to select the cases where 'group' is equal to 1 or where 'group is equal to 2. Note that based on your particular example you may use multiple logical operators. These include:

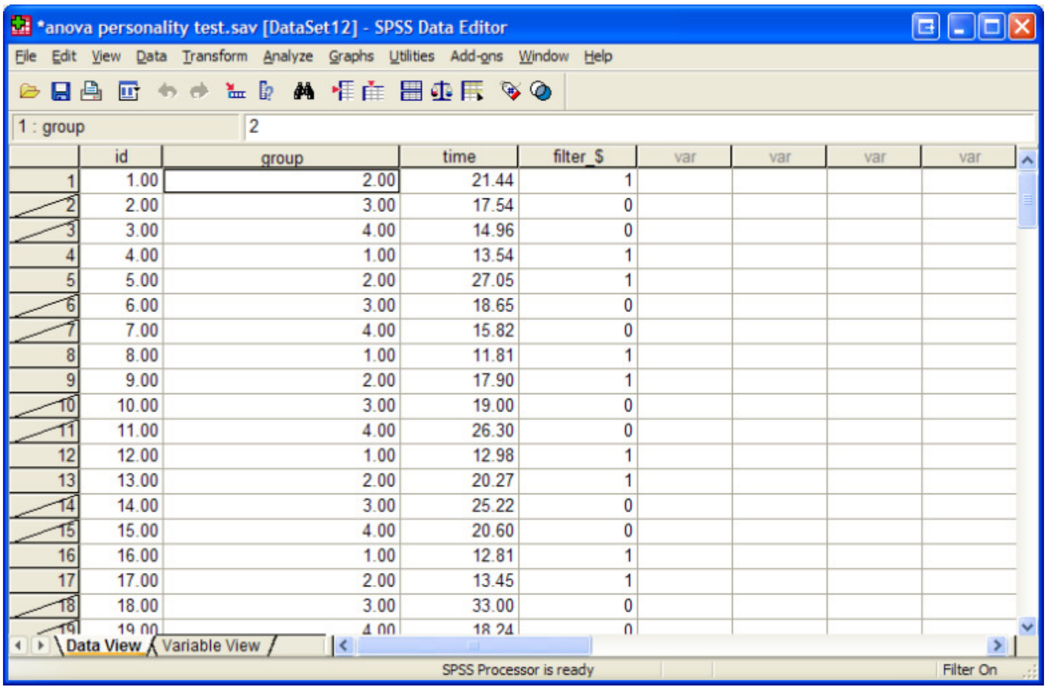
Operator	Meaning
<	less than
<=	less than or equal
>	greater than
>=	greater than or equal
&	And
=	Equal
~=	Not equal



group = 1 | group = 2

Output

Applying the filter has a number of effects. A new variable called 'filter_$' is created, where filtered cases are assigned a value of 0 and non filtered cases are assigned a value of 1. Diagonal lines get drawn across filtered case numbers. A 'Filter On' message appears at the bottom of the screen in the status bar. While the filter is on all, analyses will only be performed on non-filtered cases.



1 : group 2

	id	group	time	filter_$	var	var	var	var
1	1.00	2.00	21.44	1				
2	2.00	3.00	17.54	0				
3	3.00	4.00	14.96	0				
4	4.00	1.00	13.54	1				
5	5.00	2.00	27.05	1				
6	6.00	3.00	18.65	0				
7	7.00	4.00	15.82	0				
8	8.00	1.00	11.81	1				
9	9.00	2.00	17.90	1				
10	10.00	3.00	19.00	0				
11	11.00	4.00	26.30	0				
12	12.00	1.00	12.98	1				
13	13.00	2.00	20.27	1				
14	14.00	3.00	25.22	0				
15	15.00	4.00	20.60	0				
16	16.00	1.00	12.81	1				
17	17.00	2.00	13.45	1				
18	18.00	3.00	33.00	0				
19	19.00	4.00	18.24	0				

Split files

Theory

Related to filtering cases, you may wish to repeat certain analyses in a number of groups. Examples of when you might want to use this technique include the following:

- Getting the correlation between two variables in different groups (e.g., separate correlations for males and females)
- Obtaining graphs for each group separately
- Performing t-tests or ANOVAs comparing groups (e.g., training versus no training) in each of several other groups separately (e.g., younger people and older people)
- Obtaining information on each individual in your data file (in this case splitting by ID)

This technique is very powerful when you work in an organisation and you have to provide overall information but you also have to provide an individual report for each department, team or individual. Split File can make this process very efficient.

SPSS

Case Study

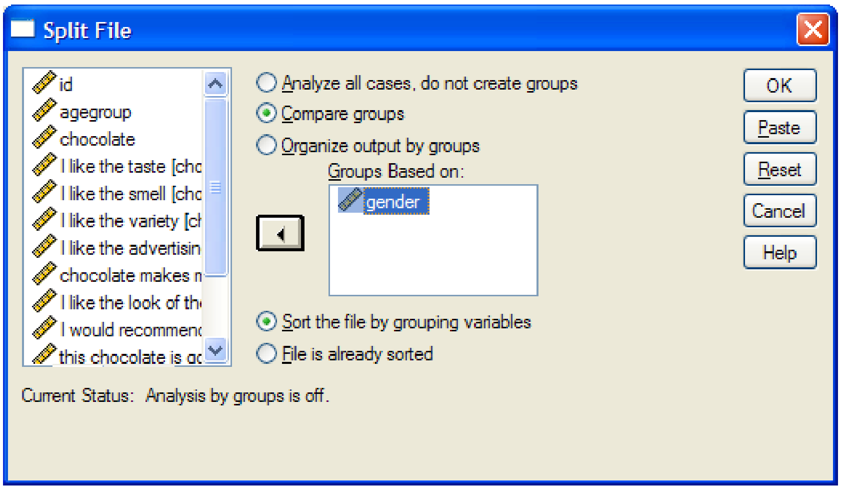
Imagine you wanted to know the correlation between two questions regarding liking chocolate. The first question asked about the taste, and the second question asked about the smell. However, you wanted to see the results separately for male and females

Running

Data >> Split file

Click on 'Compare Groups' or 'Organize output by groups'

Place Gender into "Groups Based on:"



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

Groups Based on: gender

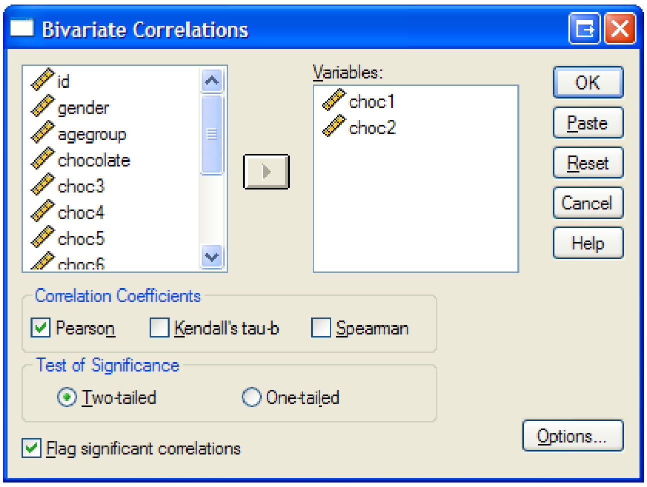
● Sort the file by grouping variables
○ File is already sorted

Current Status: Analysis by groups is off.

This will bring up the text "Split File On" in the status bar.

Run the Correlation as usual using: Analyze >> Correlate >> Bivariate

Place the two variables into the 'variables' box.



Variables: choc1, choc2

Correlation Coefficients
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

Output

The result is the table below which provides information on the correlation between the variables for males and females separately.

Correlations

gender			I like the taste	I like the smell
.00 male	I like the taste	Pearson Correlation	1	.466**
		Sig. (2-tailed)		.000
		N	119	119
	I like the smell	Pearson Correlation	.466**	1
		Sig. (2-tailed)	.000	
		N	119	119
1.00 female	I like the taste	Pearson Correlation	1	.523**
		Sig. (2-tailed)		.000
		N	131	131
	I like the smell	Pearson Correlation	.523**	1
		Sig. (2-tailed)	.000	
		N	131	131

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

Aggregate

Theory

Sometimes we have data on a series a group of individuals in teams and we decide that instead of analysing the individuals we want to analyse the teams. In this situation we want to get the team scores. To do this we could aggregate the individuals to get an average or total team score.

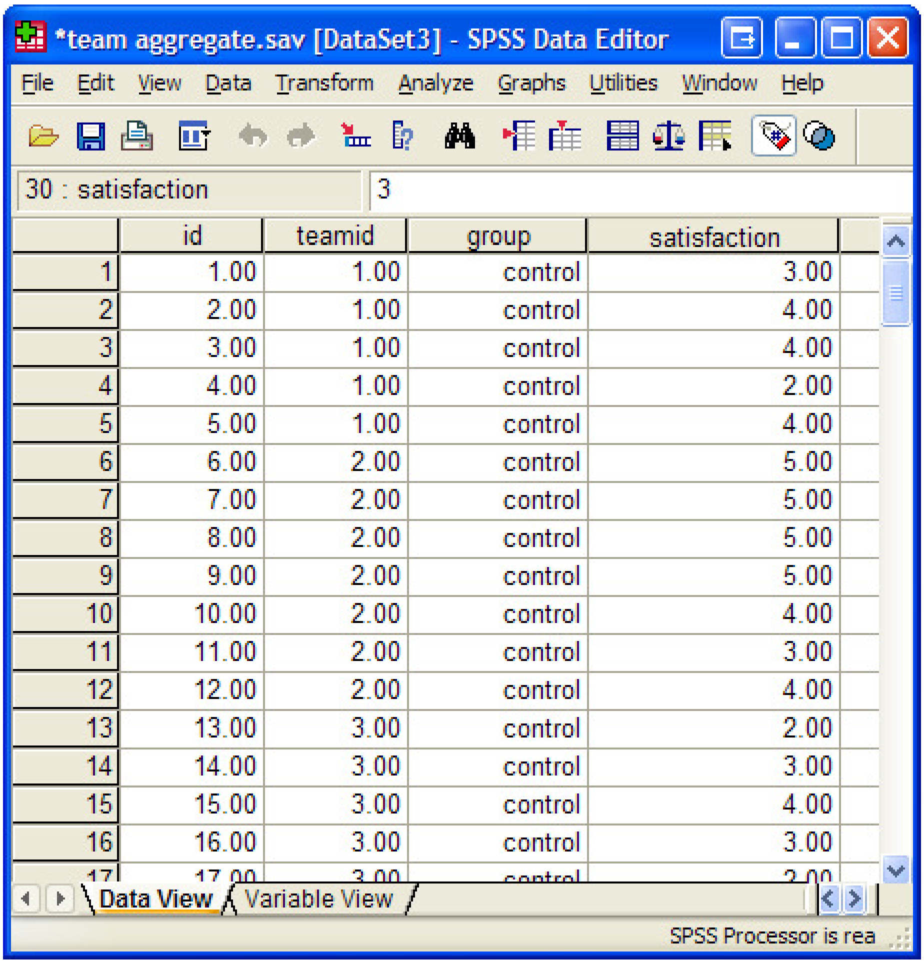
SPSS

Case Study

Imagine that your organisation was exploring the effectiveness of a training program that tried to improve the satisfaction of teams. Imagine also that there were different number of people in each team. The design of the study involved some teams receiving the training program and others not. Satisfaction was measured on every team member. The initial SPSS data file is such that every row of the data file is an individual. You want to make every row a team and obtain the average level of satisfaction for each team.

Running

The initial data file is structured such that the individual is the unit of analysis. Every individual has an ID number, but importantly every individual also belongs to a team. This team membership variable is called 'teamid'.



30 : satisfaction 3

	id	teamid	group	satisfaction
1	1.00	1.00	control	3.00
2	2.00	1.00	control	4.00
3	3.00	1.00	control	4.00
4	4.00	1.00	control	2.00
5	5.00	1.00	control	4.00
6	6.00	2.00	control	5.00
7	7.00	2.00	control	5.00
8	8.00	2.00	control	5.00
9	9.00	2.00	control	5.00
10	10.00	2.00	control	4.00
11	11.00	2.00	control	3.00
12	12.00	2.00	control	4.00
13	13.00	3.00	control	2.00
14	14.00	3.00	control	3.00
15	15.00	3.00	control	4.00
16	16.00	3.00	control	3.00
17	17.00	3.00	control	2.00

To convert this data such that each row is a case, we can use the SPSS aggregate tool.

Data >> Aggregate

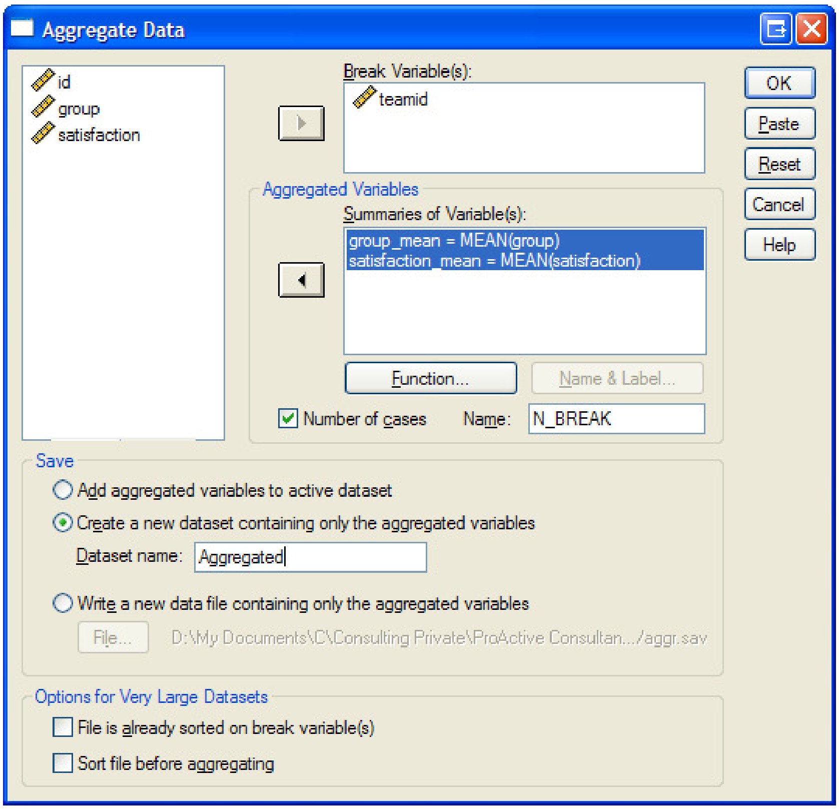
Place 'teamid' in the 'Break variable' box. This tells SPSS that each unique team ID will represent a single case in the aggregated data file.

Place the variables you want to retain in the "Aggregated variables" box.

Specify for each aggregated variable what summary statistic you wish to use. The most common is the mean and this is the default. More options such as sums and medians can be accessed by clicking on the functions button.

Select 'Number of cases'. It can be useful to know how many individuals were in each team as the size of the team may be a variable that influences satisfaction.

Select where the new data should be placed. It is generally best to "Create a new dataset containing only the aggregated variables"



Aggregate Data

id
group
satisfaction

Break Variable(s):
teamid

Aggregated Variables

Summaries of Variable(s):
group_mean = MEAN(group)
satisfaction_mean = MEAN(satisfaction)

☑ Number of cases Name: N_BREAK

Save

○ Add aggregated variables to active dataset

● Create a new dataset containing only the aggregated variables

Dataset name: Aggregated

○ Write a new data file containing only the aggregated variables

D:\My Documents\C\Consulting Private\ProActive Consultan.../aggr.sav

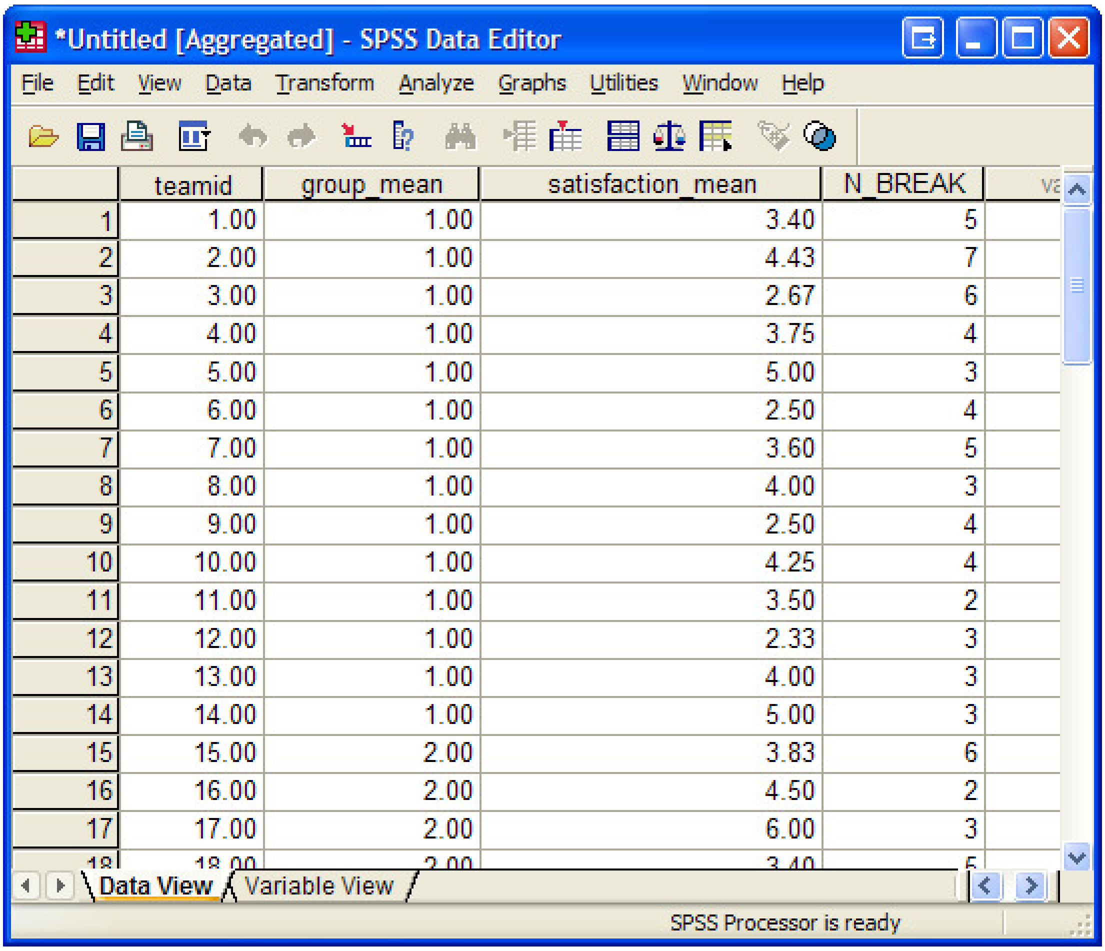
Options for Very Large Datasets

□ File is already sorted on break variable(s)

□ Sort file before aggregating

Output

The new data file is shown below. Now, each case is a single team. Group_mean is the same as group for the individuals because every individual in a team are in the same group anyway. 'Satisfaction_mean' is now the mean satisfaction levels for that particular team. This data could then be used to examine whether teams in the control and the training condition differ in terms of team satisfaction.



*Untitled [Aggregated] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

	teamid	group_mean	satisfaction_mean	N_BREAK
1	1.00	1.00	3.40	5
2	2.00	1.00	4.43	7
3	3.00	1.00	2.67	6
4	4.00	1.00	3.75	4
5	5.00	1.00	5.00	3
6	6.00	1.00	2.50	4
7	7.00	1.00	3.60	5
8	8.00	1.00	4.00	3
9	9.00	1.00	2.50	4
10	10.00	1.00	4.25	4
11	11.00	1.00	3.50	2
12	12.00	1.00	2.33	3
13	13.00	1.00	4.00	3
14	14.00	1.00	5.00	3
15	15.00	2.00	3.83	6
16	16.00	2.00	4.50	2
17	17.00	2.00	6.00	3
18	18.00	2.00	3.40	5

10. ONE SAMPLE T-TEST

Theory

When we know the population mean, but not the population standard deviation, we may want to know whether the mean obtained in our sample is significantly different from the population. In this situation we use the one sample t-test.

Assumptions

Normality

It is assumed that the dependent variable is normally distributed in the population. The procedure is relatively robust to modest violations of this assumption

Homogeneity of variance

The variance in the comparison group and in your sample is assumed to be the same. This can be tested statistically with Levene's test of Homogeneity of variance.

SPSS

Overview of Study

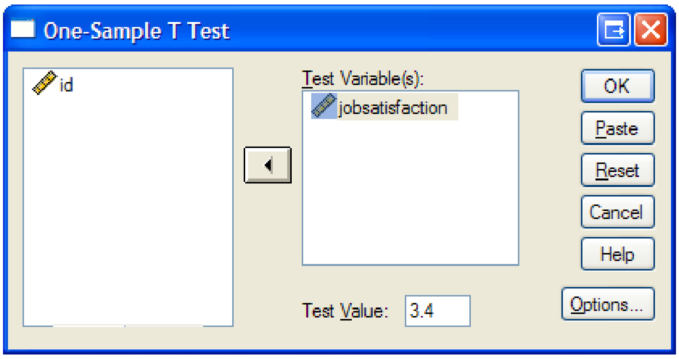
Imagine you had collected job satisfaction data from the employees in your organisation on a five-point scale. You have data from a benchmarking agency to show that the mean job satisfaction level for your industry is 3.4. You want to know whether your organisation's level of job satisfaction is significantly different from the benchmark.

Running

Analyze >> One Sample T Test

Place the dependent variable into "Test Variables"

Enter the value of the population mean into "Test Value"



Test Variable(s): jobsatisfaction

Test Value: 3.4

Output

The first table provides descriptive statistics about the mean of your sample. We can see that in the organisation job satisfaction is lower (2.99) in comparison to the population benchmark (3.4). The next table tells us whether this is statistically significant. The sig (2-tailed) section provides a p value less than .05. Thus, we can conclude that the difference between our sample and the population mean is statistically significant.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
jobsatisfaction	100	2.9900	1.02981	.10298

One-Sample Test

	Test Value = 3.4					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
jobsatisfaction	-3.981	99	.000	-.41000	-.6143	-.2057

Write-up

A comparison was made between the job satisfaction results for the organisation (mean=2.99, sd=1) and the benchmark for the industry (mean =3.4). A one sample t-test showed that this was a statistically significant difference, t (99) = -3.98, p<.001.

11. INDEPENDENT GROUPS T-TEST

Theory

Overview

The independent-groups t-test is used to test whether two groups differ in terms of their means on a continuous dependent variable. Examples of the kinds of questions that could be answered include: Are there gender differences in intelligence? Does a group that receives training perform better than a group that does not? Is there a difference in job satisfaction between older and younger workers?

Assumptions

Independence of Observations

Independence of observations is the assumption that there is no relationship between one observation and the next. This assumption is usually satisfied.

Homogeneity of Variance

Homogeneity of variance is the assumption that the within group variance is equal across groups in the population.

SPSS provides Levene's test, which test the assumption. If Levene's test has a significance level less than .05, then the assumption is typically held to be violated. T-tests are relatively robust to violations of the assumption especially if group sizes are relatively equal.

Normality of the Dependent Variable

Normality is the assumption that the scores on the dependent variable are normally distributed in each group.

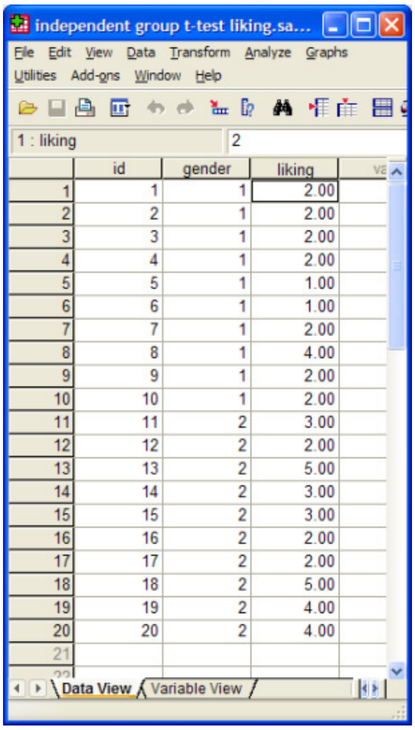
SPSS

Overview of Study

Imagine you were in Market research and you wanted to know whether there was any difference between males and females in liking for you new advertising campaign. You showed the advertisement to 10 females and 10 males. Each participant rated there liking for the advertising on a 5-point scale where higher scores indicated greater liking of the advertising.

Running

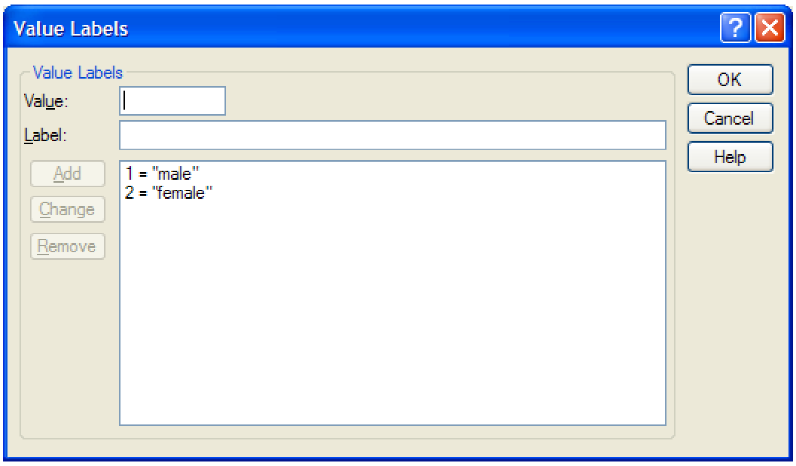
The raw data file would appear as below. One column represents the independent variable (gender), and one variable represents the dependent variable (liking).


independent group t-test liking.sa...
File Edit View Data Transform Analyze Graphs
Utilities Add-ons Window Help
Data View
Variable View

1 : liking 2

	id	gender	liking
1	1	1	2.00
2	2	1	2.00
3	3	1	2.00
4	4	1	2.00
5	5	1	1.00
6	6	1	1.00
7	7	1	2.00
8	8	1	4.00
9	9	1	2.00
10	10	1	2.00
11	11	2	3.00
12	12	2	2.00
13	13	2	5.00
14	14	2	3.00
15	15	2	3.00
16	16	2	2.00
17	17	2	2.00
18	18	2	5.00
19	19	2	4.00
20	20	2	4.00
21			

It is useful to go into Variable View and specify under Values for gender that a value of 1 should have the label "Male" and 2 should have "Female".

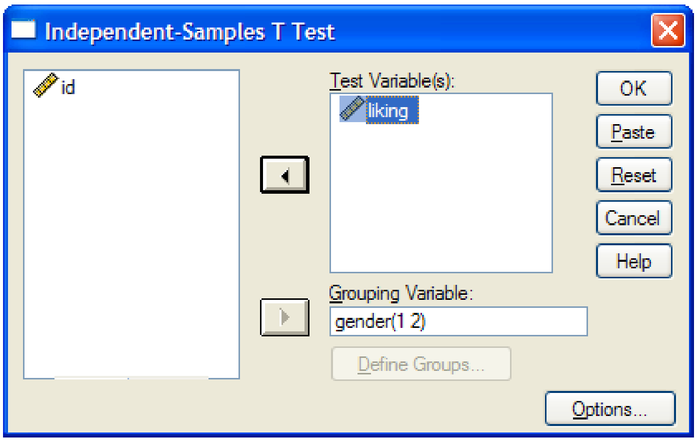


1 = "male"
2 = "female"

You then go to: Analyze >> Compare Means >> Independent-Samples T Test

Place gender into the Grouping Variable and press Define Groups to specify what numbers represent the groups in the data file.

Place 'liking' in the Test Variable (Dependent Variable)



Test Variable(s): liking

Grouping Variable: gender(1 2)

Output

This first table shows the descriptive statistics (mean, standard deviation and sample size) associated with the two groups. From an initial look at the means it would appear that females liked the advertisement more than males. But is this difference statistically significant?

Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
liking	male	10	2.0000	.81650	.25820
	female	10	3.3000	1.15950	.36667

The table below shows two rows of data. One assumes homogeneity of variance and one does not. The process is to first look at Levene's test to see whether homogeneity of variance is a reasonable assumption. As the p-value is not less than .05, we can assume homogeneity of variance. We then proceed to analyse "Equal variances assumed" row. The p-value associated with the test is statistically significant at the .05 level.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Liking	Equal variances assumed	3.875	.065	-2.899	18	.010	-1.30000	.44845	-2.24217	-.35783
	Equal variances not assumed			-2.899	16.164	.010	-1.30000	.44845	-2.24990	-.35010

Write-up

An independent-groups t-test was performed examining difference between males and females on liking for a new advertising campaign. A significant difference was found between males and females for liking of the program, t (18) = 2.90, p = 0.1. Examination of the means showed that females (mean = 3.3, sd = 1.2) preferred the advertising to males (mean = 2.0, sd = 0.8).

12. REPEATED MEASURES T-TEST

Theory

Overview

The repeated measures t-test is used to test whether means differ on two levels of a variable obtained from the same set of individuals. Examples could include looking at knowledge before and after training, health levels before and after receiving a drug, or satisfaction with two different computer interfaces. Repeated measures designs are generally more powerful than between-subjects designs, because we are able to remove the effect due to individual differences.

Assumptions

Independence of observations

Assumption is that observations within a treatment are independent. It is not assumed that observations on the same person are independent. By definition people will be related to themselves from time 1 to time 2.

Normality of difference scores

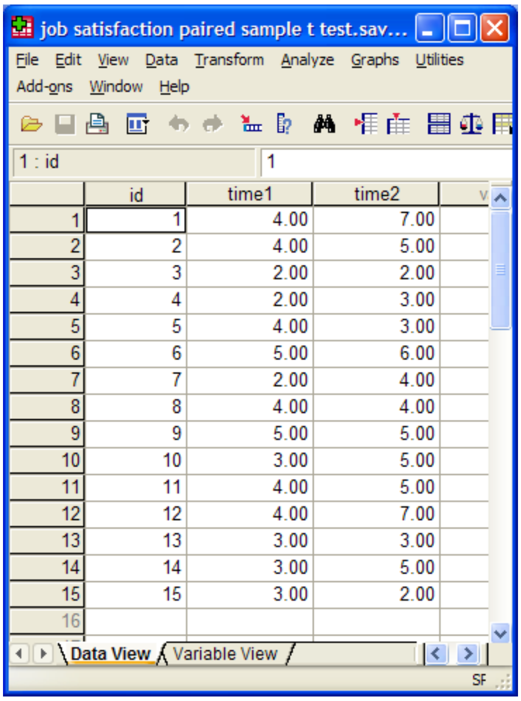
The assumption is that in the population the difference scores are normally distributed. A difference score is the difference between a participants score in the first condition minus the second condition. This can be computed using SPSS: Transform >> Compute; then write an expression that reflects one variable minus the other. This can then be plotted as a histogram.

SPSS

Overview of Study

Imagine a scenario where you are a Human Resource manager and have just implemented a program to attempt to increase job satisfaction. You have measured job satisfaction at time 1 prior to your program. Six months later (time2) after implementing your program, you have then measured job satisfaction on the same employees. Job satisfaction was measured on a 7 point scale where higher scores indicate greater job satisfaction.

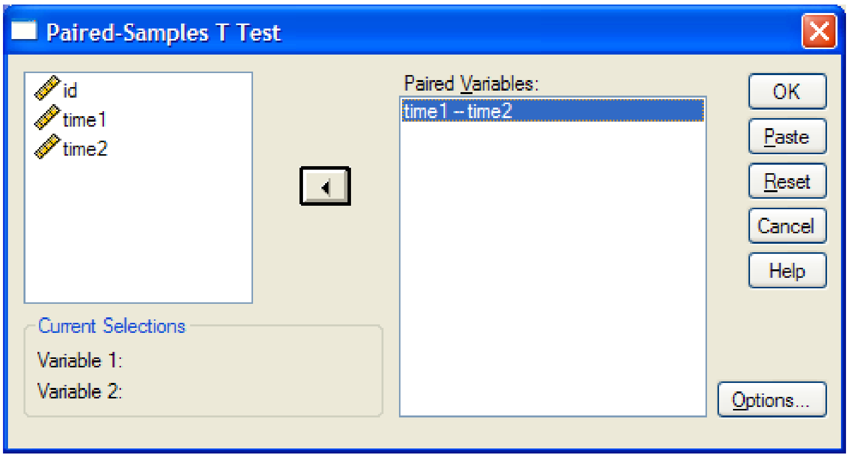
Running


job satisfaction paired sample t test.sav...
File Edit View Data Transform Analyze Graphs Utilities
Add-ons Window Help
1 : id
1
Data View
Variable View
SF

	id	time1	time2
1	1	4.00	7.00
2	2	4.00	5.00
3	3	2.00	2.00
4	4	2.00	3.00
5	5	4.00	3.00
6	6	5.00	6.00
7	7	2.00	4.00
8	8	4.00	4.00
9	9	5.00	5.00
10	10	3.00	5.00
11	11	4.00	5.00
12	12	4.00	7.00
13	13	3.00	3.00
14	14	3.00	5.00
15	15	3.00	2.00
16			

Analyze >> Compare Means >> Paired Samples T Test

Click on time1 then on time2 and copy across to paired variables.



Paired Variables:
time1 -- time2

Current Selections
Variable 1:
Variable 2:

Output

The following table shows the means, sample size and standard deviation of job satisfaction at the two time points. Which time point looks like it had higher job satisfaction? It looks like job satisfaction went up between time1 and time2.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	time1	3.4667	15	.99043	.25573
	time2	4.4000	15	1.59463	.41173

This table shows the correlation between scores at time1 and time2. This shows whether the people who were more satisfied at time1 were also more satisfied at time2. This is the stable individual difference factor that is removed in repeated measures design. Thus, the larger the correlation, the more stable individual differences are, and the more powerful the repeated measures design is.

Paired Samples Correlations

	N	Correlation	Sig.
Pair 1 time1 & time2	15	.597	.019

This table shows whether the difference between the two time points is statistically significant. Based on an alpha of .05, if the significance column is less than .05, we would conclude that there was a significant difference

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Dev iation	Std. Error Mean	95% Conf idence Interv al of the Difference				
				Lower	Upper			
Pair 1 time1 - time2	-.933	1.279	.33046	-1.64211	-.22456	-2.824	14	.014

Write-up

A repeated measures t-test was performed to assess whether job satisfaction changed over time following an intervention that aimed to increase job satisfaction. Prior to the intervention, mean job satisfaction was 3.4 (sd = .99) and after intervention mean job satisfaction was 4.4 (sd = 1.59). Based on a paired samples t-test this represented a significant increase in job satisfaction over time, t(14) = -2.82, p = .014.

13. ONE-WAY ANOVA

Theory

Overview

Whereas t-tests are limited to comparing means across 2 groups, ANOVA can be used when there are 2 or more groups. For example, if we were interested in the effect of training on performance and we had three groups (practical training, theoretical training, and no training) we could use ANOVA to compare the groups.

The independent variable can be called a factor. Each factor has a number of levels. Levels are sometimes also called treatments or groups. In the previous example where there are three groups, we could describe this study as having three levels of training.

ANOVA is a very powerful procedure that can include one or more factors. When there is one factor, the technique is called one-way ANOVA, and when there are two factors, it is called two-way ANOVA, etc.

ANOVA can also involve between subjects and repeated measures factors. Between subjects factors occur when different individuals are in each treatment. Repeated measures factors occur when the same subjects appear in the different treatment conditions, such as when you measure people at multiple time points.

ANOVA can even incorporate continuous predictor variables that are called covariates. These can be used to control for other effects in our experiments. In this case it is called ANCOVA.

As the current course is introductory, we will be only discussing the simplest example of One-way ANOVA, where we have one between subjects variable.

ANOVA tests the null hypothesis that all the group means are the same.

Post-Hoc Tests

If you obtain a significant result for your overall ANOVA, you can conclude that not all the group means are the same. However, if you have 3 or more groups, you will not know which groups are different from which other groups. Is it that groups 1 and 2 are the same and group 3 is different or is it that all three groups are significantly different from each other?

Post-hoc tests perform each pairwise comparison between groups. This allows you to determine which groups are significantly different from which other groups. One of the most common post-hoc tests is Tukey's Honestly Significant Difference.

Assumptions

Independence of Observations

Each case is assumed to be independent of every other case.

Homogeneity of Variance

The within group variance for each group is assumed to be equal in the population. We typically use Levene's test of Homogeneity of Variance to test this assumption. If the significance level is less than .05, then the assumption is said to be violated. In general ANOVA is relatively robust to modest violations of this assumption, especially if group sizes are equal.

Normality of the dependent variable

The dependent variable is assumed to be normally distributed within each group.

Overview of Study

A study was conducted in social psychology to see the effect of self-perceptions on persistence. Four groups completed a personality test and were given feedback about what the test indicated about their future. The feedback given had nothing to do with how the person actually responded on the personality test. The feedback condition was the independent variable:

Future alone (told they would be alone later in life)

Future belonging (told they would have good relationships)

Misfortune control (told they were prone to accidents)

No-feedback control (not told anything about their future).

Participants then performed an unsolvable experimental puzzle. The dependent variable was the amount of time that the participant tried to solve the unsolvable puzzle.

The data here are not exactly the same as those published, but the results are consistent with their findings.

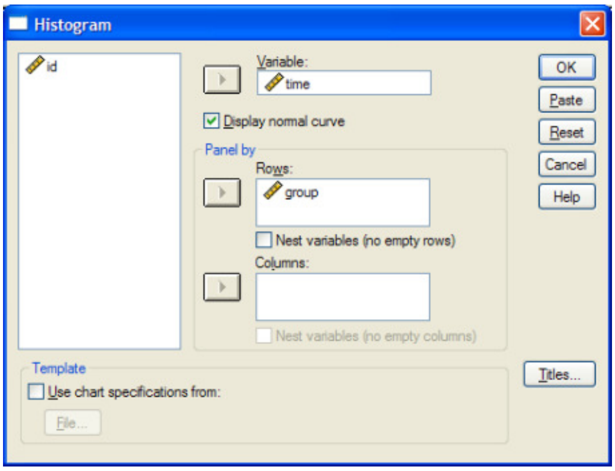
Baumeister, R. F., DeWall, C N., Ciarocco, N. J., & Twenge, J. M. (2005). Social Exclusion Impairs Self-Regulation. *Journal of Personality and Social Psychology, 88*(4), 589-604.

Assessing Normality

Graph >> Histogram

Variable = Time

Rows: Group



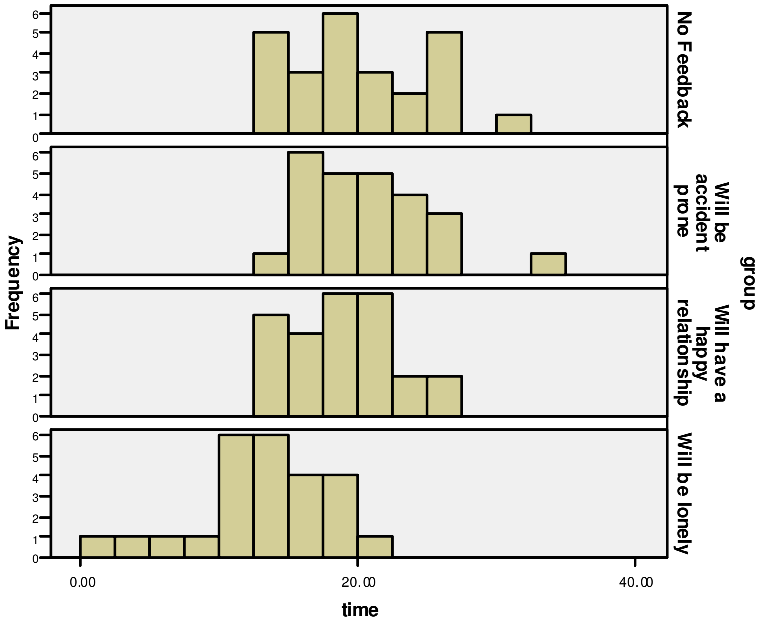
Variable: time

☑ Display normal curve

Rows: group

□ Nest variables (no empty rows)

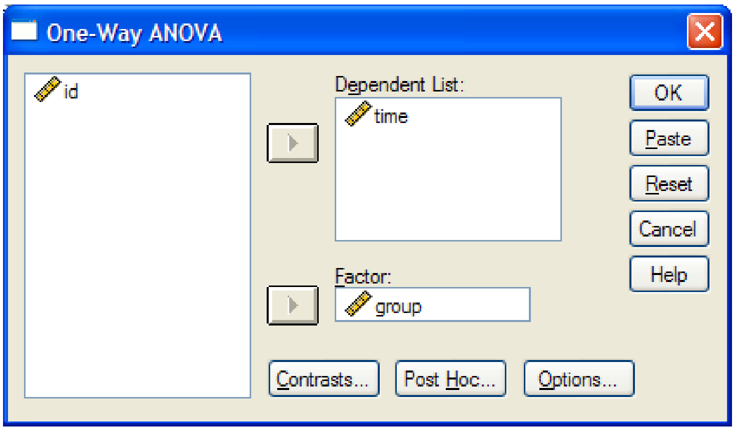
This creates a nested histogram. If we look at the distribution of the dependent variable in each of the four groups we can see that each looks roughly symmetrical.



Running

Analyze >> Compare Means >> One way ANOVA

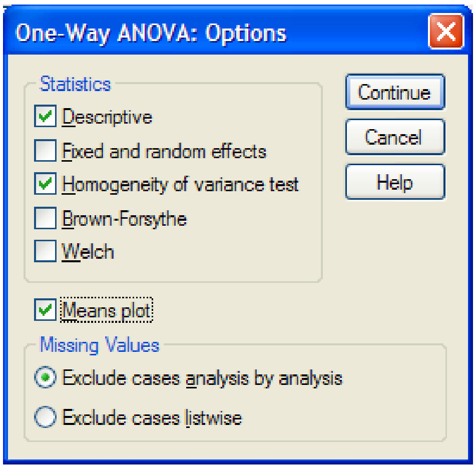
Place time in the dependent and group in the factor



Dependent List: time

Factor: group

Under ‘Options…’ select Homogeneity of Variance test and Descriptives



Statistics

☑ Descriptive

□ Fixed and random effects

☑ Homogeneity of variance test

□ Brown-Forsythe

□ Welch

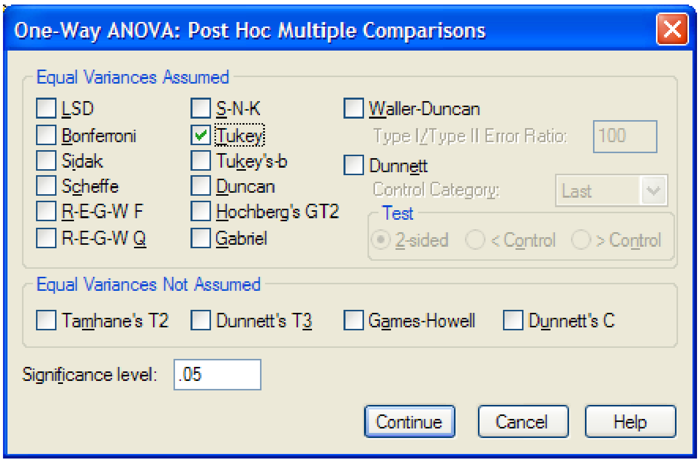
☑ Means plot

Missing Values

● Exclude cases analysis by analysis

○ Exclude cases listwise

Under ‘Post Hoc’ select the post hoc test you want. A commonly used post hoc test is Tukey’s HSD. For this example, select “Tukey”.



One-Way ANOVA: Post Hoc Multiple Comparisons

Equal Variances Assumed

□ LSD □ S-N-K □ Waller-Duncan
□ Bonferroni ☑ Tukey Type I/Type II Error Ratio: 100
□ Sidak □ Tukey's-b □ Dunnett
□ Scheffe □ Duncan Control Category: Last
□ R-E-G-W F □ Hochberg's GT2 Test
□ R-E-G-W Q □ Gabriel ● 2-sided ○ < Control ○ > Control

Equal Variances Not Assumed

□ Tamhane's T2 □ Dunnett's T3 □ Games-Howell □ Dunnett's C

Significance level: .05

Output

This first table reports the descriptive statistics for each of the groups. Initial observation of the means and standard deviations is useful to get a sense of what occurred in the study. In this example, it appears that the group that was told they will be lonely persisted less on the impossible study than the other three groups. The next step is to determine whether the differences are statistically significant.

Descriptives

time

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Will be lonely	25	13.3132	4.82455	.96491	11.3217	15.3046	1.83	22.07
Will have a happy relationship	25	18.9889	3.98093	.79619	17.3456	20.6321	12.93	27.05
Will be accident prone	25	20.6267	4.32027	.86405	18.8434	22.4101	12.69	33.00
No Feedback	25	20.0478	5.01861	1.00372	17.9762	22.1194	12.86	30.42
Total	100	18.2441	5.35260	.53526	17.1821	19.3062	1.83	33.00

This table assess the assumption of homogeneity of variance. Here, the p value (Sig.) is greater than .05. Thus, we would feel comfortable in assuming homogeneity of variance.

Test of Homogeneity of Variances

time

Levene Statistic	df1	df2	Sig.
.682	3	96	.565

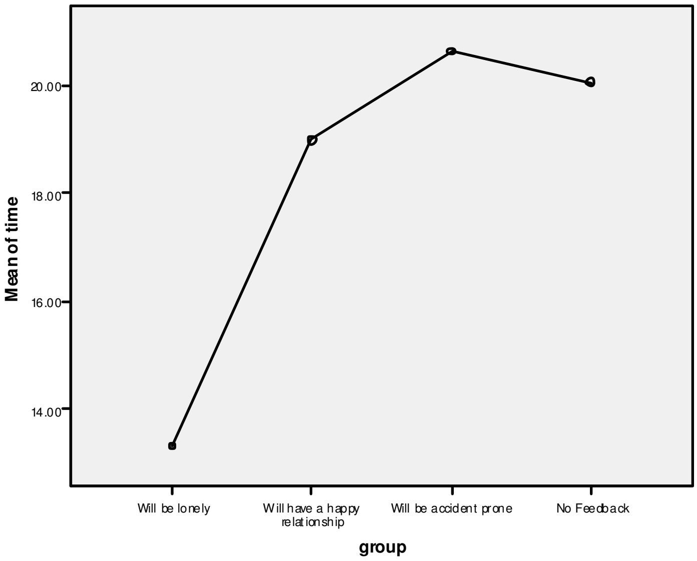
This table reports our significance test for whether the group means are all the same. The p value (Sig.) is less than .05. Therefore, it would be concluded that there is a significant difference between the means in the four groups. At this point we do not know which groups are different from which other groups, but we do know that at least one of the

ANOVA

time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	844.982	3	281.661	13.578	.000
Within Groups	1991.405	96	20.744		
Total	2836.387	99			

The graph shows the means for the four groups. It highlights that the group that were told that they would be alone appear to have persisted substantially less than the other three groups.



The post-hoc tests report the size and statistical significance of any difference between pairs of group means. In the current example, the group told that they would be lonely in the future persisted significantly less than the other three groups, but there were no significant differences between the other three groups.

Multiple Comparisons

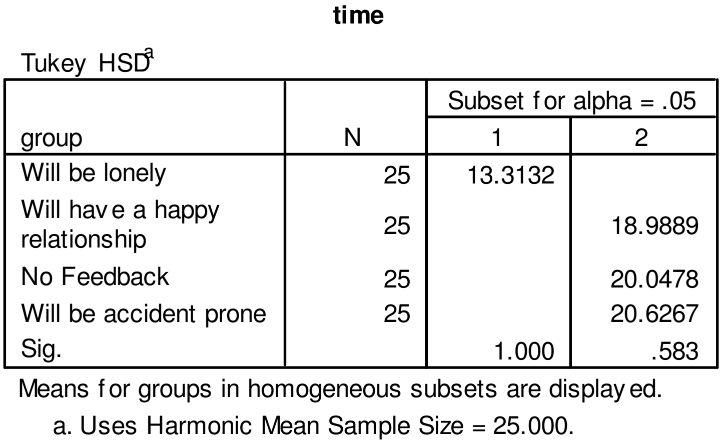
Dependent Variable: time

Tukey HSD

(I) group	(J) group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Will be lonely	Will have a happy relationship	-5.67573*	1.28822	.000	-9.0439	-2.3075
	Will be accident prone	-7.31358*	1.28822	.000	-10.6818	-3.9454
	No Feedback	-6.73466*	1.28822	.000	-10.1028	-3.3665
Will have a happy relationship	Will be lonely	5.67573*	1.28822	.000	2.3075	9.0439
	Will be accident prone	-1.63785	1.28822	.583	-5.0060	1.7303
	No Feedback	-1.05893	1.28822	.844	-4.4271	2.3093
Will be accident prone	Will be lonely	7.31358*	1.28822	.000	3.9454	10.6818
	Will have a happy relationship	1.63785	1.28822	.583	-1.7303	5.0060
	No Feedback	.57892	1.28822	.970	-2.7893	3.9471
No Feedback	Will be lonely	6.73466*	1.28822	.000	3.3665	10.1028
	Will have a happy relationship	1.05893	1.28822	.844	-2.3093	4.4271
	Will be accident prone	-.57892	1.28822	.970	-3.9471	2.7893

*. The mean difference is significant at the .05 level.

This second Post-hoc test table is another way of summarising the results from the Post-hoc test. When a group's mean appears only in one subset it is significantly different from other groups that do not appear in that subset. In the present example the "will be lonely" condition appears in subset 1 and the other three groups appear in subset 2. This shows that the "will be lonely" condition persisted significantly less than the other three groups and that there were no significant differences between the other three groups.



time

Tukey HSD[a]

group	N	Subset for alpha = .05	
		1	2
Will be lonely	25	13.3132	
Will have a happy relationship	25		18.9889
No Feedback	25		20.0478
Will be accident prone	25		20.6267
Sig.		1.000	.583

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 25.000.

Write-up

A study was performed to see whether the feedback given to participants influenced the amount of time they persisted on an impossible puzzle. An assessment of assumptions of normality and homogeneity of variance showed no major violations of assumptions. A one-way between subjects ANOVA showed that there was a statistically significant difference in mean persistence time on the puzzle across the four groups, F (3, 96) = 13.6, p <.001. Tukey's HSD post-hoc test revealed that the group that were told they would be lonely (mean =13.3, sd = 4.8) persisted significantly less (p<.05) than the three other groups (happy relationship, mean = 19.0, sd = 4.0; accident prone, mean = 20.6, sd = 4.3; no feedback, mean = 20.0, sd = 5.0).

14. CORRELATION

Theory

Introduction

Correlation coefficient summarises the linear relationship between two variables. The correlation coefficient is represented in text by the letter "r". Correlation coefficients range from -1 to +1.

First, inspect the direction of the relationship: a positive relationship suggests that an INCREASE in one variable is associated with an INCREASE in another variable; a negative relationship suggests that an INCREASE in one variable is associated with a DECREASE in another variable.

Second, examine the size of the correlation. Some rules of thumb are presented below, but it is also important to see the correlation in a context.

"r" = zero to about .20 - no or negligible correlation.

"r" = .20 to .40 - low degree of correlation.

"r" = .40 to .60 - moderate degree of correlation.

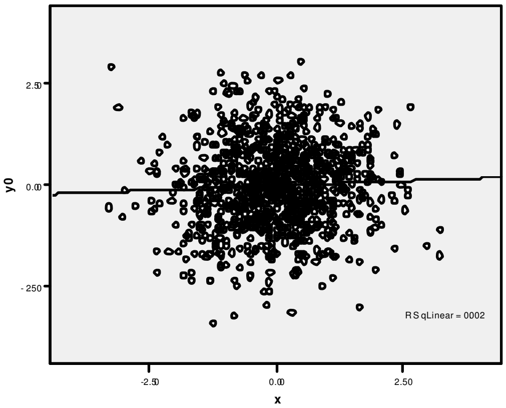
"r" = .60 to .80 - marked degree of correlation.

"r" = .80 to 1.00 - high correlation.

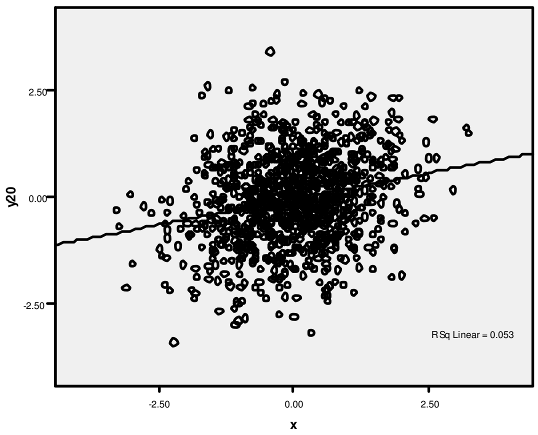
[A. Franzblau (1958), A Primer of Statistics for Non-Statisticians, Harcourt, Brace & World. (Chap. 7)

Examples

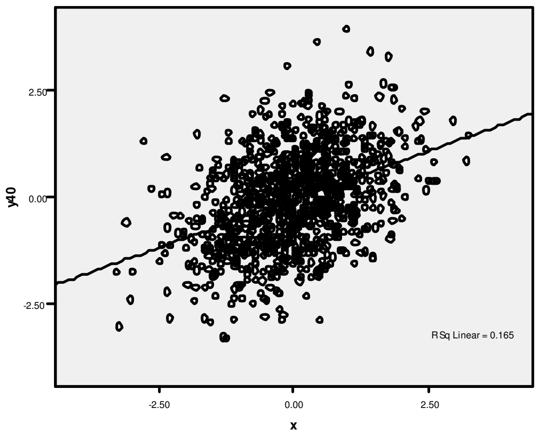
r ≈ 0



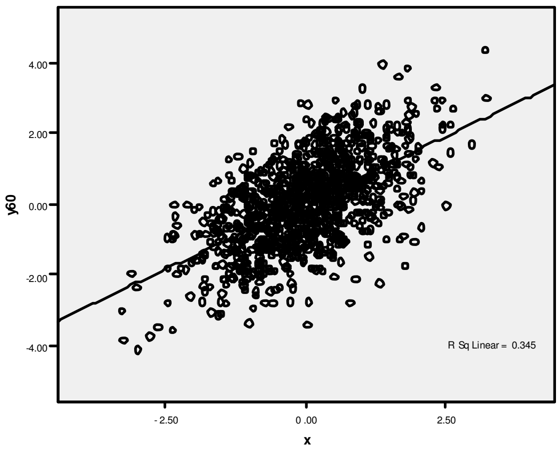
r ≈ .20


2.50
0.00
-2.50
y20
RSq Linear = 0.053
-2.50
0.00
2.50
x

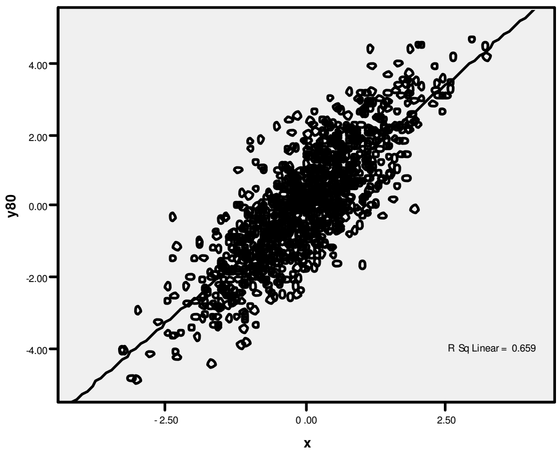
r ≈ .40


2.50
0.00
-2.50
y40
RSq Linear = 0.165
-2.50
0.00
2.50
x

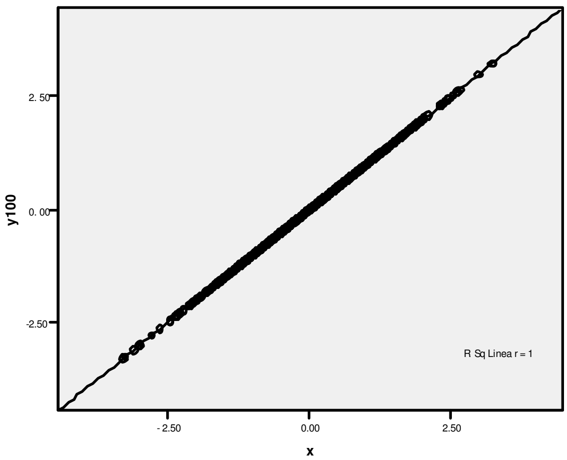
r ≈ .60


4.00
2.00
0.00
-2.00
-4.00
y60
R Sq Linear = 0.345
- 2.50
0 .00
2.50
x

r ≈ .80



r = 1.0



Further Discussion

Squaring the correlation coefficient indicates the percentage of variance explained in one variable by the other (e.g., $r = .5; r^2 = .5 * .5 = .25$; 25% of variance is variable y is explained by variable x)

A t-test can be performed to test whether a correlation coefficient is significantly different from zero in the population.

The standard form of correlation is Pearson's Product Moment Correlation. Another form of correlation coefficient is Spearman's rho, which a correlation performed on the ranks of the data. Spearman's rho only assumes that the data is ordinal. There are a range of other measures of association, which can also be used to describe the relationship between two variables.

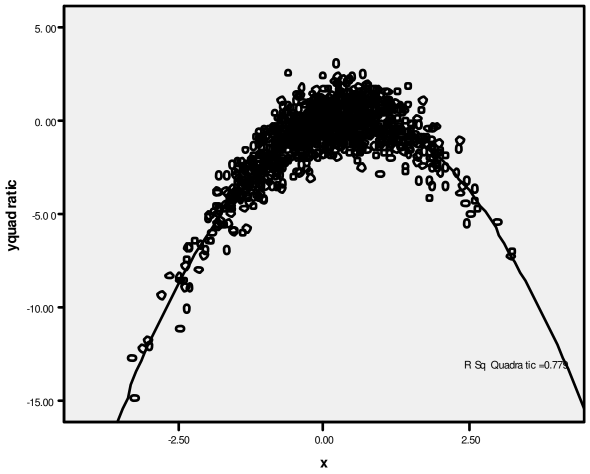
Assumptions

Linearity

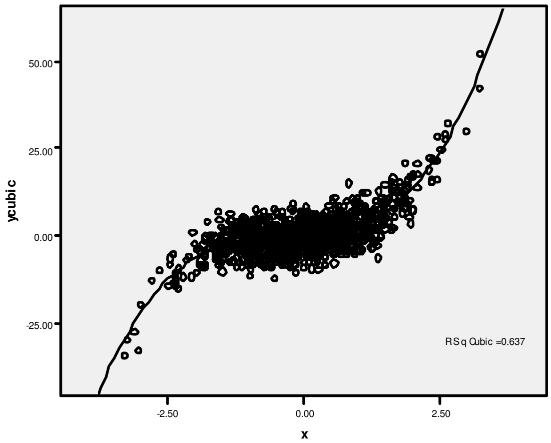
The relationship between two variables is assumed to be linear. A linear relationship is one where an increase in one variable is associated with an increase or decrease in another variable. All the examples previously displayed demonstrated this kind of relationship. You can assess this assumption by looking at a scatter plot (Graphs >> Scatter/Dot). The assumption is violated when there is a substantially non-linear relationship present.

There are many different types of non-linear relationships. These include the higher order polynomial trends such as quadratic, cubic and quartic. Other non-linear relationships include power-relationship and s-shaped relationships.

The relationship below is called quadratic or U-Shaped:



The relationship below is cubic: it has two turning points



Outliers

Outliers can be either univariate or bivariate. An outlier is an extreme data point. An outlier can have a disproportionate influence on a correlation coefficient.

Univariate Outliers

A univariate outlier can be defined as one having an extreme z-score. Z-scores above 3 or 3.5 are often regarded as outliers. In small samples (i.e.., less than 40), a case might be considered an outlier if the Z-score is greater than 2.5. Sometimes univariate outliers are bivariate outliers, but this is not always the case.

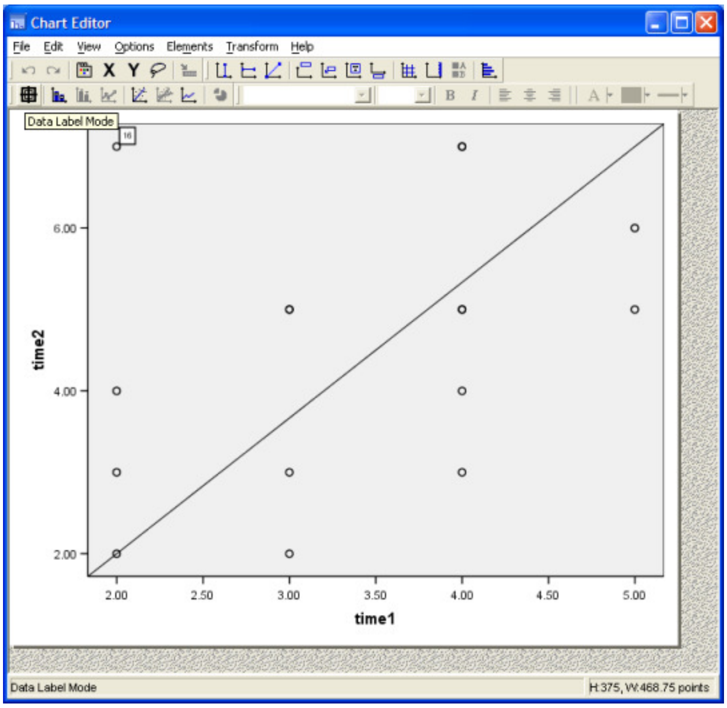
Bivariate Outliers

Bivariate outliers can be assessed by looking at the scatter plot (Graph >> Scatter/Dot). A case that is a lot further from the line of best fit than other cases may be a bivariate outlier.

In the example below I added a case to an existing data file on job satisfaction. The person I added had very low job satisfaction at time one and very high job satisfaction at time 2. This is unusual combination as most people remain somewhat stable in their job satisfaction over time. This case can be identified as the one in the top left hand corner. It is the case that is the furthest away from the line of best fit.

To show the line of best fit, go to: Elements >> Fit Line at Total

To determine which case in the data file a particular outlier data point is, go to: Elements >> Data Label Mode; then click on the data point. This will bring up the row in the data file represented by the data point.



After identifying an outlier, you may want to consider the reason why it is an outlier. You may consider excluding this case from analyses or adjusting its score to a less extreme value.

Range restriction

Frequently in our studies we do not have a full range of population values when performing a correlation. If we examine the relationship between ability and performance of current staff, we are likely to have range restrictions for ability. This is because all the people who would have lower ability and would not be able to get the job have been excluded. Obtaining a correlation between two variables where one of the variables has been measured with a restricted range may increase or decrease the correlation that is obtained relative to the unrestricted range. When generalising findings about a correlation to a broader population, we have less confidence in our generalisations to populations with a different range of values on one of the variables.

Continuous or binary variables

Correlations can be obtained between interval, ratio, and binary variables. You can not perform correlations on categorical data, where there are three or more categories.

Correlation is not causation

The presence of a correlation does not mean that one variable causes the other. The classic example of a non-causal correlation is that between ice cream consumption and the number of people who drown on a day. We would not infer that eating ice creams causes people to drown. Rather, it is likely that a third variable is explaining the relationship (e.g., hot weather is leading people to eat more ice creams and go to the beach more).

SPSS

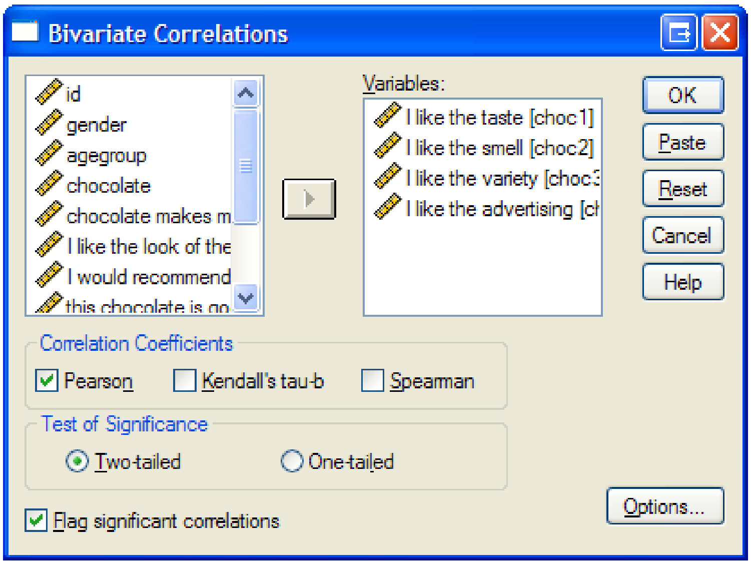
Overview of Study

A market research study was exploring consumer attitudes towards a particular brand of chocolate, participants were asked about the degree to which they agreed with eight different statements regarding the chocolate. The researcher wanted to know whether the responses to one item were related to another items. Did people who like the taste of the chocolate, also like the smell.

Running

Analyze >> Correlate >> Bivariate

Place the variables you want to correlate in the "variables" box. You can choose to flag statistically significant correlations. You might also want to examine Spearman's rho, if you are concerned about outliers or normality.



Bivariate Correlations

id
gender
agegroup
chocolate
chocolate makes m
I like the look of the
I would recommend
this chocolate is go

Variables:
I like the taste [choc1]
I like the smell [choc2]
I like the variety [choc3
I like the advertising [ch

Correlation Coefficients
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

Output

The output below shows the correlation matrix for four items on liking of the brand of chocolate. Each cell of the table contains a correlation, significance and sample size for the two variables being explored. For example, the correlation between liking of taste and smell is .550 with a sample size of 250 and this is statistically significant at .01.

Correlations

		I like the taste	I like the smell	I like the variety	I like the advertising
I like the taste	Pearson Correlation	1	.550**	.550**	.575**
	Sig. (2-tailed)		.000	.000	.000
	N	250	250	250	250
I like the smell	Pearson Correlation	.550**	1	.549**	.503**
	Sig. (2-tailed)	.000		.000	.000
	N	250	250	250	250
I like the variety	Pearson Correlation	.550**	.549**	1	.544**
	Sig. (2-tailed)	.000	.000		.000
	N	250	250	250	250
I like the advertising	Pearson Correlation	.575**	.503**	.544**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	250	250	250	250

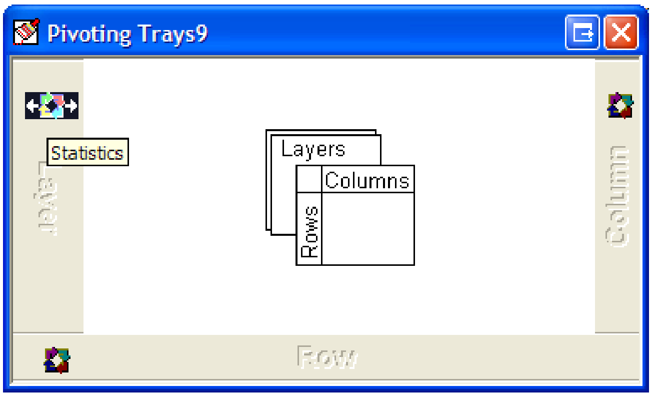
**. Correlation is significant at the 0.01 level (2-tailed).

You could improve this output by hiding the exact p-values and N for each cell and just showing the correlation matrix. To do this:

Double click on the table

Go to: Pivot >> Pivot Trays

Drag and drop the Table Element representing the statistics from the rows section to the Layer section



The result should look the table below.

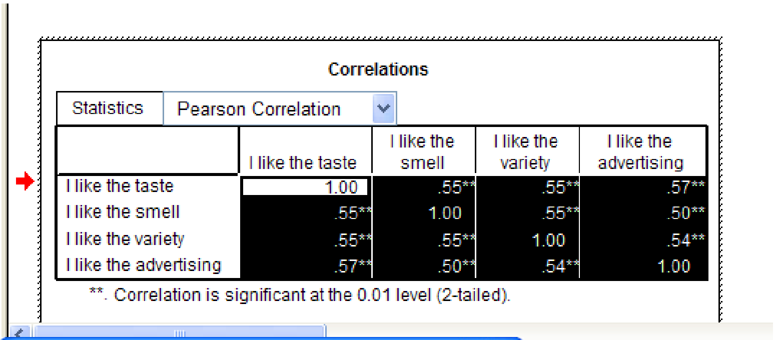
Correlations

Pearson Correlation

	I like the taste	I like the smell	I like the variety	I like the advertising
I like the taste	1	.550**	.550**	.575**
I like the smell	.550**	1	.549**	.503**
I like the variety	.550**	.549**	1	.544**
I like the advertising	.575**	.503**	.544**	1

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

You might then highlight all the cells and right click and go to 'Cell Properties' and select only two decimal places as this is a common convention for when reporting correlations.

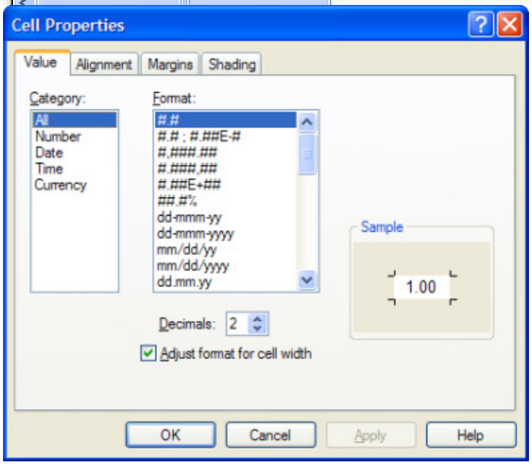


Correlations

Statistics: Pearson Correlation

	I like the taste	I like the smell	I like the variety	I like the advertising
I like the taste	1.00	.55**	.55**	.57**
I like the smell	.55**	1.00	.55**	.50**
I like the variety	.55**	.55**	1.00	.54**
I like the advertising	.57**	.50**	.54**	1.00

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



Cell Properties

Category: All

Format: #.#

Decimals: 2

☑ Adjust format for cell width

Sample: 1.00

15. Chi-Square

Overview

In some situations we are concerned with explaining proportions and counts. When we have two categorical variables, we often want to know whether one variable is related to another variable. Chi-square can be used to see if there is a relationship between two categorical variables. The question we are asking is: does being in one group affect the probability of being in another group in another variable?

There are two common applications of the chi square test to categorical data. First, we can compare the observed counts

Goodness of fit

Overview

Imagine we wanted to know whether there were more days of rain in a particular month. If the month did not make a difference on the number of days of rain, then we would expect an equal number of days of rain in each month. The chi-square goodness of fit test assesses whether the observed counts depart in statistically significant way from this expectation.

The chi-square goodness of fit test involves one categorical variable. It tests whether observed counts depart significantly from expected counts. Observed counts are obtained from the raw data. Expected counts involve multiplying the expected probability by the observed total count. For example, in the case of days of rain where the null hypothesis is that there are equal days of rain in each month. This means that the expected probability of each month is 1 divided by 12 equals 0.083. If there were 100 total days of rain for the year, the expected count for each month would be 0.083 * 100 equals 8.3. The values for observed and expected are then applied in the formula below.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o is the observed cell frequency

f_e is the expected cell frequency

Degrees of freedom are the number of categories minus one.

As the discrepancy between observed and expected gets larger, chi square gets bigger and the differences observed are more likely to be statistically significant. We are not limited to expected counts that are equal. It is common to compare categories to some known expectations. For example, if it was known from previous years what the expected number of days of rain was for each month, the current years data could be compared to these expectations.

Assumptions

Overview

Chi-square is a non-parametric test and as such it does not make particular assumptions about the distribution of the data. The following assumptions apply equally to chi-square goodness of fit test and test for independence.

Expected cell count greater than 5

It is an assumption that the expected cell count for each cell be greater than 5. SPSS will print a warning if this assumption is not satisfied. It is also possible to examine the expected count directly.

Independence of observations

This assumption is that each observation is not effected by other observations. This will generally be satisfied. It is an assumption related to experimental design.

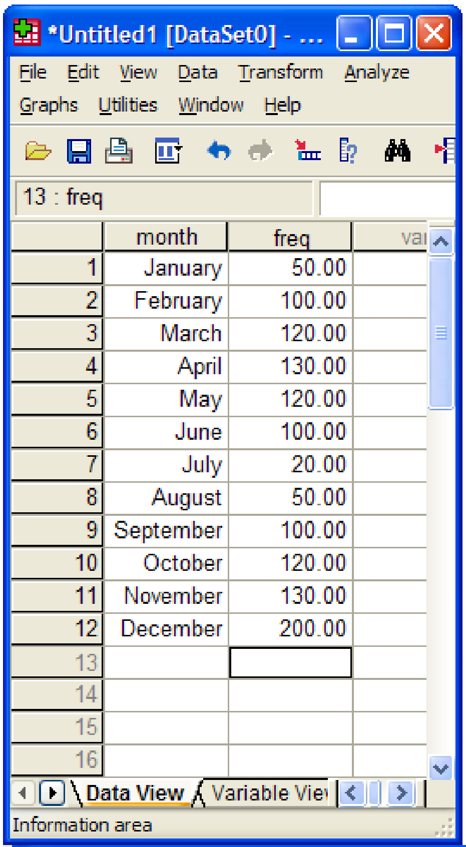
SPSS

Overview of Study

Imagine you ran a shoe store and were interested to see whether the sales for a particular year were significantly different across months. Could the variation in shoe sales be explained by random variation, or is a better explanation that there were systematic factors explaining the variability?

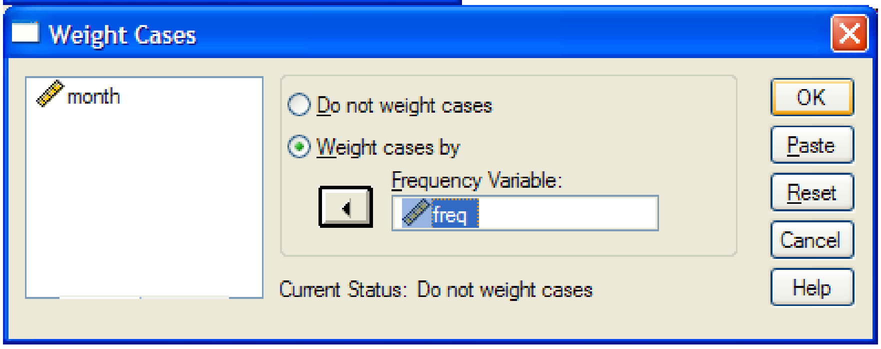
Data structuring

There are two ways to represent category counts in SPSS. The standard way is that every score represents a row in the SPSS data file. In the present example, this would mean that every row represents a shoe sale. In each row the month where the sale was recorded would be noted. An alternative way of structuring the data can often be more efficient, particularly if you are working with data that has already been summarised into totals for each category. In this case each row represents a category and one variable reflects the category number or name and another variable represents the frequency which the category occurred. We then tell SPSS to weight the data file by the frequency variable


*Untitled1 [DataSet0] - ...
File Edit View Data Transform Analyze
Graphs Utilities Window Help
Data View
Variable View
Information area

13 : freq

	month	freq	var
1	January	50.00	
2	February	100.00	
3	March	120.00	
4	April	130.00	
5	May	120.00	
6	June	100.00	
7	July	20.00	
8	August	50.00	
9	September	100.00	
10	October	120.00	
11	November	130.00	
12	December	200.00	
13			
14			
15			
16			



○ Do not weight cases

● Weight cases by

Frequency Variable: freq

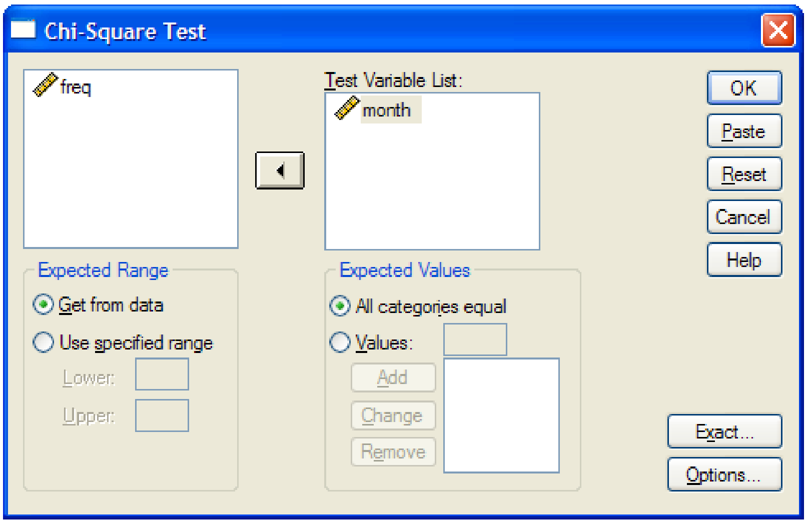
Current Status: Do not weight cases

Chi square

Analyze >> Nonparametric Tests >> Chi Square

Place categorical variable into ‘test variable list’ box

Under expected values: in this case, the null hypothesis is that all categories are equal. Alternatively you could specify the expected cell counts or percentages. In this case you enter each predicted percentage in the order of the values for the categorical variable.



Chi-Square Test

freq

Test Variable List: month

Expected Range

● Get from data

○ Use specified range

Lower:

Upper:

Expected Values

● All categories equal

○ Values:

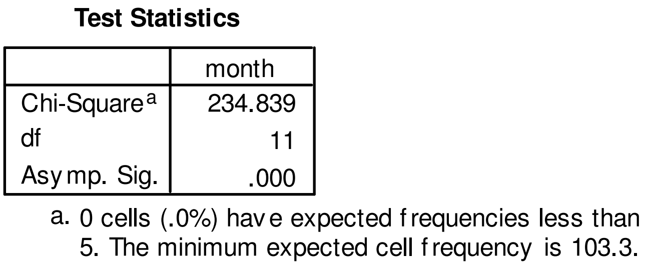
Output

The first piece of output shows the observed number of sales for each month and the number of sales expected based on the uniform sales per month hypothesis. The residual is the difference between observed and expected. Negative sales reflect months where there were fewer sales than expected, and positive residuals reflect months where there were more shoe sales than expected. All else being equal, larger residuals leads to a greater likelihood that the subsequent chi-square test will be statistically significant. Residuals can also be used to indicate where the particular differences from the uniform sales hypothesis are occurring. In the present example, January, July and August are periods of substantially lower sales and December is a period of substantially higher sales.

month

	Observed N	Expected N	Residual
1.00 January	50	103.3	-53.3
2.00 February	100	103.3	-3.3
3.00 March	120	103.3	16.7
4.00 April	130	103.3	26.7
5.00 May	120	103.3	16.7
6.00 June	100	103.3	-3.3
7.00 July	20	103.3	-83.3
8.00 August	50	103.3	-53.3
9.00 September	100	103.3	-3.3
10.00 October	120	103.3	16.7
11.00 November	130	103.3	26.7
12.00 December	200	103.3	96.7
Total	1240		

The next table tests the null hypothesis that all categories have equal probabilities in the population. Using a typical alpha of .05, we can see that the chi-square is significant. This could be written like this: χ^2(df = 11, n = 1240) = 234.8, p < .001.



Test Statistics

	month
Chi-Square[a]	234.839
df	11
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 103.3.

Test for Independence - Two-way frequency tables

Overview

A second application of the chi-square test is two two-way frequency tables. This analysis is used when there are frequencies on two categorical variables. It is used to ask the question: Is there an association between two categorical variables? Does knowing the category of one variable increase the probability of being in a particular category on a second categorical variable?

Two-way frequency tables are an excellent way of exploring the relationship between two categorical variables.

Chi-square test for independence is similar to chi-square goodness of fit in the sense that it is based on discrepancies between observed and expected. The difference is that in chi-square test for independence the expected values are based on row and column probabilities rather than some predefined probabilities.

$$f_e = \frac{f_r \times f_c}{n}$$

F_r is the row frequency

F_c is the column frequency

n is the total frequency in the table

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o is the observed cell frequency

f_e is the expected cell frequency

The same assumptions of expected cell count greater than 5 and independence of observations mentioned for single categorical variables also hold for tests for independence.

SPSS

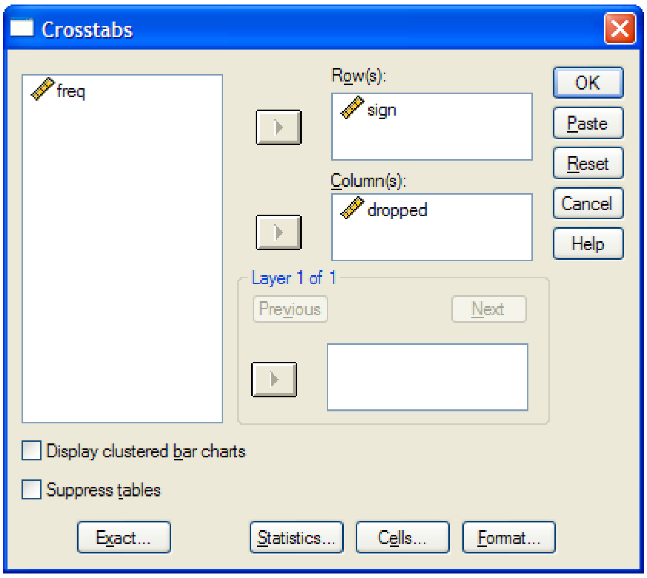
Case Study

Imagine that a littering was a problem in a particular city. An intervention was proposed by government whereby signs would be erected to remind people to not litter. Before rolling out the program, the government wanted to know whether the signs were effective at reducing littering. A study was designed whereby leaflets were handed out to people on the street either in a context where they would see the sign or where there was no sign discouraging littering. The dependent variable was whether the leaflet was dropped.

Running

Analyze >> Descriptives >> Crosstabs

Place one of the variables in rows and one in columns.



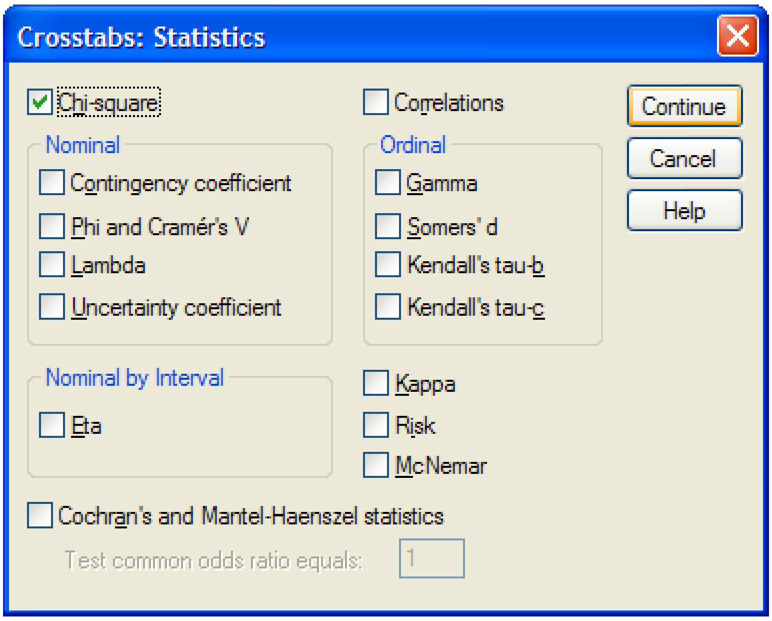
Row(s): sign

Column(s): dropped

□ Display clustered bar charts

□ Suppress tables

Click ‘statistics’ and select chi-square



☑ Chi-square

□ Correlations

□ Contingency coefficient

□ Phi and Cramér's V

□ Lambda

□ Uncertainty coefficient

□ Gamma

□ Somers' d

□ Kendall's tau-b

□ Kendall's tau-c

□ Eta

□ Kappa

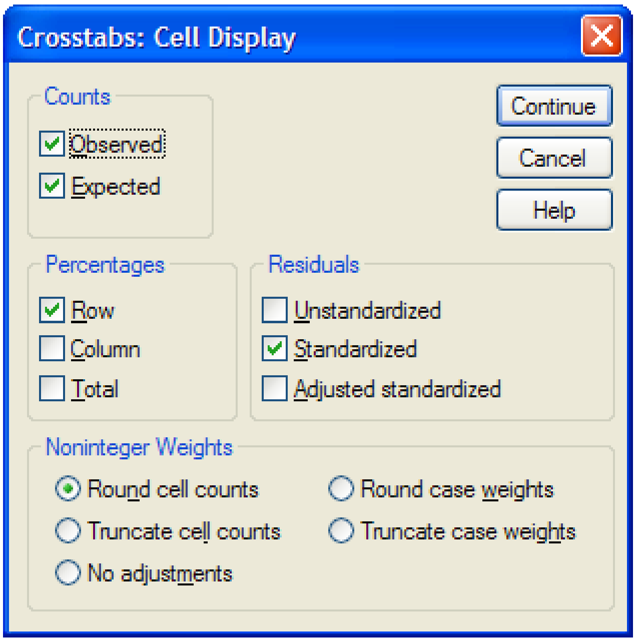
□ Risk

□ McNemar

□ Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals: 1

You might also click Cells and request a range of additional information such as expected counts, row percentages and standardized residuals.



Crosstabs: Cell Display

Counts

☑ Observed

☑ Expected

Percentages

☑ Row

□ Column

□ Total

Residuals

□ Unstandardized

☑ Standardized

□ Adjusted standardized

Noninteger Weights

● Round cell counts

○ Truncate cell counts

○ No adjustments

○ Round case weights

○ Truncate case weights

Output

This first table shows the raw counts, expected count, row percentages and standardised residuals. The raw counts highlight that dropping the letter was more common when there was no sign. The row percentages are particularly useful in highlighting the trend showing the 2 thirds of people in the 'no sign' condition dropped the paper but only half in the 'sign' condition dropped the paper. Standardised residuals provide a way of understanding where potential significant deviations from expectations might be occurring. Standardised residuals greater than plus or minus 2 are in some senses a statistically significant deviation.

It is not clear whether this could just be explained by random sampling (i.e., chance) or whether a better explanation is that the sign was effective. To find the answer to this question, we turn to the chi-square test.

sign * dropped Crosstabulation

			dropped		
			1.00 dropped	2.00 did not drop	Total
sign	1.00 Sign	Count	30	30	60
		Expected Count	35.0	25.0	60.0
		% within sign	50.0%	50.0%	100.0%
		Std. Residual	-.8	1.0	
	2.00 No sign	Count	40	20	60
		Expected Count	35.0	25.0	60.0
		% within sign	66.7%	33.3%	100.0%
		Std. Residual	.8	-1.0	
Total		Count	70	50	120
		Expected Count	70.0	50.0	120.0
		% within sign	58.3%	41.7%	100.0%

This table provides a range of different statistical tests. The typical one to focus on is the Pearson Chi-square value and the associated significance level. In this case the chi-square was approaching, but was not quite significant at .05.

Chi-Square Tests

	Value	df	Asy mp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.429[b]	1	.064		
Continuity Correction[a]	2.777	1	.096		
Likelihood Ratio	3.447	1	.063		
Fisher's Exact Test				.095	.048
Linear-by -Linear Association	3.400	1	.065		
N of Valid Cases	120				

a. Computed only f or a 2x2 table

b. 0 cells (.0%) hav e expected count less than 5. The minimum expected count is 25. 00.

16. MULTIPLE REGRESSION

Theory

Overview

Multiple regression is used when there are one or more predictor variables and one continuous dependent variables. It allows for the testing of a theoretical model about how a set of predictors account for a particular outcome variable of interest. It can be used to assess the relative importance of predictor variables. It develops a weighted linear composite that maximises the variance accounted for in the dependent variable.

The simplest version of a regression equation looks like this:

y = a + bx + e

Where

y = score on dependent variable

a = predicted value of y when x is zero

b = regression coefficient; increase in y expected for each unit increase in x

e = difference between what is predicted an individual cases actual score

Simple regression is when there is only one predictor variable. In this instance, regression is quite similar to correlation.

The dependent variable should be continuous or at least ordinal. Predictor variables can be continuous or binary. Nominal predictors need to be dummy coded before they can be used as predictors.

In multiple regression the model can include multiple predictor variables. In abstract it looks like this:

y = a + b1bx1 + b2x2 + ... + bpxp + e

where we have as many pairs of and x as there are predictors (p).

R squared

R-squared describes the percentage of variance in the dependent variable explained by the regression equation. It ranges from 0 to 1. It is used to evaluate the effectiveness of the regression to predict.

R-squared somewhat overestimates the population variance explained in the dependent variable by the regression model. The degree of this overestimation is greater when there are more predictors and when the sample size is smaller. Adjusted r-squared attempts to obtain an unbiased estimate of variance explained in the population.

An ANOVA test is used to test for a significant r-squared. The p-value obtained is the probability of getting the r-square value in the sample if there was no relationship in the population. The larger r-squared, the bigger the sample size and the smaller the number of predictors, the more powerful this test will be.

Regression coefficient

Unstandardised coefficients reflect the predicted increase in the dependent variable for each unit (i.e., increase by one) of the dependent variable. They are more meaningful when the scale of the independent and dependent variable are meaningful (e.g., height, weight, cost in dollars, years, etc.).

Standardised coefficients reflect the unstandardised coefficients that would be obtained if all variables were converted to z-scores. They can be interpreted as the effect on the dependent variable in standard deviations that would be predicted for a one standard deviation increase on the predictor. They have the benefit of not being dependent on the specific scale that was used in the study. For this reason, the relative importance of each predictor can often be gauged more effectively by examining the standardised as opposed to the unstandardised coefficients.

A t-test is presented for each regression coefficient to test whether it is significantly different from zero. It has the same degrees of freedom as the residual from the overall ANOVA.

Standard error of the estimate

The regression equation makes a prediction for each case. Each case will have a residual or error associated with it. The residual reflects the difference between the observed value and the prediction. The standard error of the estimate is effectively the standard deviation of errors. Thus, when you make a prediction, the standard error of the estimate gives you a sense of the typical amount of error to expect. The standard deviation of the dependent variable gives information of error in prediction when only knowing the mean of the dependent variable. If the standard error of the estimate is substantially less than the standard deviation, then the model represents an improvement in prediction.

Zero-order, partial and semi-partial (part) correlations

Multiple regression describes the relationship between a weighted sum of predictors and a dependent variable. Correlation describes the relationship between just two variables. In SPSS you can obtain three different types of correlations, which are useful intrinsically and as an additional aid to interpretation of the multiple regression. These are called zero-order, partial, and semi-partial (part) correlations.

Zero-order correlations are the standard correlation measure that you would obtain if you ran a correlation in SPSS. It describes the relationship between two variables. In the context of multiple regression output, this is typically the relationship between a predictor variable and the dependent variable independent of the regression model. Before discounting a variable that may not have a significant regression coefficient in a model, it is worth looking at the zero-order correlation to see what the relationship is between the predictor and the model independent of the model.

Partial correlations are correlations between two variables (X and Y) that adjust variables X and Y for other variables. The variance of X and Y that is shared with the other variables is removed. These adjusted versions of X and Y are then correlated to form the partial correlation. This can be useful when wanting to explore the relationship between variables after controlling for other extraneous variables.

Semi-partial correlations which are labelled "part" in SPSS are similar to partial correlations. The difference is that it is only the predictor variables that is adjusted for the influence of the other predictors. The dependent variable is not adjusted. This leads the semi-partial correlation to reflect the relationship with the dependent variable that is unique to the particular predictor variable. A common step is to square the semi-partial correlation. The squared semi-partial correlation has the nice interpretation of being the unique variance explained in the dependent variable by the predictor. This is commonly used to assess the relative importance of a predictor in the particular multiple regression.

Assumptions

Independence of observations

It is assumed that each case in the regression is independent of every other case.

Normality

The statistics for testing the significance of regression coefficients are reasonably robust to modest violations of normality, especially with larger samples.

Normality of Residuals

Normally distributed residuals are a desirable property for a regression model. This can be examined by looking at histogram of standardised residuals. Regression is relatively robust to modest violations of the assumption. It is only at relatively severe violations of normality that this should present a major problem to the validity of inferences.

Homoscedasticity

Homoscedasticity is the extension to multiple regression of the homogeneity of variance assumption in ANOVA and t-tests. It is an assumption of multiple regression that residuals are spread around the regression equation prediction evenly for all levels of prediction. Violations of homoscedasticity may indicate better prediction at certain points of the distribution. It may also result from non-linear relationships between predictors and the dependent. The assumption is typically examined by plotting standardised predicted values on the x-axis and standardised or studentised residuals on the y-axis. The absence of any systematic pattern supports the assumption of homoscedasticity. If the plot shows a u-shaped pattern or a fanning effect, this may indicate violation of the assumption.

Multicollinearity and Singularity

The ideal regression is one where there are strong relationships between the predictor variables and the dependent variables but week to no correlations between predictor variables. Correlations between independent variables can make it more difficult to assess the relative importance of each predictor variable. If correlations get to large, assessing the relative importance of regression coefficients can become impossible. Singularity occurs when the other predictors perfectly predict another independent variable. To assess for excessive multicollinearity, correlations between predictor variables are examined. Correlations between independent variables that are large (i.e., greater than .8) may present problems for regression coefficient interpretation. Tolerance and Variance Inflation Factor (VIF) is also examined to assess multicollinearity. Tolerance represents the unique variance of a particular predictor variable that is not shared with other predictors. If you were to run a regression predicting the predictor variable from all other predictors, the variance not accounted for by the other predictors, which in this case would be 1 – r-squared, is the value of tolerance. Tolerance ranges from 0 to 1 with values closer to 1 being preferable. In terms of rules of the thumb, values less than .1 suggest major problems, and less than .2 minor problems. VIF is the inverse of tolerance (i.e., 1 / tolerance). If one or more predictor variables are identified as having problems of multicollinearity, the strategy is usually to remove one of the predictors that are causing the problem. Alternatively a composite variable can be created with converts the variables with correlations into a single variable.

Problematic cases

Regression can be sensitive to particular influential cases which may bias the regression model that is produced. There are several different types of outliers.

Residuals are the difference between the observed and predicted value for a particular case. Standardised residuals convert residuals into z-scores. Values greater than 2.5 in small samples and 3 in larger samples reflect cases that are not well predicted by the regression model.

Leverage reflects how unusual a case is on the independent variables. It is not related to the dependent variable. Leverage values greater than 3*p/N in small samples or 2*p/N in larger samples, where p is the number of predictors and N is the sample size, are potentially problematic.

Influence is arguably the most important diagnostic of a problematic case. Influence relates to the effect the particular case is having on regression parameters. Influence is commonly expressed in terms of Cook's Distance. Values greater than 1 indicate an excessively influential case.

SPSS allows the saving of these statistics for each case in the data file. The typical procedure is to first assess the minimum and maximum values are all inside acceptable limits. If the minimum or maximum values are outside the recommended range, the case is identified and considered for deletion from the model.

Linearity

By default multiple regression assumes that relationships between predictor and dependent variables is linear. To the extent that there are non-linear relationships present such as higher order polynomials (e.g., quadratic or cubic), the regression coefficients can be misleading. Using more advanced regression procedures it is possible to model such terms. The assumption of linearity can be examined by looking at scatter plots of the predictor on the dependent variable. It is also possible to look at partial correlation plots which show the relationship between a predictor and the dependent variable taking out the effect of the other predictors.

SPSS

Case study

The following data file is taken from J.F. Fraumeni, "Cigarette Smoking and Cancers of the Urinary Tract: Geographic Variations in the United States," Journal of the National Cancer Institute, 41, 1205-1211. http://lib.stat.cmu.edu/DASL/Stories/cigcancer.html

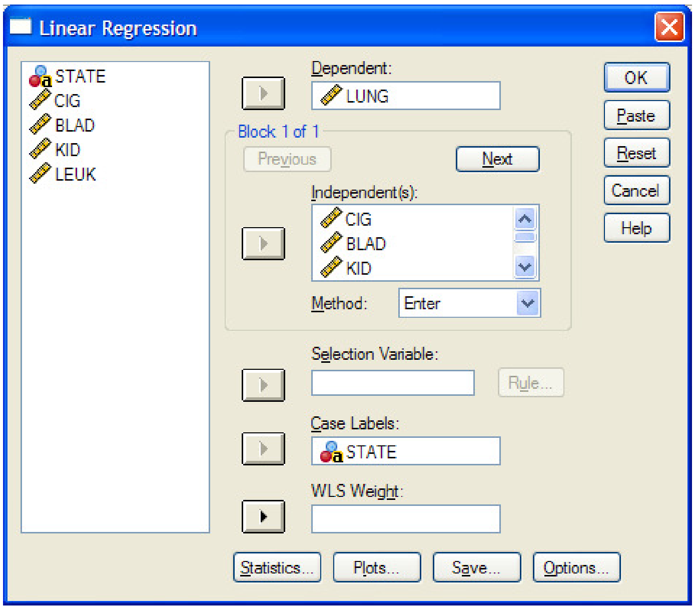
The study obtained data in 1960 on smoking consumption and death rates from lung cancer, bladder cancer, kidney failure and leukaemia across 44 US states. It might be hypothesised that state smoking consumption would be related to lung cancer death rates. It might also be thought that general properties of the state increase incidence of death from other diseases might also predict lung cancer death rates. Thus, a multiple regression could be performed predicting lung cancer death rates from state smoking consumption and the other diseases.

Running

Analyze >> Regression >> Linear

Place the dependent variable in 'dependent'

Place the independent variables in 'independent'



Linear Regression

STATE
CIG
BLAD
KID
LEUK

Dependent: LUNG

Block 1 of 1

Independent(s): CIG, BLAD, KID

Method: Enter

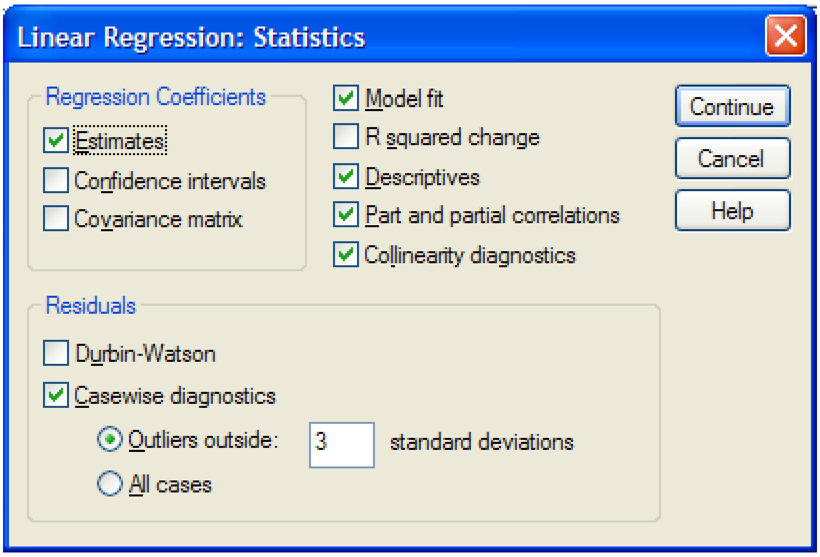
Selection Variable:

Case Labels: STATE

WLS Weight:

Under statistics button select:

'Estimates', 'Model fit', 'Descriptives', 'Part and partial correlations', 'Collinearity diagnostics', and 'Casewise diagnostics'



Regression Coefficients

☑ Estimates

□ Confidence intervals

□ Covariance matrix

☑ Model fit

□ R squared change

☑ Descriptives

☑ Part and partial correlations

☑ Collinearity diagnostics

Residuals

□ Durbin-Watson

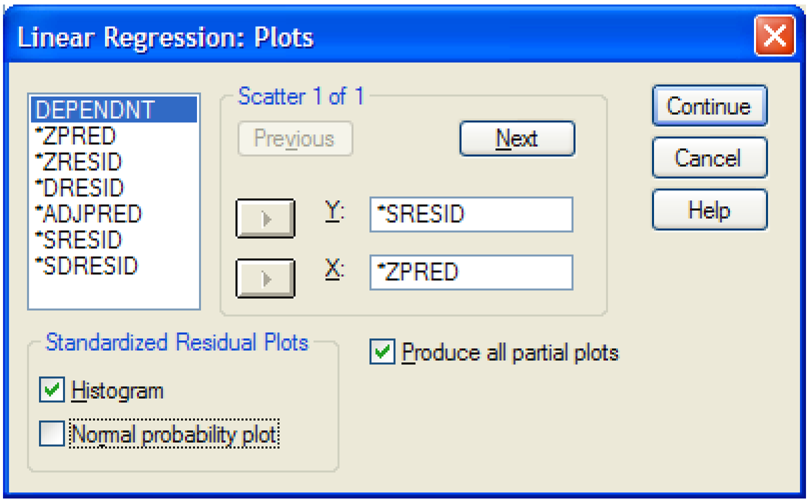
☑ Casewise diagnostics

● Outliers outside: 3 standard deviations

○ All cases

Under plots select:

Histogram, produce all partial plots and create a plot by placing SRESID (studentised residual) on the y axis and ZPRED (standardised predicted) on the x axis.



Scatter 1 of 1

Y: *SRESID

X: *ZPRED

Standardized Residual Plots

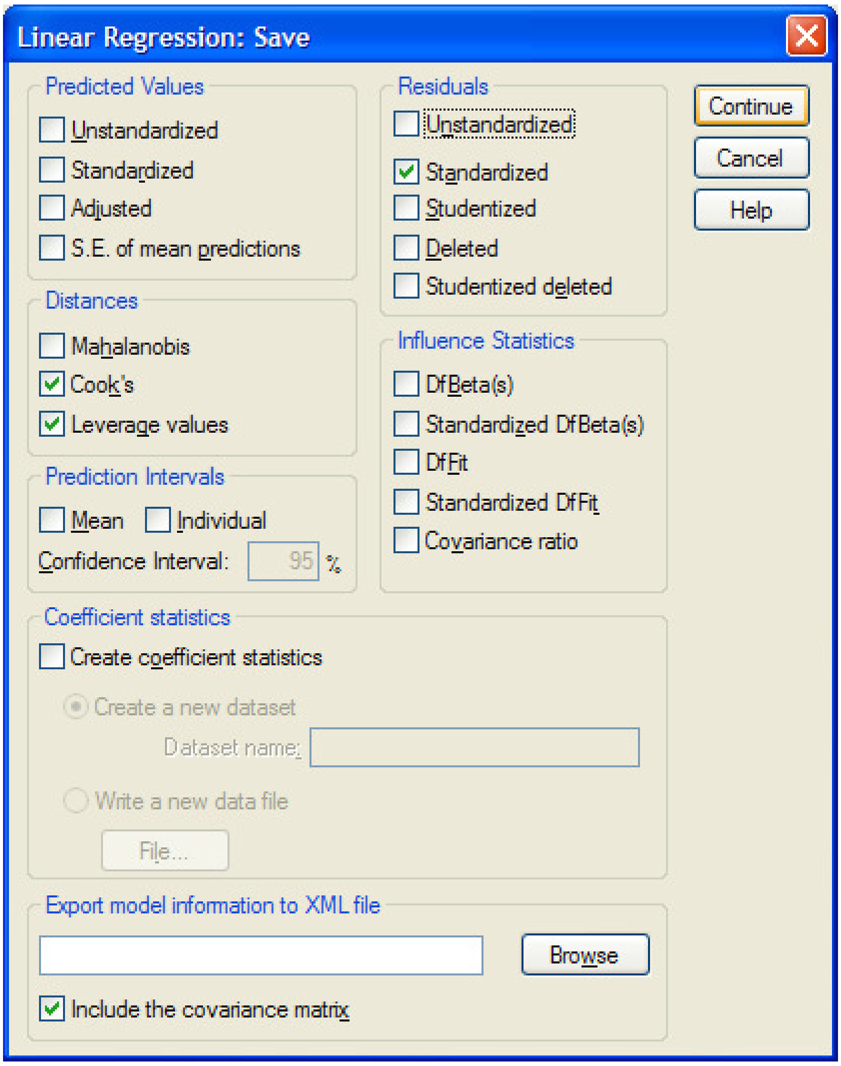
☑ Histogram

□ Normal probability plot

☑ Produce all partial plots

Under save select:

Cook’s D, leverage values and standardized residuals



Linear Regression: Save

Predicted Values
- □ Unstandardized
- □ Standardized
- □ Adjusted
- □ S.E. of mean predictions

Distances
- □ Mahalanobis
- ☑ Cook's
- ☑ Leverage values

Prediction Intervals
- □ Mean □ Individual
- Confidence Interval: 95 %

Residuals
- □ Unstandardized
- ☑ Standardized
- □ Studentized
- □ Deleted
- □ Studentized deleted

Influence Statistics
- □ DfBeta(s)
- □ Standardized DfBeta(s)
- □ DfFit
- □ Standardized DfFit
- □ Covariance ratio

Coefficient statistics
- □ Create coefficient statistics
- ● Create a new dataset
- Dataset name:
- ○ Write a new data file
- File...

Export model information to XML file
- Browse
- ☑ Include the covariance matrix

Output

This first section of output shows the basic descriptive statistics for the sample. This could be examined to see what the average number of deaths per 100 thousand was. It also confirms that the sample size was 44 states.

Descriptive Statistics

	Mean	Std. Deviation	N
Deaths per 100K population from lung cancer	19.6532	4.22812	44
Number of cigarettes smoked (hds per capita)	24.914	5.5733	44
Deaths per 100K population from bladder cancer	4.121	.9649	44
Deaths per 100K population from kidney failure	2.7945	.51908	44
Deaths per 100 K population from leukemia	6.8298	.63826	44

The next piece of output shows the correlations between all the variables in the study. The first row and column is the dependent variable. We want to see some strong correlations between the predictor variables and our dependent variable. In this case number of cigarettes smoked and bladder cancer both have strong correlations with lung cancer. We can also assess the assumption of multicollinearity by looking at the correlations between the predictor variables. None of the correlations are very high, but the correlation between bladder cancer and cigarette smoking (r=.704) might be of some concern.

Correlations

Pearson Correlation

	Deaths per 100K population from lung cancer	Number of cigarettes smoked (hds per capita)	Deaths per 100K population from bladder cancer	Deaths per 100K population from kidney failure	Deaths per 100 K population from leukemia
Deaths per 100K population from lung cancer	1.000	.697	.659	.283	-.152
Number of cigarettes smoked (hds per capita)	.697	1.000	.704	.487	-.068
Deaths per 100K population from bladder cancer	.659	.704	1.000	.359	.162
Deaths per 100K population from kidney failure	.283	.487	.359	1.000	.189
Deaths per 100 K population from leukemia	-.152	-.068	.162	.189	1.000

The model summary summarises how well the overall regression model predicts the dependent variable. In this case the regression model based on the four predictors account for 57.6% of variance. The standard error of the estimate is 2.89 as compared to a standard deviation of 4.2 in the dependent variable. Both these indicators suggest that the model is good at predicting lung cancer death rates in US states.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.759[a]	.576	.533	2.89085

a. Predictors: (Constant), Deaths per 100 K population from leukemia , Number of cigarettes smoked (hds per capita) , Deaths per 100K population from kidney failure, Deaths per 100K population from bladder cancer

b. Dependent Variable: Deaths per 100K population from lung cancer

The ANOVA table shows that the R^2 value above is a statistically significant difference from 0, F(4, 39) = 13.24, p < .001.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	442.788	4	110.697	13.246	.000[a]
	Residual	325.923	39	8.357		
	Total	768.712	43			

a. Predictors: (Constant), Deaths per 100 K population from leukemia , Number of cigarettes smoked (hds per capita) , Deaths per 100K population from kidney failure, Deaths per 100K population from bladder cancer

b. Dependent Variable: Deaths per 100K population from lung cancer

The coefficients table here is shown split over two tables. This first table shows the regression coefficients in unstandardised and standardised form with tests of statistical significance. Based on this table the unstandardised regression equation would be:
Lung = 13.5 + .31*cigarettes + 1.8 * bladder + (-.2) * kidney + (-1.2) * leukaemia
If we look at the standardised predictors, it would appear that cigarettes and bladder cancer were the bigger predictors. If we look at the tests of statistical significance for each of the regression coefficients we can see that only cigarettes and bladder cancer had regression coefficients which showed a statistically significant difference from 0.

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13.502	5.335		2.531	.016
	Number of cigarettes smoked (hds per capita)	.312	.126	.411	2.466	.018
	Deaths per 100K population from bladder cancer	1.797	.674	.410	2.665	.011
	Deaths per 100K population from kidney failure	-.243	1.007	-.030	-.241	.811
	Deaths per 100 K population from leukemia	-1.221	.751	-.184	-1.627	.112

a Dependent Variable: Deaths per 100K population from lung cancer

The second half of the coefficients table shows the different correlations between predictors and the dependent variable. It also includes diagnostic statistics for multicollinearity. The zero-order correlation replicates part of the table shown above. The part column shows the semi-partial correlations. We can that if we square this we get the unique percentage of variance in lung cancer rates that are predicted by the respective predictor. For example the semi-partial correlation for cigarettes is .257. When we square .257 (i.e., .257*.257), we get .066. We could then conclude that 6.6% of the variance in lung cancer rates is uniquely accounted for by cigarettes.

An examination of the tolerance values show nothing below .1 or below .2. Thus, the assumption of multicollinearity does not appear to be violated in any substantial way. It should be noted that initial correlation between cigarettes and bladder is probably bringing about the relatively low tolerance levels for these variables.

Model		Correlations			Collinearity Statistics	
		Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)					
	Number of cigarettes smoked (hds per capita)	.697	.367	.257	.392	2.552
	Deaths per 100K population from bladder cancer	.659	.392	.278	.459	2.178
	Deaths per 100K population from kidney failure	.283	-.039	-.025	.711	1.407
	Deaths per 100 K population from leukemia	-.152	-.252	-.170	.847	1.181

a Dependent Variable: Deaths per 100K population from lung cancer

The following table shows summary information for the residual statistics. The first step is to look at the maximum and minimum values to see if there are any cases outside acceptable limits. If there is, we will need to examine the data file to see which cases may be causing the problem

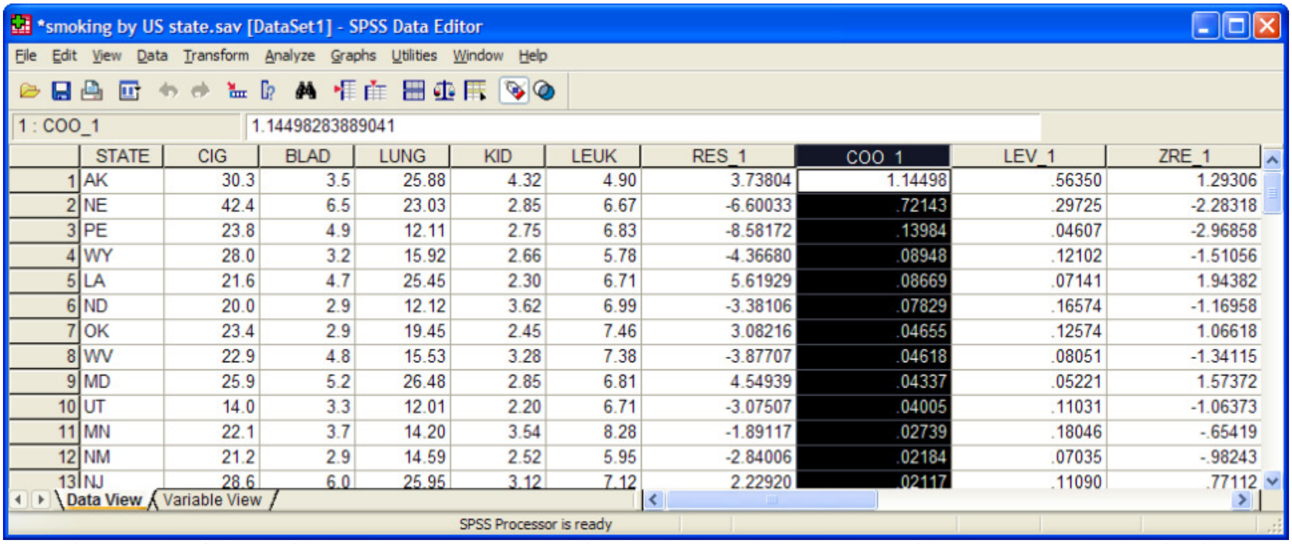
An examination of the standardised residuals shows that at least one case has a standardised residual less than -2.5. An examination of Cook's distance shows that there is at least one case has a value greater than 1. The critical value for leverage was 2*p/N, 2*4/44=.18. There was at least one case with a leverage value above .18.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	15.0851	29.6303	19.6532	3.20896	44
Std. Predicted Value	-1.424	3.109	.000	1.000	44
Standard Error of Predicted Value	.499	2.213	.924	.315	44
Adjusted Predicted Value	15.5553	32.7361	19.6270	3.40979	44
Residual	-8.58172	5.61929	.00000	2.75311	44
Std. Residual	-2.969	1.944	.000	.952	44
Stud. Residual	-3.076	2.042	.001	1.047	44
Deleted Residual	-9.70610	9.03396	.02623	3.42237	44
Stud. Deleted Residual	-3.489	2.133	-.012	1.105	44
Mahal. Distance	.306	24.230	3.909	4.032	44
Cook's Distance	.000	1.145	.060	.200	44
Centered Leverage Value	.007	.563	.091	.094	44

a. Dependent Variable: Deaths per 100K population from lung cancer

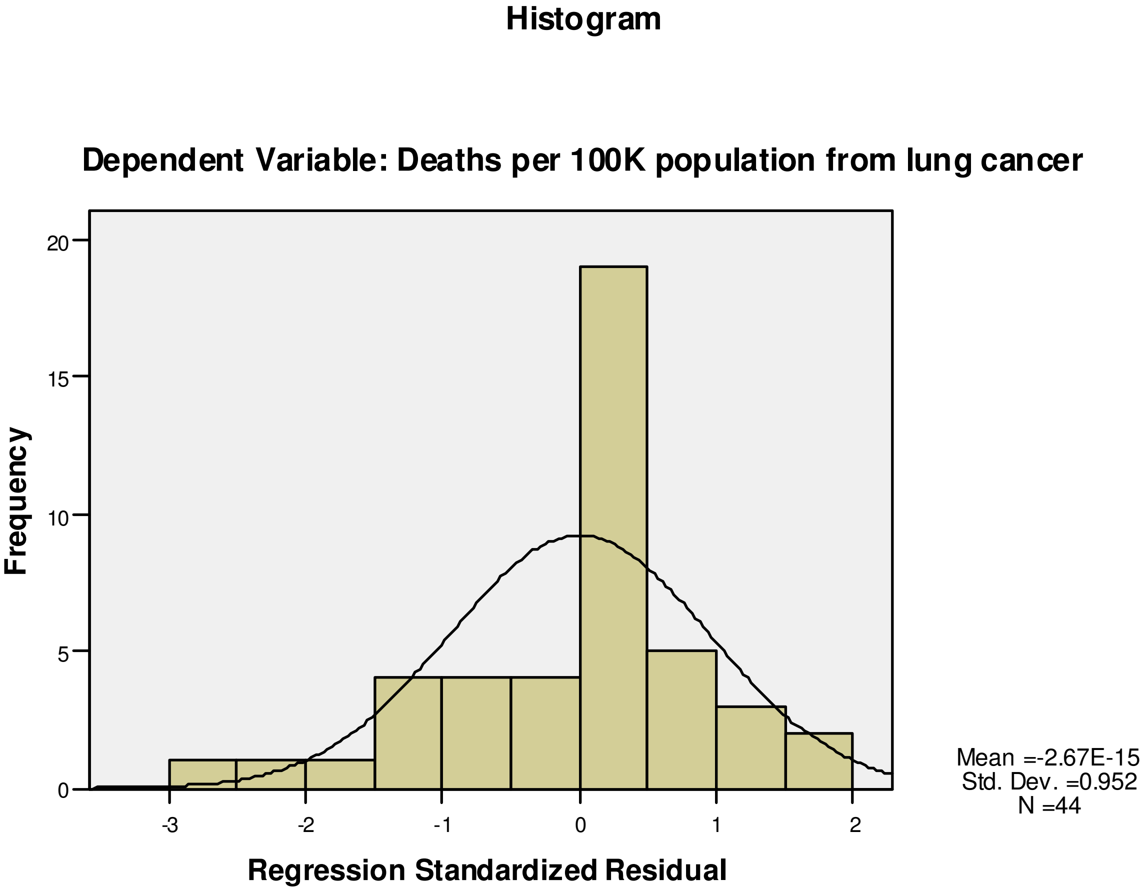
Because there were problematic cases in the data file, we turn now to the data file to which cases were problematic. Because we selected the appropriate variables in the 'Save' box earlier on, SPSS has created influence, leverage and residual statistics for each case. For the example of Cook's distance, we go to the variable and right click and sort descending. We can see below that this has highlighted one case with a Cook's D above 1. This was for the state of Alaska (AK). We would repeat this process for standardised residuals and for leverage. We might then choose to re-run the analysis without Alaska in the model.


*smoking by US state.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

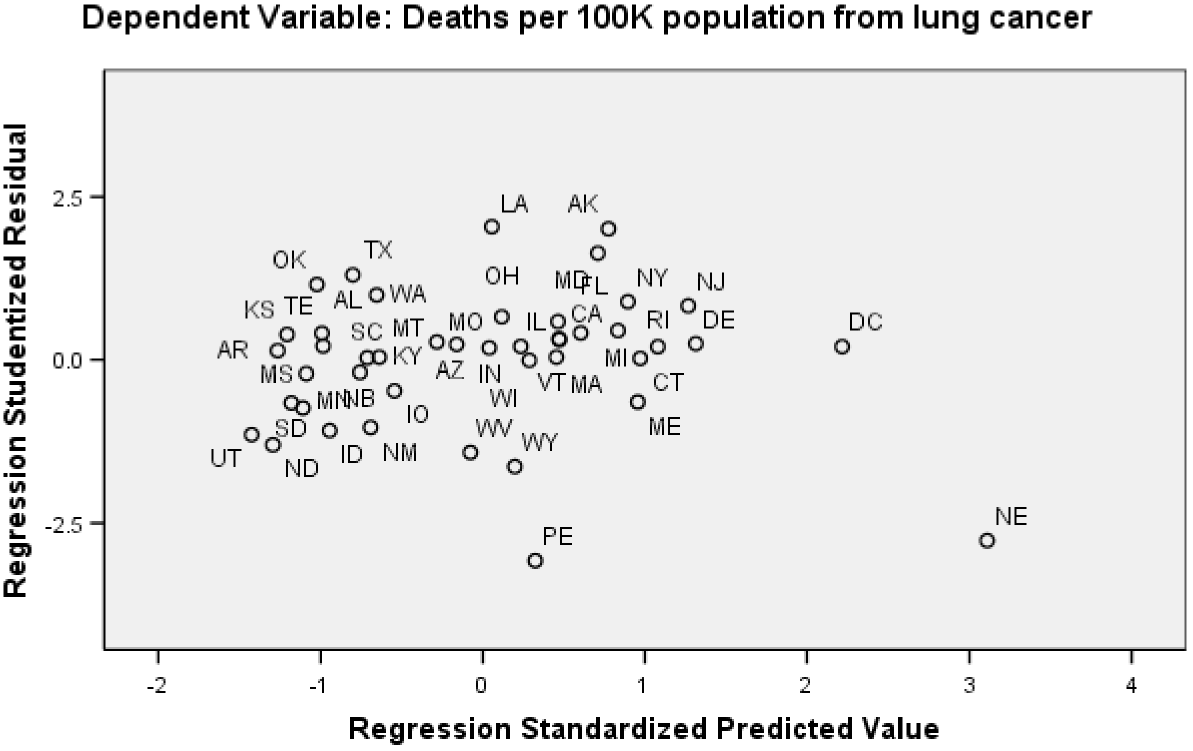
1 : COO_1 1.14498283889041

	STATE	CIG	BLAD	LUNG	KID	LEUK	RES_1	COO_1	LEV_1	ZRE_1
1	AK	30.3	3.5	25.88	4.32	4.90	3.73804	1.14498	.56350	1.29306
2	NE	42.4	6.5	23.03	2.85	6.67	-6.60033	.72143	.29725	-2.28318
3	PE	23.8	4.9	12.11	2.75	6.83	-8.58172	.13984	.04607	-2.96858
4	WY	28.0	3.2	15.92	2.66	5.78	-4.36680	.08948	.12102	-1.51056
5	LA	21.6	4.7	25.45	2.30	6.71	5.61929	.08669	.07141	1.94382
6	ND	20.0	2.9	12.12	3.62	6.99	-3.38106	.07829	.16574	-1.16958
7	OK	23.4	2.9	19.45	2.45	7.46	3.08216	.04655	.12574	1.06618
8	WV	22.9	4.8	15.53	3.28	7.38	-3.87707	.04618	.08051	-1.34115
9	MD	25.9	5.2	26.48	2.85	6.81	4.54939	.04337	.05221	1.57372
10	UT	14.0	3.3	12.01	2.20	6.71	-3.07507	.04005	.11031	-1.06373
11	MN	22.1	3.7	14.20	3.54	8.28	-1.89117	.02739	.18046	-.65419
12	NM	21.2	2.9	14.59	2.52	5.95	-2.84006	.02184	.07035	-.98243
13	NJ	28.6	6.0	25.95	3.12	7.12	2.22920	.02117	.11090	.77112

The histogram of the residuals allows us to graphically assess whether the residuals are normally distributed. In this case, the assumption looks reasonable, but there would appear to be a bit kurtotic with the possibility of a few outliers. Thus, given the small sample, it is difficult to assess exactly.



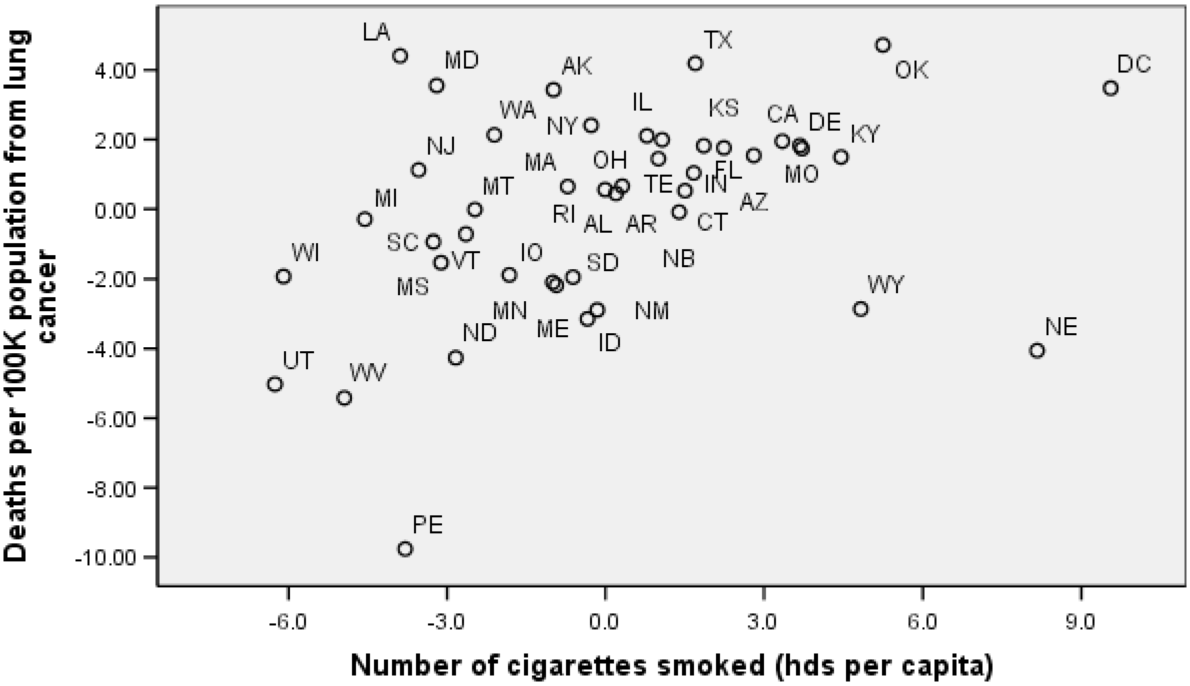
This next plot shows the relationship between predicted scores and residuals. There does not appear to be any systematic patterns. There are a couple of strange cases such as NE and PE but no substantial violation of the general assumption of homoscedasticity.



What follows in the SPSS output are partial regression plots for each of the predictors. For the sake of space, only the one for cigarettes is shown here. We can see here that the general relationship is relatively linear.

Partial Regression Plot

Dependent Variable: Deaths per 100K population from lung cancer



17. FACTOR ANALYSIS & PRINCIPAL COMPONENTS ANALYSIS

Theory

Overview

Factor Analysis and Principal Components Analysis are both used to reduce a large set of items to a smaller number of dimensions and components. These techniques are commonly used when developing a questionnaire to see the relationship between the items in the questionnaire and underlying dimensions. It is also used in general to reduce a larger set of variables to a smaller set of variables that explain the important dimensions of variability. Factor Analysis involves a number of steps and assumptions.

Factorability

The first issue is whether factor analysis is appropriate for the data. An examination of the correlation matrix of the variables used should indicate a reasonable number of correlations of at least medium size (e.g., > .30). A good general summary of the applicability of the data set for factor analysis is the Measure of Sampling Adequacy (MSA). These have the following rules of thumb:

in the .90s	marvellous
in the .80s	meritorious
in the .70s	middling
in the .60s	mediocre
in the .50s	miserable
below .50	unacceptable

If MSA is too low, then factor analysis should not be performed on the data.

Method of Extraction

Principal Components Analysis uses a different mathematical procedure to factor analysis. Factor analysis extraction methods in SPSS include: Maximum Likelihood, Generalised Least Squares, and Unweighted Least Squares. The pros and cons of each method are beyond the scope of this course. The most established is Maximum Likelihood and it is the one recommended for most contexts. If you are curious, try your analysis with different extraction methods and see what effect is has on your substantive interpretation. Frequently in practice, the method of extraction does not make much difference in the results achieved. If you are interested in extracting underlying factors, it would make more sense to use a true factor analytic method, such as Maximum Likelihood. If you want to create a weighted composite of existing variables, principal components may be the more appropriate method.

Number of Dimensions

Overview

There are several approaches for deciding how many factors to extract. Some approaches are better than the others. A good general strategy is to determine how many factors are suggested by the better tests (e.g., scree plot, parallel test, theory). If these different approaches suggest the same number of factors, then extract this amount. If they suggest varying numbers of factors, examine solutions with the range of factor suggested and select the one that appears most consistent with theory or the most practically useful.

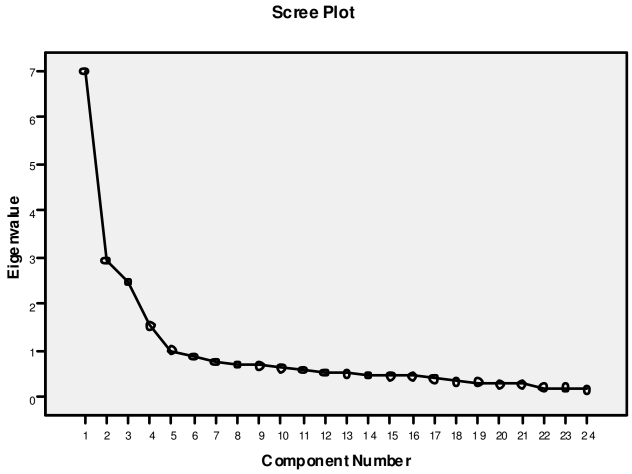
Maximum number of factors

Based on the requirement of identification, it is important to have at least three items per factor. Thus, if you have 7 variables, this would lead to a maximum of 2 factors (7/3 = 2.33, rounded to 2). This is not a rule for determining how many factors to extract. It is just a rule about the maximum number of factor to extract.

Scree Plot

The scree plot shows the eigenvalue associated with each component. An eigenvalue represents the variance explained by each component. An eigenvalue of 1 is equivalent to the variance of a single variable. Thus, if you obtain an eigenvalue of 4, and there are 10 variables being analysed, this component would account for 4 / 10 or 40% of the variance in items. The nature of principal components analysis is that it creates a weighted linear composite of the observed variables that maximises the variance explained in the observed variables. It then finds a seconds weighted linear composite which maximises variance explained in the observed variables, but based on the condition that it does not correlate with the previous dimension or dimensions. This process leads to each dimension accounting for progressively less variance. It is typically assumed that there will be certain number of meaningful dimensions and then a remaining set which just reflect item specific variability. The scree plot is a plot of the eigenvalues for each component, which will often show a few meaningful components that have substantially larger eigenvalues than later components followed which in turn show a slow steady decline. We can use the scree plot to indicate the number of important or meaningful components to extract. The point at which the components start a slow and steady decline is the point where the less important components commence. We go up one from when this starts and this indicates the number of components to extract.

Looking at the figure below highlights the degree of subjectivity in the process. Often it is not entirely clear when the steady decline commences. In the figure below, it would appear that there is a large first component, a moderate 2nd and 3rd component, and a slightly smaller 4th component. From the 5th component onwards there is steady gradual decline. Thus, based on the rule that the 5th component is the start of the unimportant components, the rule would recommend extracting 4 components.



Eigenvalues over 1

This is a common rule for deciding how many factors to extract. It generally will extract too many factors. Thus, while it is the default option in SPSS, it generally should be avoided.

Parallel Test

The parallel test is not built into SPSS. It requires the downloading of additional SPSS syntax to run.

http://flash.lakeheadu.ca/~boconno2/nfactors.html

The parallel test compares the obtained eigenvalues with

Theory

Based on knowledge of the content of the variables, a researcher may have theoretical expectations about how many factors will be present in the data file. This is an important consideration.

Method of Rotation

Rotation serves the purpose of redistributing the variance accounted for by the factors so that interpretation is clearer. A clear interpretation can generally be conceptualised as each variable loading highly on one and only one factor.

Two broad categories of rotation exist, called oblique and orthogonal. Oblique rotations in SPSS are Direct Oblimen and Promax. These allow for correlated factors. Orthogonal rotations in SPSS are Varimax, Quartimax, and Equamax and force factors to be uncorrelated.

The decision on whether to perform an oblique or orthogonal rotation can be influenced by whether you expect the factors to be correlated.

Assumptions

Factorability

See discussion above.

Sample Size

Factor analysis performs better with big samples. As a general rule, factor analysis requires a minimum of around 150 participants in order to get a reliable solution. If correlations between items and the factor loadings are large (e.g., several correlations >.5), sample size can be less and the opposite if the correlations are low. The more items per factor, the fewer participants required.

Normality

Significance tests used in factor analysis assume variables are univariate, bivariate and multivariate normally distributed. Factor analytic solutions may also be improved when normality holds in the data.

Linearity

Factor analysis is based on analyses of correlations and covariances. Correlations and covariances measure the linear relationship between variables. Linear relationships are usually the main forms of relationships for the kinds of purposes that factor analysis is typically applied. If the relationships between variables are non-linear, factor analysis probably is not an appropriate method.

Continuous variables

Factor analysis can be performed on continuous or binary data.

Haywood cases

Haywood cases can occur when computational problems arise when extracting a solution in factor analysis. The main indicator of a Haywood case is an unrotated factor loading that is very close to one (e.g., .99). When this occurs the solution provided should not be trusted. A common cause of Haywood cases is the extraction of too many factors. Thus, a resolution to the problem of Haywood cases is to extract fewer factors. Another resolution is to try a different method of extraction.

Case Study

Imagine an employee climate survey was administered to a large number of employees. The items appeared as follows:

item	Theorised Factor	Description
i1	Job Satisfaction	Overall I am very satisfied
i2	Job Satisfaction	I like my job
i3	Job Satisfaction	I find my job interesting
i4	Job Commitment	I am committed to my job
i5	Job Commitment	I respect the values of the organisation
i6	Job Commitment	I plan to stay with the organisation
i7	Performance	I perform better than the average employee
i8	Performance	I contribute effectively to the organisation's performance
i9	Performance	I add substantial value to the organisation
i10	Leadership	I respect my immediate supervisor
i11	Leadership	My immediate supervisor is very effective
i12	Leadership	Central management are performing well

The survey included 12 items and the researchers had hypothesised that the survey included four factors. The hypothesised relationship between factors and items is shown above. The researchers wanted to find out what the actual factors were and whether their theory was correct.

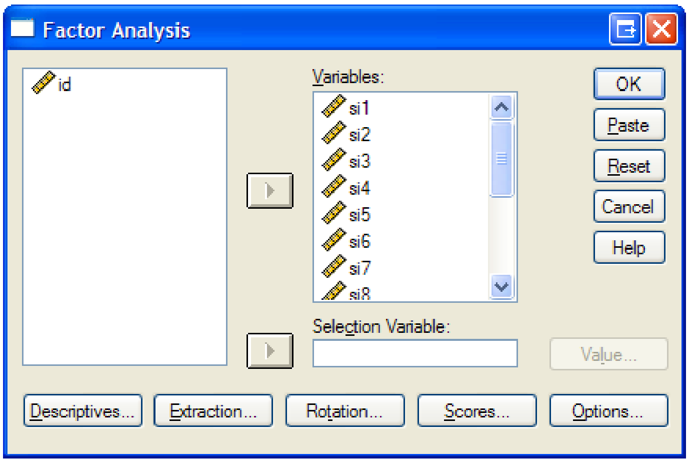
Running & Output

Factorability

The first question to assess before performing factor analysis is to assess its appropriateness to factor analysis. The sample size was 376, which is adequate for factor analysis. The appropriateness of the correlation matrix needs to be assessed.

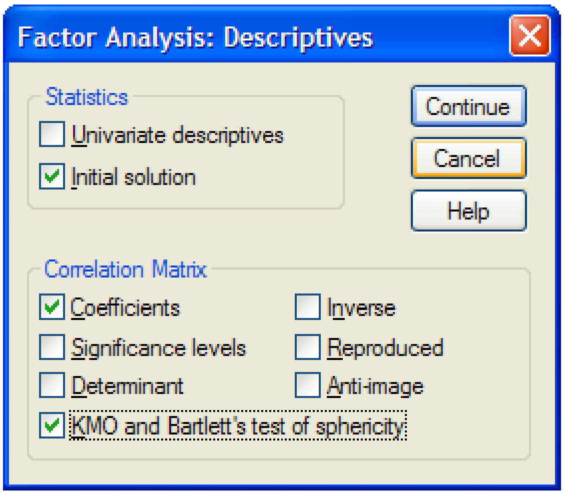
Analyze >> Data Reduction >> Factor

Place all items in the survey in the 'variables' box



Click Descriptives

Select coefficients and KMO and Bartlett's test of sphericity


Factor Analysis: Descriptives

Statistics

☐ Univariate descriptives
☑ Initial solution

Correlation Matrix

☑ Coefficients ☐ Inverse
☐ Significance levels ☐ Reproduced
☐ Determinant ☐ Anti-image
☑ KMO and Bartlett's test of sphericity

The correlation matrix shown below is an abbreviation of the whole matrix. We can see from a glance at the size of the correlations that there is a number of medium to large correlations between items. This suggests factor analysis is appropriate.

		overall I am very satisfied	I like my job	I find my job interesting	I am committed to my job
Correlation	overall I am very satisfied	1.000	.614	.642	.565
	I like my job	.614	1.000	.616	.526
	I find my job interesting	.642	.616	1.000	.598
	I am committed to my job	.565	.526	.598	1.000
	I respect the values of the organisation	.410	.402	.475	.572
	I plan to stay with the	.325	.303	.389	.461

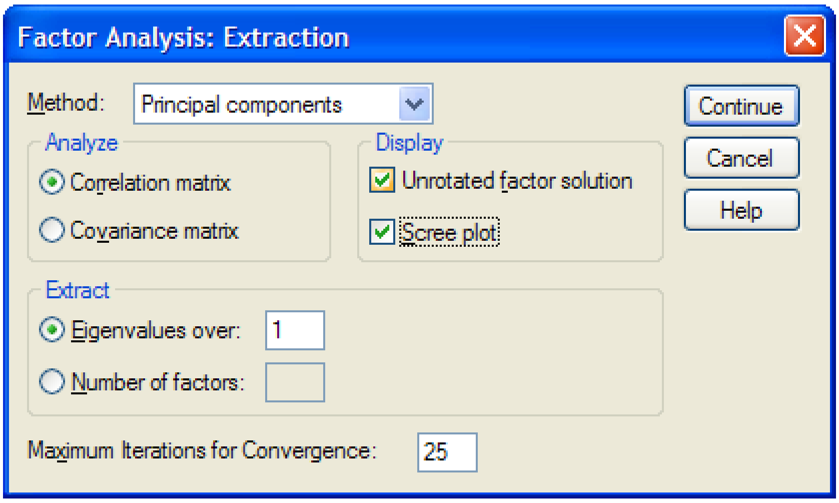
KMO's Measure of sampling adequacy further supports the idea that factor analysis is appropriate for this data.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.878
Bartlett's Test of Sphericity	Approx. Chi-Square	1479.577
	df	66
	Sig.	.000

Number of factors

We then need to determine the appropriate number of factors. We know that theory suggests four factors. We can examine what the scree plot suggests by running the factor analysis again, but this time clicking the 'extraction' button in the factor analysis dialog box and selecting 'Scree plot'.



Factor Analysis: Extraction

Method: Principal components

Analyze

● Correlation matrix

○ Covariance matrix

Display

☑ Unrotated factor solution

☑ Scree plot

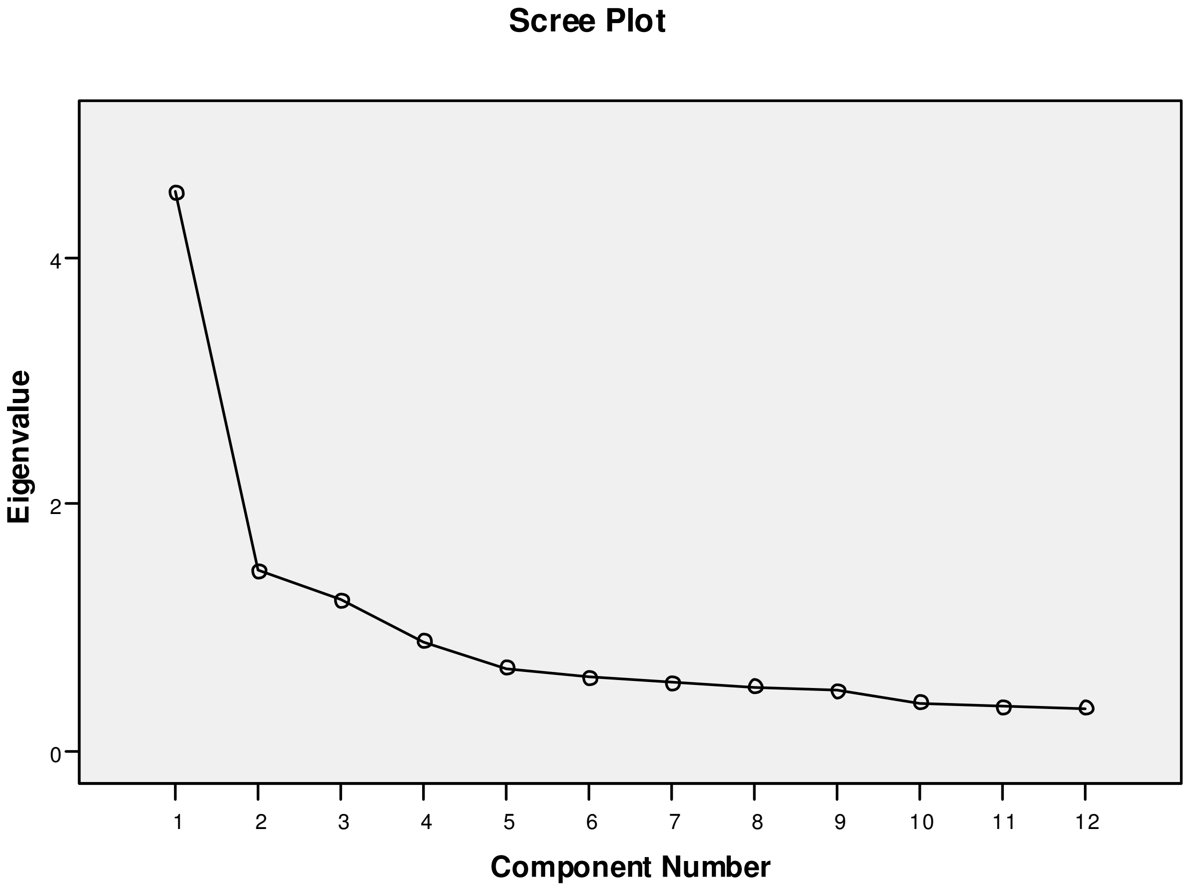
Extract

● Eigenvalues over: 1

○ Number of factors:

Maximum Iterations for Convergence: 25

In addition to the output we previously obtained, we now also get a scree plot. It shows that there is one large factor in the survey. This is followed by a number of smaller factors which maybe of theoretical relevance. It highlights that the scree plot is somewhat subjective in interpretation. It is difficult to determine when the slope moves to the steady decline which would indicate the unimportant components. This may occur at 2, 4 or 5 components, which would suggest 1, 3, or 4 components to be extracted.



A final test we can rely on to guide the decision of number of factors is the parallel test. After downloading it from the website and opening the file we are presented with a large amount of text, some of which we need to modify. The start of the file looks like this:

```
* Parallel Analysis program.

set mxloops=9000 printback=off width=80  seed = 1953125.
matrix.

* enter your specifications here.
```

```
compute ncases   = 50.
compute nvars    = 9.
compute ndatsets = 100.
compute percent  = 95.
```

We need to modify 'ncases' to reflect the number of active participants in our data file (i.e.,376) and 'nvars' to reflect the number of variables in our data file (i.e., 12). The first bit of the file will now look likethis

```
* Parallel Analysis program.

set mxloops=9000 printback=off width=80  seed = 1953125.
matrix.

* enter your specifications here.
compute ncases   = 376.
compute nvars    = 12.
compute ndatsets = 100.
compute percent  = 95.
```

You then click: Run >> All
The output looks like this:

```
Specifications for this Run:
Ncases    376
Nvars      12
Ndatsets  100
Percent    95

Random Data Eigenvalues
         Root        Means     Prcntyle
     1.000000     1.298365     1.357946
     2.000000     1.221398     1.267165
     3.000000     1.161334     1.201760
     4.000000     1.103907     1.139741
     5.000000     1.058563     1.088738
     6.000000     1.014155     1.048447
     7.000000      .972220     1.003783
     8.000000      .924913      .956740
     9.000000      .881952      .910338
    10.000000      .841706      .874425
    11.000000      .790012      .830600
    12.000000      .731476      .774675
```

We then go back to the factor analysis output and compare the above output to the 'Total Variance Explained' table. The rule is that we only take components that have total eigenvalues larger than the column labelled 'Prcntyle', which corresponds to the 95% percentile of eigenvalues for random data. In this case, the third component accounts for 1.224 and the 95th percentile of random data is 1.202. Because this observed eigenvalue is larger than the random data, we would retain at least 3 components. When we go the fourth component, we see that .888 is not larger than 1.1397. Thus, the parallel test recommends 3 components.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.534	37.787	37.787	4.534	37.787	37.787
2	1.457	12.139	49.926	1.457	12.139	49.926
3	1.224	10.198	60.124	1.224	10.198	60.124
4	.888	7.400	67.523			
5	.675	5.623	73.146			
6	.589	4.910	78.056			
7	.549	4.573	82.628			
8	.513	4.278	86.907			
9	.490	4.080	90.987			
10	.386	3.219	94.205			
11	.353	2.939	97.144			
12	.343	2.856	100.000			

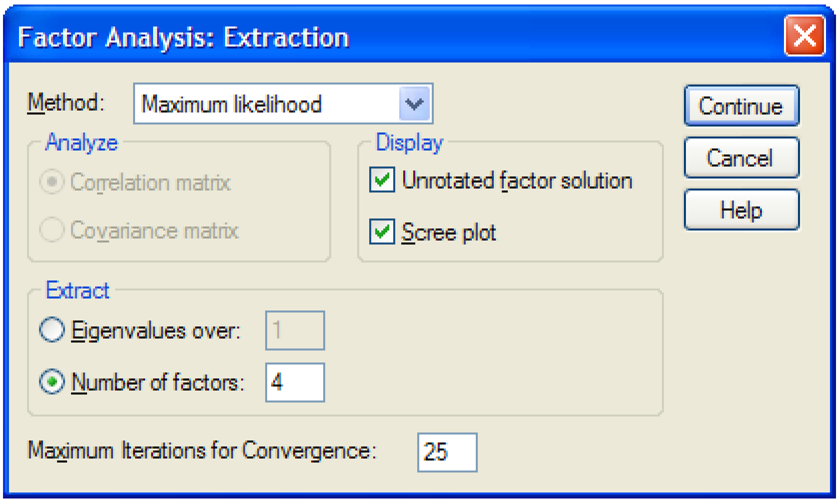
Extraction Method: Principal Component Analysis.

Integrating all this information we have options of 4 factors based on theory, between 1 and 4 factors based on the scree plot and 3 factors based on the parallel test. Thus, it would probably be advisable to explore a number of these options. In particular the 3 and 4 factor solutions would be good candidates. For simplicity, we are only going to test the four factor solution.

Main Analysis

Because we are interested in underlying factors, it is preferable to use a common factor extraction method. The most established of these approaches is arguably maximum likelihood.

This is specified in the Extraction window. We will also specify the number of factors to be 4.



Factor Analysis: Extraction

Method: Maximum likelihood

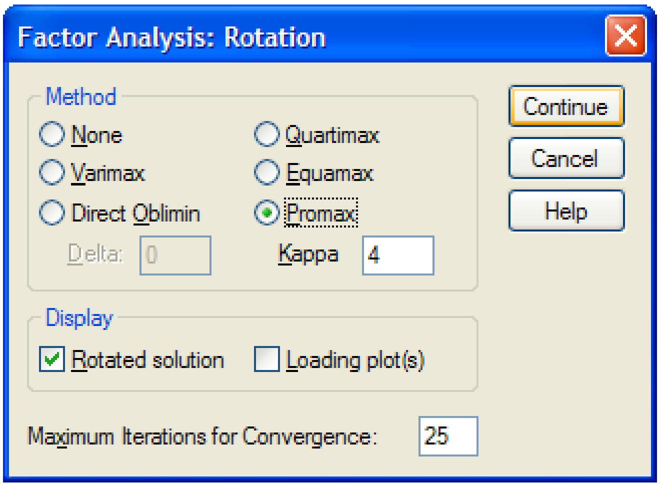
Analyze
● Correlation matrix
○ Covariance matrix

Display
☑ Unrotated factor solution
☑ Scree plot

Extract
○ Eigenvalues over: 1
● Number of factors: 4

Maximum Iterations for Convergence: 25

Rotation will aid interpretation of factor loadings. It is generally better to start with an oblique rotation such as Promax, especially when the factors are anticipated to be related.



Factor Analysis: Rotation

Method

○ None ○ Quartimax

○ Varimax ○ Equamax

○ Direct Oblimin ● Promax

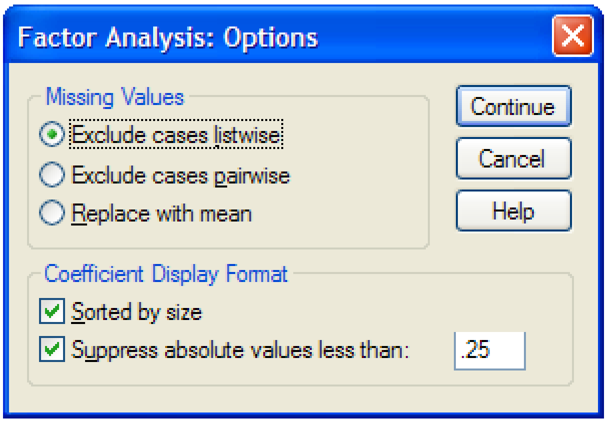
Delta: 0 Kappa 4

Display

☑ Rotated solution □ Loading plot(s)

Maximum Iterations for Convergence: 25

Finally, we will selection under 'options' 'sorted by size' and specify 'suppress absolute values less than' to be .25. This will make the meaning of factors to be easier to interpret by sorting them into groups based on factor loadings and hiding the small loadings.



Factor Analysis: Options

Missing Values

● Exclude cases listwise

○ Exclude cases pairwise

○ Replace with mean

Coefficient Display Format

☑ Sorted by size

☑ Suppress absolute values less than: .25

The first new bit of output is the communalities. This shows the percentage of variance accounted in each item by the extracted factors. We would be concerned if there were some items with particularly small values here (perhaps <.2). In this data, these values look reasonable. We can also get a sense of which items are better represented by the extracted factors.

Communalities

	Initial	Extraction
overall I am very satisfied	.539	.651
I like my job	.487	.598
I find my job interesting	.559	.656
I am committed to my job	.558	.676
I respect the values of the organisation	.395	.498
I plan to stay with the organisation	.271	.367
I perform better than the average employee	.300	.561
I contribute effectively to the organisation's performanc	.243	.362
I add substantial value to the organisation	.238	.338
I respect my immediate supervisor	.338	.468
My immediate supervisor is very effective	.371	.573
Central management are performing well	.260	.373

Extraction Method: Maximum Likelihood.

In this table we can see the percentage of variance accounted for by each factor. Overall 51% of the variance in the items is accounted for by four factors.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	4.534	37.787	37.787	4.082	34.015	34.015	3.425
2	1.457	12.139	49.926	.908	7.570	41.585	3.141
3	1.224	10.198	60.124	.695	5.791	47.376	2.633
4	.888	7.400	67.523	.435	3.624	51.001	2.057
5	.675	5.623	73.146				
6	.589	4.910	78.056				
7	.549	4.573	82.628				
8	.513	4.278	86.907				
9	.490	4.080	90.987				
10	.386	3.219	94.205				
11	.353	2.939	97.144				
12	.343	2.856	100.000				

Extraction Method: Maximum Likelihood.

a. When factors are correlated, sums of squared loadings cannot be added to obtain a total variance.

The factor matrix shows the correlations between the items and the factors prior to rotation. The interpretation of a factor is determined by which items load highly on it and the direction of the loadings. We can see that the first factor is related to all the items in the survey. The second factor includes the performance items, the third factor reflects satisfaction with leadership and the fourth factor reflects two items from commitment. Because this is prior to rotation the interpretation is generally not as clear. By definition the first factor is forced to account for the most variance.

Factor Matrix[a]

	Factor			
	1	2	3	4
I am committed to my job	.782			
I find my job interesting	.780			
overall I am very satisfied	.763			
I like my job	.712			
I respect the values of the organisation	.627			.311
My immediate supervisor is very effective	.554		.458	
I respect my immediate supervisor	.526		.375	
I plan to stay with the organisation	.485			.308
I perform better than the average employee	.390	.624		
I add substantial value to the organisation	.355	.460		
I contribute effectively to the organisation's performance	.339	.459		
Central management are performing well	.410		.416	

Extraction Method: Maximum Likelihood.

a. 4 factors extracted. 4 iterations required.

Rotation serves to distribute the variance accounted for by the four factors more evenly so as to improve properties of what is referred to as simple structure. This should lead to a closer to one-to-one mapping of items to factors, which should in turn make interpretation of the relationship between items and factors clearer. In the present case interpretation has converged perfectly with theory. The first factor corresponds to job satisfaction, the second factor to commitment, the third factor to leadership and fourth factor to performance.

Pattern Matrix[a]

	Factor			
	1	2	3	4
I like my job	.771			
overall I am very satisfied	.741			
I find my job interesting	.656			
I respect the values of the organisation		.659		
I plan to stay with the organisation		.648		
I am committed to my job		.612		
My immediate supervisor is very effective			.718	
Central management are performing well			.643	
I respect my immediate supervisor			.617	
I perform better than the average employee				.765
I contribute effectively to the organisation's performance				.586
I add substantial value to the organisation				.559

Extraction Method: Maximum Likelihood.
Rotation Method: Promax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

The factor correlation matrix reflects the fact that we chose an oblique rotation. This allows the factors to be correlated. We can see that all of the factors are correlated with each other. For example, factor 1 (job satisfaction) is correlated .676 with factor 2 (commitment). This reinforces the decision to use an oblique rotation that allows for correlated factors.

Factor Correlation Matrix

Factor	1	2	3	4
1	1.000	.676	.561	.435
2	.676	1.000	.535	.412
3	.561	.535	1.000	.324
4	.435	.412	.324	1.000

Extraction Method: Maximum Likelihood.
Rotation Method: Promax with Kaiser Normalization.

We might then consider creating scales based on these factors. Overall, the results showed good support for the theorised factor structure, although it should be noted that one factor did a reasonable job of accounting for most of the variance in the items. This suggests that there might be one overall satisfaction-based measure with a number of more specific domains.

18. RELIABILITY ANALYSIS

Theory

An important property of a measurement instrument is that it is reliable.

The most commonly reported measure of reliability is chronbach's alpha. It is a measure of internal consistency. Thus, if you have six items that are all meant to measure the same thing, chronbach's alpha will give an estimate.

Rules of thumb for interpretation:

>.8 excellent

>.7 good

>.6 mediocre

<.5 poor

It is an estimate of the correlation between the observed score and the true score.

Strictly speaking, reliability is not a property of a test. It is a property of a test applied to a particular sample in a particular context.

To know whether to calculate reliability, you can ask the question: am I using a total score that is made up of components?

Assumptions

Sample Size

In order to get a reasonable estimate of the reliability, you need a reasonable sample size. As a rough rule of thumb you might desire at least 80 to 100 people before calculating reliability.

Item reversal

Any reverse scored items have been reversed.

For example, if you have three items measuring 'happiness': 1) Are you happy; 2) Do you like your life; 3) Do you sometimes feel unhappy. The first two aim to measure happiness and the third attempts to measure the opposite of happiness. Thus, prior to adding the items up to form a total or running reliability analysis, you would need to reverse the negatively worded items so that high scores on item 3 now reflect happiness.

Items are continuous or binary

To perform chronbach's alpha the items need to be ordered or binary. While it is possible to use other reliability tools for categorical data, these can not be analysed with the main reliability tool in SPSS.

SPSS

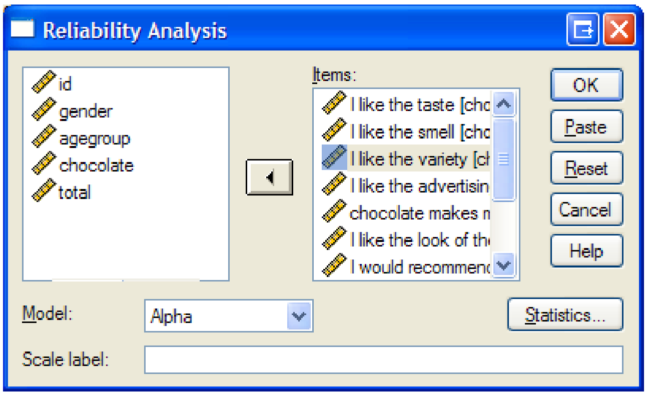
Case Study

Once again we look at the market research study with eight items looking at liking of chocolate. Imagine that you want to create an overall 'liking of chocolate' scale. The idea is to combine the eight items together to form an overall measure.

Running

Scale >> Reliability Analysis

Place the items that make up the scale in "items"



Reliability Analysis

- id
- gender
- agegroup
- chocolate
- total

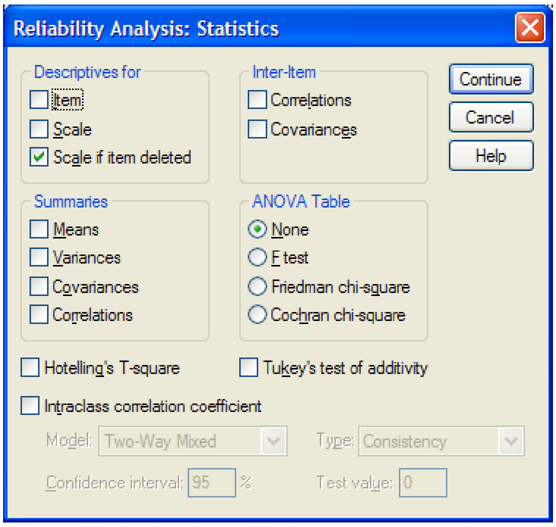
Items:

- I like the taste [cho
- I like the smell [cho
- I like the variety [ch
- I like the advertisin
- chocolate makes m
- I like the look of th
- I would recommen

Model: Alpha

Scale label:

Under 'Statistics', there are a number of other options that can be useful, but the main option to select is 'scale if item deleted'.



Reliability Analysis: Statistics

Descriptives for
- □ Item
- □ Scale
- ☑ Scale if item deleted

Inter-Item
- □ Correlations
- □ Covariances

Summaries
- □ Means
- □ Variances
- □ Covariances
- □ Correlations

ANOVA Table
- ● None
- ○ F test
- ○ Friedman chi-square
- ○ Cochran chi-square

□ Hotelling's T-square

□ Tukey's test of additivity

□ Intraclass correlation coefficient

Model: Two-Way Mixed

Type: Consistency

Confidence interval: 95 %

Test value: 0

Output

These first two tables set out the number of participants and then display the internal consistency reliability and the number of items in the scale. Based on the rules of thumb presented earlier, a Chronbach's Alpha of .90 constitutes an excellent reliability.

Case Processing Summary

		N	%
Cases	Valid	250	100.0
	Excluded[a]	0	.0
	Total	250	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
.900	8

The next table shows a number of bits of information. ‘Corrected Item-Total Correlation’ shows the correlation between the scale total if the item was not used to form the total. If you see a negative number here, this may suggest that you have not reversed an item. In general you want this number to be closer to 1, but anything above approximately .2 or .3 is reasonable. ‘Chronbach’s Alpha if Item Deleted’ tells you what your reliability would be if the item was not in the scale. If any off these values are higher than our alpha in the table above, then you might consider removing the item from the scale, because it is reducing the reliability of the scale. In this case none of the items show a value above the overall Chronbach’s Alpha.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
I like the taste	24.1680	48.116	.691	.887
I like the smell	24.2560	49.043	.672	.888
I like the variety	24.2720	47.902	.722	.884
I like the advertising	24.2200	47.771	.692	.887
chocolate makes me feel good	24.2760	49.911	.603	.894
I like the look of the chocolate	24.2520	47.490	.745	.882
I would recommend this chocolate	24.1360	48.447	.676	.888
this chocolate is good value	24.2640	47.930	.682	.888

An analysis of the internal consistency reliability was performed on the eight items measuring liking for chocolate. The overall chronbach’s alpha was .90, which represents an excellent reliability. Analysis of corrected item-total correlations and alpha if item deleted statistics revealed that all of the items contributed positively to the overall reliability of the scale.

19. FINAL DISCUSSION

Getting further assistance

Overview

In general it is desirable to have two types of books when doing data analysis. It is desirable to have a solid multivariate statistics book to provide answers to the more sophisticated questions that might arise. It is also beneficial to have a book that guides you through the basic steps of running and using SPSS. The present text largely fills this role, but others are listed below.

Statistics Books

Tabachnick & Fidell - Using Multivariate Statistics (4th Edition)

Hair, Anderson, Tatham, & Black – Multivariate Data Analysis

SPSS books

Coakes & Steade - SPSS without Anguish (5th Edition)

Julie Pallan - SPSS Survival Manual 2nd Ed

Internet

http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

Google search: name of statistics technique

Revision of the Research Process

Review Literature

Research typically starts with information gathering. This typically involves brainstorming to clarify the research question; researching and finding information about how the question has been answered previously; and synthesising the information to form possible ideas to test and explore.

Design Study

At this point, the sample has to be specified in terms of the number of people desired and the characteristics of the desired sample. Sampling and recruitment procedures need to be decided upon. Measurement instruments need to be designed or selected. Administration and procedures need to be designed. Much more could be written on this process.

Define Hypotheses and corresponding statistical analyses

It is important to clarify your research questions. What is the purpose of conducting the study? What are the main variables that will be measured? How can the research design effectively test these ideas? How have the abstract questions been operationalised into an empirical study?

Enter Data

Think about how the data will be analysed. Remember to use an ID number. Develop a coding scheme for defining the different variables that will be used.

Verify accuracy of data

Check that values make sense. Review the section on data validation.

Perform preliminary data manipulations

Create scales

Restructure data if necessary

Deal with missing data problems

Explore descriptive statistics

Examine measures of central tendency and spread for variables

Examine graphs

Perform Main Analyses

Determine what the appropriate statistical procedures are. A starting point for determining this involves determine what types of variables you have. Are the variables categorical, ordinal, binary, interval or ratio? A second question is whether you are interested in differences between groups or associational measures such as correlation, regression and factor analysis.

Once techniques have been selected, it is important to assess assumptions and decide about the various options in how to conduct the particular method.

Write up results

Report some measure of effect size

Use examples of other researcher's reporting style as a guide

Tell a story with the results

Be selective in the information presented

Demonstrate the meaning

Don't be more complex than necessary

Provide a reasoned argument

Demonstrate an understanding of options and why the approach you have adopted is the best

Provide references to justifications for statistical decisions (e.g., assumptions, sample size requirements, method selection) – preferably journal articles

Integrate the results with the hypotheses and build on them in the discussion

Integrate Results with Theory

Remember that statistics is a tool that is used to answer a research questions

Some Golden Rules

Stay close to the data

Link analyses back to theories

Try it both ways and if it doesn't make a difference then it probably doesn't matter